An analysis of dropout students in the German higher education system using modern data mining techniques

Dissertation
zur Erlangung des Doktorgrades
„Dr. rer. pol."

der Fakultät für Wirtschaftswissenschaften
der Universität Duisburg-Essen

vorgelegt von
Marco Giese
aus
Olpe, Deutschland

Betreuer:
Prof. Dr. Andreas Behr
Lehrstuhl für Statistik

Essen, 10.12.2020

Gutachter

Erstgutachter: Prof. Dr. Andreas Behr

Zweitgutachter: Prof. Dr. Christoph Hanck

Drittgutachter: Prof. Dr. Rainer Kasperzak

Tag der mündlichen Prüfung:

13.04.2021

DOI: 10.17185/duepublico/74317

URN: urn:nbn:de:hbz:464-20210526-091817-1

Acknowledgment

Since July 2017, I have worked as a research assistant at the chair of statistics. From July 2017 to March 2020, I was a member of the project "determinants and models for predicting study dropouts" (DMPS), funded by the "Bundesministerium für Bildung und Forschung" (BMBF). I am grateful to the BMBF for allowing me to work on the very specific research field of educational data mining for such a long period.

During this time, and also after the project expired in March 2020, I received support from various people in the department. First of all, I would like to thank my supervisor Prof. Behr, who was also the mentor of the DMPS project and supported this thesis in many ways. Further big thanks goes to Dr. Katja Theune, the head of the project DMPS, and my college Hervé Donald Teguim Kamdjou, who both were a great help for the first five articles of this thesis.

Furthermore, I would like to thank Prof. Hanck, who agreed to take over the second review of the thesis.

Contents

List of Figures

List of Tables

1 General introduction

1.1 Motivation

This thesis deals with the useful application of advanced statistical data mining in the area of higher education research. The focus lies on the statistical models and their results but these models also provide meaningful starting points to improve the higher education system.

Study success and study dropout are of growing interest. This is particularly caused by the rapidly increasing enrollment rates in recent years and the constantly large proportion of dropout students (detailed numbers are provided in the next section and in the seven articles in the main body of this thesis). Furthermore, there is a growing demand for highly qualified specialists in the labor market, especially in Science, Technology, Engineering, and Mathematics (STEM fields). This is why the private sector is also interested in an increasing number of higher education graduates.

The higher education system is a major public cost factor and a study dropout is generally seen as a waste of public resources. Additionally, a study dropout is often regarded as a personal failure of the student. This is particularly the case if the dropout is non-voluntary, for example, if it is caused by too many failures in an examination. In such cases, extra tutorials could help students who have problems with the subject to pass their examinations. Other students may drop out voluntarily because they are no more interested in the study field or realize that graduates in their study field have poor job prospects. This might be due to students not being sufficiently well informed when starting their studies. Additional information events could help such students to find the right study field for them or even recommend a vocational training program.

Due to the rising interest of policymakers, the "Bundesministerium für Bildung und Forschung (BMBF)" (Federal Ministry of Education and Research) started the funding line "Study success and study dropout" in 2016, which was extended by a further funding line in 2020 (BMBF, 2020). This indicates the huge importance that the topic currently has and that research in this field can help to improve the actual situation.

Many strategies focus to help students who have problems in their studies and are at risk to drop out. A common difficulty in tertiary education is to identify such students. Compared to primary and secondary education, there is usually less contact between the teaching staff and the students in tertiary education. The only feedback the teachers receive about their students is often limited to the grade in the examination at the end of the semester. At this point, it is often too late to help such students.

This is the point where my research gets relevant. In the research field of educational data mining, big educational datasets are analyzed with modern data mining and machine learning methods. This can help, for example, to identify students at risk at an early stage of study just before failing important examinations. Such models can be used to integrate early warning systems at tertiary education institutions.

A further aspect of this thesis are the reasons why students leave university without a degree. Methods of cluster analysis help to find different types of study dropouts.

The research field is the interface between the traditional educational research, where predominantly basic statistical methods are used, and an application of advanced statistical and mathematical models.

To get a general impression of the dropout phenomenon at tertiary education, the next section presents some numbers and actual trends in tertiary education.

1.2 Extents and trends of higher education entrance and graduation rates

To give an overview of dropout rates in Germany, which are presented in Table 1.1, I refer to the study of Heublein et al. (2017). In Germany, a distinction is made between general universities, and universities of applied sciences, which concentrate more on practical

aspects. Since dropout is not defined consistently in the literature and complicated to measure, organizations as the Organization for Economic Co-operation and Development (OECD) or the federal statistical office of Germany do not state official dropout numbers. In this thesis higher education dropout is consistently defined from a macro-perspective as cases where students finally leave the higher education system without obtaining a first degree. I mainly concentrate on the first higher education degree, which is in general a Bachelor's degree. This means that changes of university or study program are not considered as dropouts. My definition is equivalent to the definition by Heublein et al. (2017). For their estimation of dropout rates in Germany, they compare the cohort of a specific graduation year with the enrollment numbers of all corresponding freshmen years.

Table 1.1: Dropout rate in Germany in percent based on the numbers of Heublein et al. (2017)

Degree	2006	2008	2010	2012	2014
Bachelor total	30	28	28	28	29
Bachelor universities	25		35	33	32
Bachelor universities of applied sciences	39		19	23	27
Master universities				11	15
Master universities of applied sciences				7	19

Dropout rates are generally larger in Bachelor programs because in Master programs students usually know what they can expect and have already proven their abilities when obtaining a first degree. Usually, dropout rates in universities of applied sciences are smaller compared to general universities. This result is confirmed by various studies, as shown in the literature review provided in chapter 2. The large proportion of dropouts in Bachelor programs of universities of applied sciences in 2006 and in Master programs at universities of applied sciences in 2014 must be considered with caution since they are not supported by other empirical findings.

In the following, I will present some actual numbers providing insight into the recent development of higher education entrance and graduation rates in Germany, the United Kingdom (UK), and the United States (US). Table 1.2 shows the estimated graduation and entry rates for Germany, the UK, and the US in 2005 and from 2011 to 2014. There is a rising trend in graduation, and in entry rates in Bachelor and Master programs in all three countries. Germany has lower graduation rates in Bachelor and Master programs than the UK and the US.

Table 1.2: Percentage of people estimated to graduate with a Bachelor, Master or Doctoral degree in Germany, the UK and the USA from 2005 to 2014 once in their lifetime (entry rates for Germany and the UK in parentheses); OECD (2020)

	Germany			UK			USA		
Year	Bachelor	Master	Doctor	Bachelor	Master	Doctor	Bachelor	Master	Doctor
2005	13.89 (23.00)	13.58 (22.67)	2.32				32.52	17.27	1.39
2011	(41.39)	(20.23)	(5.21)	41.21 (61.55)	24.20	2.45	37.39	19.86	1.43
2012	(47.93)	(22.90)	(5.38)	44.15 (63.24)	25.51	2.38	37.74	20.37	1.46
2013	27.49 (48.13)	16.31 (24.65)	2.72 (5.41)	44.54 (60.17)	26.66 (30.87)	3.00 (3.99)	37.87	20.25	1.52
2014	30.18 (51.94)	17.08 (27.77)	2.79 (5.52)	49.92 (63.74)	26.37 (31.94)	2.88 (4.14)	38.16	20.03	1.57

In 2005 the relatively small percentage of the German population estimated to earn at least one Bachelor's degree in their lifetime. This is because the Bachelor and Master programs in Germany were introduced after the Bologna reform in 1999. Before the Bologna process, the Diploma was the standard degree in Germany, which was characterized by a long study duration of 8-10 semesters. The Diploma is equivalent to a Master's degree (ISCED2011 level 7) (OECD, 2020). The second major alteration of the Bologna process was the implementation of the European Credit Transfer System (ECTS). These reforms aimed to unify the European tertiary education system.

To get a detailed overview of the study fields, Table 1.3 shows the relative numbers of students in different study fields by degree type. Some study fields reveal a high rate of Bachelor graduates, but a smaller rate of Master graduates and persons with a doctoral degree in this field. In Germany, for example, this is true for Engineering, Manufacturing, and Construction, where in many professions a Bachelor's degree is sufficient. In other subjects, as Health and Welfare, it is very common in Germany to get a doctoral degree.

As mentioned in section 1.1, the dropout of a student is often considered as a waste of public resources. Each student in tertiary education institutions costs on average 18,486 US Dollars per year in Germany, which is a rather small amount compared to the United Kingdom with 28,144 US Dollars and the United States with 33,063 US Dollars but more than the OECD average of 16,329 US Dollars (OECD, 2020). The huge difference is due

Table 1.3: Percentage of graduates in a specific education field in Germany, the UK and the USA in 2014; OECD (2020)

	Germany			UK			USA		
Field	Bachelor	Master	Doctor	Bachelor	Master	Doctor	Bachelor	Master	Doctor
Education	12.1	10.8	3.1	4.5	18.8	4.5	5.3	18.1	15.3
Humanities and Arts	8.3	19.3	7.5	20.7	9.9	15.2	16.8	7.4	11.6
Social Sciences, Business and Law	32.2	25.9	15.1	28.8	38.6	15.7	38.7	37.5	19.9
Science, Mathematics and Computing	11.7	16.1	31.9	20.6	10.6	32.5	11.2	6.1	26.6
Engineering, Manufacturing and Construction	24.1	15.9	11.1	8.4	9.2	14.1	6.3	6.5	15.4
Agriculture and Veterinary	2.1	1.6	3.0	0.9	0.6	1.2	1.1	0.7	1.5
Health and Welfare	4.7	8.5	27.5	14.3	11.4	16.7	11.7	20.9	7.9
Services	4.7	1.9	0.8	1.8	0.9	0.2	8.8	2.9	1.8

to the fact that contrary to the US and the UK, students in Germany do not have to pay tuition fees for public higher education institutions.

1.3 Brief overview of the seven articles

A major part of this thesis, namely the first five articles, have been written during the project "Determinanten und Modelle zur Prognose von Studienabbrüchen (DMPS)" (determinants and models for predicting study dropouts). The project was financed by the BMBF[1] as part of the BMBF funding priority line "Study success and study dropout". These five articles were written in cooperation with my college Hervé Donald Teguim Kamdjou, the principal investigator Dr. Katja Theune and Prof. Dr. Andreas Behr, who was the mentor of this project.

To provide a general impression of the research field, the first article gives a comprehensive overview of the literature in the research field of higher education dropout. The central aim of the DMPS project was the modeling of the complex dropout process, which usually depends on various determinants. Therefore, methods of the statistical field of data mining have been used. The research field of educational data mining has been growing rapidly in recent years. But before the start of the BMBF funding line, there was a lack of German studies investigating the dropout process with such modern methods of data mining. The first research priority during the project was the modeling of students who graduate and students who leave the higher education system without

[1]Gefördert vom Bundesministerium für Bildung und Forschung (BMBF), Förderkennzeichen: 01PX16006

obtaining a first tertiary degree, which is, in general, a Bachelor's degree. This is a binary classification problem. In a first step, decision trees and random forests were used for classification, since their results are easily comprehensible in contrast to other "black-box" (Pochiraju and Seshadri, 2018) machine learning methods. In a further step different machine learning methods, including black-box methods like support vector machines, were compared with regard to their model performance. The last research interest of the project was to find different types of study dropouts. This was done with a clustering approach.

The last two articles of this thesis were not written in the context of the DMPS project. In the sixth article, I forecast the final grade of students for their first higher education degree on the basis of pre-study and early-study determinants. In the seventh article, a sequential dropout and graduation model was fitted, making use of the panel structure of the data. Both of these two articles use statistical methods that reduce the potential bias evoked by panel attrition, which is a well-known problem for survey panels.

All studies, except the literature review, used the starting cohort 5 of the National Education Panel Study (NEPS) as data basis. This covers in total 17,910 freshman students of the winter semester 2010/11 in German higher education institutions. In irregular time spans new survey waves are available in the NEPS data. As we started writing the second article, which is the first empirical article, the NEPS covers 9 waves, while in the latest article 14 waves were available. A more detailed description of the NEPS is provided in the individual articles.

Since this is a cumulative thesis, each article must be independently readable to be published. Therefore, some sections as the data description and parts of the literature review are very similar in each article. Furthermore, the problem of panel attrition, which is a very common problem in panel surveys, is widely discussed in most articles. Higher education dropout is defined consistently in all studies, according to Larsen et al. (2013b), as students who leave the higher education system without a degree and do not return to higher education at a later point in time.

1.3.1 Dropping out of university: a literature review

The first article provides a comprehensive literature review of pertinent research in the field of higher education dropout in Europe. The article is structured in three main parts, one on theoretical dropout models, one dealing with determinants influencing the dropout decision, and one devoted to an outlook on the following research.

Theoretic dropout models can roughly be divided into economic, psychological, and sociological models. Economic models focus on the monetary costs (e.g. opportunity costs of studying) and the returns of a higher education degree, which is strongly related to the human-capital-theory. Psychological models mainly consider the academic and social integration of students, their personal attitudes, persistence, and their perceived value of the higher education degree. In sociological models, the university dropout is seen as a longitudinal process where the institutional equipment and the social system (e.g. financial aid system for students) affect students' satisfaction and willingness to complete their study program. The model of Tinto (1975) is one of the most important models in the current dropout literature.

From more than 200 articles in the field of higher education dropout, 35 satisfied our standards and were hence selected for this literature review.

The most relevant determinants that frequently occur in the selected literature, were structured in three groups. The first group of variables concerns the national education (and financial aid) system. Students from the upper secondary education track (in Germany the "Gymnasium") have, for example, larger graduation chances. Higher grants for students can reduce socio-economic-disadvantages.

Variables of the institutional level concern the type of institution, the study field, and study conditions. In Germany, universities of applied sciences reveal lower dropout rates than general universities. In fields where strong mathematical skills are relevant, like Engineering, Mathematics and Natural Sciences, the highest dropout rates were observed.

The third group of variables covers students' individual variables, including pre-university determinants such as gender, age, school grades, family and migration background, and study related determinants like study organization, self-esteem, satisfaction, and potential off-campus work.

The review ends with the identification of research gaps, which are addressed in the following articles.

1.3.2 Dropping out from higher education in Germany - an empirical evaluation of determinants for Bachelor students

The first empirical article with the NEPS data starts with a binary analysis to detect variables that largely differ between the two groups of higher education graduates and dropouts. The mean in the two groups is compared using two effect size measures, namely Cohen's d and the point biserial correlation. Variables in five different thematic fields are ranked according to the largest absolute effect sizes.

Surprisingly, the group of graduates spends more time with off-campus work during the semester break, but less time during the semester compared to the group of dropout students. Female students are more frequently in the group of graduates. Members of the graduation group are also better prepared for study, have better grades at secondary school, study more frequently at universities of applied sciences, are more satisfied, find the degree course more interesting, and reveal a higher extrinsic motivation. These are only some examples of almost 200 variables that are analyzed, whereof 52 variables are presented in the article.

The simple bivariate analysis provides information about the fact that, for example, women are more likely to complete their studies, but this may also be caused by other determinants such as the field of study. Therefore, a multivariate logistic regression should provide information about possible interdependencies. Interactions between two variables are hence included in the model. The LASSO method (Tibshirani, 1996) should reduce the number of relevant features since this rapidly increases if interactions are included in the model. The logistic regression model reaches an accuracy of 73.35% in the classification of dropouts and graduates, a mean squared error (MSE) of 0.318, and an area under the curve value (AUC) of 0.796.

The empirical analysis provides valuable starting points for developing intervention strategies. For instance, dropout students are less informed about the study. This problem can be tackled by additional (mandatory) information days for students who want to enroll in a higher education program.

1.3.3 Early prediction of university dropouts - a random forest approach

This article aims to predict study dropouts with the NEPS data as early as possible. The random forest based on conditional inference trees (Hothorn et al., 2006) seems to be an appropriate method for this binary classification problem because it can handle missing values (using surrogate splits), is able to use all common types of variables (e.g. nominal and metric variables), and has no problems with extreme outliers. Furthermore, the random forest reveals an unbiased importance raking of variables if subsampling (drawing without replacement) is used instead of bootstrapping (Strobl et al., 2007).

To model the dropout process at a very early time, only variables from the pre-study period were used in the first classification step. This includes variables that are relevant up to the end of secondary education or vocational training, but before the start of the study decision process. This has the advantage that institutions can initiate countermeasures for potential dropout students at a very early time (just before students decide to study) on the costs of relatively poor model accuracy. In a second step variables of the study decision phase were added as explanatory variables (after secondary school but before the start of the higher education program). Finally, the third step additionally considers variables of the early study phase (the first two semesters), leading to overall 81 explanatory variables in the last step.

Separate models are fitted for four major subject groups and a further model for all students in the NEPS data. Cross-validation is applied to get out-of-sample predictions. The predictive performance for the complete sample increases from an AUC value of 0.77 in the pre-study episode, to 0.86 in the model containing also variables of the early study phase. Additional variables of the study decision phase only lead to a minor model improvement. Best predictions are possible in the field of mathematics and natural sciences since hard (mathematical) skills are better covered by the data than, for example, the talent for arts.

The most relevant variable in all models is the final grade at secondary school. Further important variables are the age (younger students have better graduation chances), the overall satisfaction, and students' helplessness.

1.3.4 Predicting dropout from higher education - a comparison of machine learning algorithms

This empirical work follows directly from the previous article. Here, different machine learning algorithms are compared in order to find the best algorithm for this specific classification problem. Only one model is deployed, including all 81 variables that emerged as explanatory variables of the study-related episode in the previous investigation.

Five machine learning models are compared, namely the naive Bayes model, which serves as a benchmark, the logistic regression, the support vector machine, and two tree-based classifiers, namely the random forest based on classification and regression trees and the AdaBoost algorithm. Since not all of these models can handle missing values, these values are imputed via median imputation, which surprisingly leads to the best model performances and outperforms more sophisticated imputation strategies. Potential tuning parameters of the different models are optimized in the inner loop of a nested cross-validation.

The best models for these data turn out to be the two tree-based classifiers, which both receive an AUC value of 0.87, respectively and a root mean squared error (RMSE) of 0.24 for the model including all subject fields. The poorest results were reached with the naive Bayes classifier with an AUC value of 0.81 and an RMSE of 0.35, which is caused by the naive and violated assumption of stochastic independent variables.

Further model improvement can be reached by model stacking, where different models are combined to a single classifier. This approach leads to a small increase of the AUC value to 0.88, while the RMSE takes a value of 0.24, just like the tree-based classifiers. The main disadvantage of this "black-box" technique is that it is difficult to communicate to potential users due to its complexity and its huge computation-intensity.

Potential users, such as higher education institutions intending to incorporate early warning systems, will presumably prefer one of the tree-based classifiers due to their easy implementation, simple comprehensibility, and comparably good model performance.

1.3.5 Motives for dropping out from higher education - an analysis for Bachelor students in Germany

The focus here is on 24 motives of 662 students for leaving the higher education system. To reduce the dimension of the data the 24 reasons for dropping out were clustered into six thematic fields, namely study conditions, performance and requirements, interest and expectations, job alternatives and career, personal and family aspects, and financial aspects.

The results reveal that only 14.8% of students have only one major reason for dropping out, while generally a bundle of different reasons accumulates and persuades students to leave higher education without a degree. The major single dropout motives are lacking interest in the study field, wrong expectations, failed examinations, high study requirements, and the wish for more practical work.

Students who drop out in the later study years more often fail in examinations. In the fields of Engineering, Mathematics and Natural Sciences dropout is more often the result of financial problems, while in Linguistics and Cultural Sciences students leave university more frequently due to poor job opportunities.

Clustering the six groups of reasons for leaving the higher education system without a degree leads to eight different dropout clusters. More than 56% of all dropouts state reasons from nearly all six thematic fields for leaving the university. Almost 27% state that a lack of interest, wrong expectations, poor performance and high study requirements are their main reasons for dropout. Only a minority of 5 % state that mainly personal and family reasons are responsible for their dropout. For 8% of students, only a lack of interests and the wrong expectations are their predominant dropout reasons.

The results show the main reasons for leaving higher education, and thus indicate where suitable countermeasures to dropouts should start. Since a lack of interest in and wrong expectations about the study field can be due to freshmen being insufficiently informed, additional information events could help to prevent students from dropping out.

1.3.6 Predicting higher education grades using strategies correcting for panel attrition

In this article, I forecast the final grade of the first higher education degree, which is generally a Bachelor's degree, at two different stages. In the first stage, only pre-university variables are used to generate very early predictions that could also help students with their study decision. The second stage additionally contains determinants from the early study phase. Note that the variables used for this article are a bit different from those in previous articles, which aim to predict dropout students.

Inverse probability weights of students' response probability and the Heckman correction should reduce the bias in the estimated parameters caused by panel attrition. The usual ordinary least squares estimator (OLS) serves as a benchmark model.

I also distinguish between two scenarios. In the first scenario dropout students are excluded from the model. The second scenario includes dropout students with a grade of 5.0. The problem in the second scenario is that there is a huge gap between the worst graduates (4.0) and dropouts (5.0) leading to a mixture of a continuous- (only graduates) and a discrete distribution. After a transformation of the grades (dropouts should be 0) this leads to a zero-inflated continuous distribution. Generalized linear models allow also other distributions than the Gaussian, which is used in the OLS model, as long as the distribution belongs to the exponential family. The Tweedie distribution is one of these distributions, which generalizes some other distributions including the Poisson-Gamma mixture distribution which is used here.

Only 16.9% of the variance can be explained by the model if only pre-university variables are included in the scenario where dropouts are excluded. The R^2 improves to 52.2% if variables of the early study phase are added to the model. If dropouts are included in the latter scenario, the R^2 rises from 0.415 for the benchmark OLS model to 0.550 for the Tweedie model.

The inverse probability weight estimator and the Heckman correction provide slightly different parameter estimates than the benchmark model, which indicates that the bias is not too large even when the attrition problem would be ignored.

The most striking result in the educational context is that some parameters are significant in the model where dropouts are included, but not significant if dropouts are excluded. These are mainly variables that influence the dropout process, and also found

to be significant in the actual dropout literature, but have no influence on study performance if a student graduates.

1.3.7 Prediction of time-dependent dropout and graduation rates in higher education under the presence of panel attrition

In the final article of this thesis, I take the panel character of the data into account to construct a time dependent dropout and graduation model. Since especially dropout students are affected by panel attrition, this group is synthetically oversampled to generate dropout and graduation proportions that are comparable to the numbers of Heublein et al. (2017).

The classification problem in this situation has three possible classes: dropout, graduation, and study continuation. The model predicts the status of the next wave with variables of the actual wave. Therefore, the actual status of a student must be "still studying" and the student must participate in both waves. This reduces the number of instances in the model from wave to wave because more and more students graduate, drop out, or finally leave the panel.

A short model comparison reveals that the random forest is again well suited for this classification problem and even outperforms artificial neuronal networks. Best predictions are possible at the end of the standard period of study where the difference between graduates and dropouts increases. In later waves, the model suffers under a relatively small sample size and the model performance decreases.

This model also provides information about the important variables at different points in time in the study. At the beginning of the study determinants regarding prior education, family and migration background are more important, but their importance decreases in later waves. Study related variables are of high relevance during the complete study period.

The model sheds light on the time dependent development of individual graduation and dropout probabilities. This can help institutions to update their models regularly, e.g. after each semester, and improve their early warning systems for students at risk.

2 Dropping out of university: a literature review

Dropping out of university: a literature review

Andreas Behr, Marco Giese, Herve D. Teguim K., Katja Theune
Chair of Statistics
University of Duisburg-Essen, 45117 Essen, Germany

Abstract

This study provides a comprehensive review of the student dropout phenomenon from tertiary education. Student withdrawal is the result of a long decision-making process and complex interaction between several determinants. First, we provide an overview of definitions, theoretical models and perspectives of dropouts. Referring to previous theoretical and empirical evidence from a wide range of disciplines, we focus on a detailed discussion of determinants affecting student dropout decisions. There are three superior aspects why students leave the higher education system without a degree. These are 1) the national education system, e.g. the countries financing policy, 2) the higher education institutions, e.g. the type of the institution or the teaching quality, and 3) the students themselves, whereby the last point is subdivided in pre-study determinants, e.g. the secondary school type and study-related aspects, e.g. off-study work. Based on these findings, we discuss the implications for further research. Here, especially the application of modern data mining techniques on comprehensive data sets covering a wide range of relevant determinants may lead to new insights into the dropout process. The results shall provide helpful tools for universities wishing to implement early warning systems and to support students at risk at an early stage of study.

Keywords: student dropout, higher education, dropout prediction, educational data mining, review

2.1 Introduction

Since there is a growing number of first year students and an increasing demand for academically qualified persons, study success and student dropout from tertiary education constitutes a very current and politically relevant topic. Moreover, the percentage of people graduating with a Bachelor, Master or Doctoral degree has also risen over the recent years in many countries (OECD, 2017).

On the one hand, labour market projections claim a substantial reduction of poorly-qualified and an increase in the demand for highly-qualified occupations. A large surplus of unqualified and an increasing lack of higher education graduates are predicted for the next few years (Vogler-Ludwig et al., 2016) and already today, employers worry about the low supply of qualified workers on the labour market. On the other hand, there is indeed a rising number of students, but also a high number dropping out from higher education.

This literature review focuses on determinants of students' dropout and graduation in the tertiary educational system of European countries. Due to the Bologna process in 1999, the education systems are mostly comparable in structure. Table 2.1 provides the population share of 25-34 year-olds with a higher education degree in the different countries. For comparison, we also include the mean percentage from all 36 OECD states (OECD, 2017). Italy and Germany have a relatively small amount of tertiary education graduates.[1]

The dropout rates in higher education reported in the empirical literature vary widely because of being based on different definitions and relying on different databases. Estimations by Schnepf (2014) are presented in Table 2.1. High dropout rates will increase the expected shortage of highly educated and skilled university graduates and therefore this constitutes a very relevant aspect for labour market developments in the next decades.

Other important consequences associated with student dropout are individual, institutional as well as societal costs. Moreover, high dropout rates may point to an inefficient use of resources by universities or to a lack of teaching quality and probably decrease

[1] On the OECD homepage you can find the data for the other OECD countries and years `https://data.oecd.org/eduatt/population-with-tertiary-education.htm` (2019-10-02)

Table 2.1: Population with higher education (in %) in the cohort of 25-34 year-olds (OECD, 2017) and dropout rates in 2011 (Schnepf, 2014).

Country \ Year	1998	2003	2008	2013	2018	Dropout rate (2011)
Canada	45.6	53.2	55.6	57.6	61.8	-
Finland	-	37.9	38.3	40.0	41.3	18.5
France	29.6	37.9	40.8	44.1	46.9	17.9
Germany	21.5	21.8	23.9	30.0	32.3	14.7
Italy	9.0	12.7	19.9	22.7	27.7	34.1
Netherlands	27.5	32.1	39.8	42.9	47.6	28.3
Norway	-	-	45.6	46.6	48.2	17.4
OECD average	23.8	29.9	35.5	40.6	44.5	-
Spain	31.2	37.9	40.0	41.1	44.3	24.2
United Kingdom (UK)	26.0	33.3	43.3	48.3	50.8	16.3
USA	36.2	38.7	41.6	44.8	49.4	-

the reputation of universities. In order to motivate and encourage study success, universities are increasingly searching for programmes to address and reduce dropouts. These include expanded information activities, study advice throughout the course, and the introduction of mentoring programmes.

The current state of empirical research on student dropout from a wide range of disciplines has identified a number of relevant reasons for withdrawing from tertiary education. These include, among others, demographic (gender, age), individual (social background, school performance), psychological (motivation, attitudes), institutional (teaching quality, learning environment) and national determinants (financing policy). There is no consensus in literature regarding the importance of each of these factors. The difficulty is that withdrawing from university is hardly the result of short-term or spontaneous decisions, but rather of a long decision-making process, during which several conditions and problems accumulate and prompt students to leave university without a degree. Furthermore, there is rarely only one factor leading to dropout, but rather an inter-relationship of many factors from different areas. Many studies focus on so-called "hard" determinants, which are not within the sphere of influence of the university (for example age, gender, social background, school grades). Much less is known about "soft" determinants, e.g. attitudinal based and university-malleable factors such as motivation, satisfaction or integration, which may be positively affected by the institution (Larsen et al., 2013c). Overall, the phenomenon of student dropout is a very complex (decision-making) process, which has so far not been adequately and comprehensively described

by theories and captured by empirical research.

In this paper, we aim at providing an encompassing overview on the existing literature about the dropout phenomenon. This includes a discussion of definitions and theoretical models of tertiary education dropout. Our structured review focuses on different determinants affecting student withdrawal and previous empirical evidence related to dropout. Based on this, we discuss implications for further research and how empirical findings could be used for policy makers and higher education institutions wishing to implement "early-warning systems" for at-risk students.

The paper is structured as follows: section 2.2 addresses the definitions of the concept of student dropout and the most relevant theoretical perspectives. The different dropout determinants, which are related to the national system level, the institutional level and the individual level are presented in section 2.3. In section 2.4, we discuss implications for future research work, e.g. which kind of data is needed and which methods are useful to get new insights in the dropout process. Section 2.5 concludes.

2.2 Theoretical background

2.2.1 The concept of dropout

To describe the phenomenon of university dropout and its determinants in detail, it is necessary to operationalise the concept of dropout and to evaluate the most relevant theories in this context. Larsen et al. (2013c) discuss the different aspects and usage of the term "university dropout", which is commonly understood as leaving a university without a degree. University dropout is a very diverse concept and could therefore be operationalised in various ways. In most cases, the terms "dropout', "failure", "non-completion" or "withdrawal", are used synonymously. While the first three terms describe student dropout as a more negative and involuntary process, the term "withdrawal" stresses a more voluntary aspect of leaving university. Moreover, these terms are often used within a students' perspective, whereas "attrition" describes an institutional perspective. Positive counterparts are, for instance, "persistence", "completion" (students' perspective), "retention" or "graduation" (institutional perspective).

Especially psychologically motivated theories focus on the positive outcomes. In theoretical and empirical research, the different terms for negative as well as for positive outcomes are often used interchangeably.

Furthermore, the dropout phenomenon could be distinguished by students' motives to abandon study, and the degree of voluntariness. For example, from a student's perspective, dropout caused by academic failure would be perceived as non-voluntary. A more or less voluntary dropout would probably occur due to financial distress or other personal problems. Students might dropout entirely voluntarily because of more favourable job options (Larsen et al., 2013c). A voluntary dropout, therefore, could be seen as a revision of a disadvantageous decision because of unfavourable career possibilities or a bad match of study content and students' preferences. These various types of dropout are driven by different motives and causes. According to Tinto (1975), involuntary dropout is rather a result of insufficient academic integration, such as in the form of bad grades, whereas voluntary dropouts are mainly consequences of social isolation at university. In empirical research, many of these determinants indicating a more voluntary or a more involuntary dropout are analysed.

A further distinction should be made according to the level at which dropout occurs. Students might change their field of study (within the same subject area or between subject areas), the type of degree, the (type of) university, or students might simply leave the university system. Depending on the perspective, for instance from a student's or faculty perspective, these different types of dropouts would be perceived as transfers (e.g. from one field to another) or as a formal total dropout. The former is sometimes called "re-selection" (Larsen et al., 2013c) or "institutional departure" (Tinto, 1993, p. 36), the latter "de-selection" (Larsen et al., 2013c) or "system departure" (Tinto, 1993, p. 36). At least, dropouts should be distinguished according to their timing. Several studies state that dropouts at various stages of study are possibly driven by different factors, e.g. due to varying problems students face in the integration process (Tinto, 1988). In empirical research, one field of study or one institution is often analysed, whereby a change of the study field or of the institution is perceived as dropout. Studies with a broader perspective define dropout as leaving the higher education system, with the drawback that it is not observable if dropout students will enrol later again in their career. Moreover, many studies focus on the first semesters at university as it is the interest of institutions to prevent dropout at an early stage of study, in order to support students at risk as early as possible.

These distinctions between dropout types should be taken into consideration when analysing student departure. Tinto (1975) states: "Because of the failure to make such distinctions, past research has often produced findings contradictory in character and/or misleading in implication. Failure to distinguish academic failure from voluntary withdrawal, for instance, has very frequently led to seemingly contradictory findings that indicate ability to be inversely related to dropout, unrelated to dropout, and directly related to dropout. In other cases, failure to separate permanent dropout from temporary and/or transfer behaviours has often led institutional and state planners to overestimate substantially the extent of dropout from higher education" (Tinto, 1975, pp. 90).

2.2.2 Theoretical perspectives of students dropout and retention

We briefly describe the most relevant theoretical models on higher education (non-) completion, which have been developed over the recent decades. These theories originate from different disciplines and can mainly be divided into sociologically, psychologically, and economically orientated theories (see e.g. Sarcletti and Müller, 2011). Sociologically-oriented concepts emphasise the importance of social and academic student integration. In contrast, psychological theories focus on the role of student behaviour and attitudes in the dropout process. Economically motivated frameworks point to concepts of rational decisions and cost-benefit considerations as relevant determinants of dropout. Table 2.2 provides an overview of the most important theories and implied determinants.

Sociologically motivated models

The most influential dropout models are sociologically-motivated, for instance the well-known student attrition model developed by Vincent Tinto (Tinto, 1975, 1993). He refines and modifies a model proposed by Spady (Spady, 1970, 1971), based on Durkheim's theory of suicide (Durkheim, 1951), which suggests that social integration has a strong impact on the suicide decision. Spady's model combines psychological and sociological factors, whereby social integration plays a key role as it interacts with satisfaction and institutional commitment and thereby indirectly affects the decision to drop out. Tinto criticised psychologically grounded concepts, because they concentrate on student attributes and therefore claim dropout mainly as student failure. According to his model of

Table 2.2: Theoretical perspectives of student dropout/persistence

perspective	idea/concepts	key determinants
sociological	**dropout** as longitudinal process, students **interact** with social/academic system affecting their **social/academic integration**, modification of students' initial institutional commitment/goals; interactionalistic perspective, focus: **inside institution** (e.g. Spady 1970/1971, Tinto 1975/1993)	academic integration (e.g. grades, identification with academic norms and values), social integration (e.g. interaction with fellow students, extracurricular activities), personal goals (e.g. grades, graduation), institutional commitment, pre-entry attributes (e.g. family background, individual attributes)
	dropout as longitudinal process, interactions between students and the institution, also subjective measures of integration; **organisational perspective** (workforce turnover) (e.g. Bean 1980)	institutional structure and organisation (e.g. perceived practical value of education, quality of the institution), satisfaction, institutional commitment, pre-entry attributes (e.g. socioeconomic status)
	focus: also **outside institution** (Bean 1985, Bean and Metzner 1985)	additionally non-institutional factors (e.g. finances, friends outside university, off-campus living conditions)
	consistency of students' and institutions' habitus (**individual environment fit**) affects **retention**; cultural-capital-theory (e.g. Reay et al. 2001, Thomas 2002)	students' habitus/cultural capital (e.g. educational background of students, parents, peer group), institutional habitus (e.g. practices/norms/values of an institution)
psychological	emphasis on the role of students **psychological characteristics** for **interaction** behaviour and social/academic **integration** and **persistence**; attitude-behaviour-, coping-behavioural-, self-efficacy-, and attribution-theory (e.g. Bean and Eaton 2000/2001)	pre-entry psychological characteristics and through psychological processes developed characteristics (e.g. self-efficacy, attributions, motivation), academic and social integration, institutional fit/commitment
	students **expectations** of success and **perceived value** of college affect **persistence**; expectancy-value-theory (e.g. Ethington 1990)	expectations of study success (e.g. academic self-concept, perception of difficulty), individual valuation of a higher education degree (e.g. students' economic and social goals)
economic	**dropout/retention** as result of weighting **costs** and **benefits** of study/external alternatives; rational-choice-, human-capital-theory (e.g. Hadjar and Becker 2004, Becker and Hecken 2007)	e.g. expected returns to education (e.g. career prospects, accumulated human capital), monetary/mental costs (e.g. financial situation), expected educational success (e.g. grade performance)
phase model	complex model with **several theoretical perspectives**, **dropout** as a consequence of a **process of different phases**, different influencing factors in each phase, inter-relationship of individual qualifications and institutional conditions (e.g. Heublein 2014)	preliminary phase: background (e.g. social origin), personality (e.g. big 5), socialisation in education (e.g. school type), study decision(e.g. subject choice); current study situation: study conditions (e.g. teaching quality), information, behaviour (e.g. integration), motivation (e.g. identification), performance, psych./phys. resources (e.g. health), living conditions (e.g. financing), alternatives (e.g. vocational training); decision: individual motivation to drop out

student retention, the probability of withdrawing as opposed to persisting at university, depends strongly on academic and social integration. Both forms of integration are distinct constructs which affect each other mutually. Academic integration includes grades and identification with academic norms and values, while social integration comprises interaction with fellow students and extracurricular activities. Tinto depicts the dropout decision as a longitudinal process. His model is further inspired by the theory of rites of passage from van Gennep (1960), describing movements of individuals from one group membership to another. Applied to university dropout, students withdraw because they fail to separate from past associations and to incorporate new values and norms of the new academic environment (Tinto, 1988). According to Tinto's model, students have a set of pre-study attributes (like family background, prior education) which form initial institutional commitment, goals and intentions. When entering university, students start to interact with the academic and social system. The level of academic and social integration modifies students' initial institutional commitment, goals and intentions, which in turn determine students' decision to stay or to leave university. High integration intensifies these goals and commitments, resulting in student persistence. Low integration weakens goals and commitments, thereby promoting the decision to drop out. Tinto also states that events external to the university can affect dropout decisions, but mainly indirectly, due to their impact on student goals and institutional commitments (Tinto, 1975, 1993). Tinto was criticised for the exclusion of factors representing the non-institutional environment (Ulriksen et al., 2010).

Bean (1980) also recommends the concept of a longitudinal process of dropout, but as his work was derived from studies of turnover in work organisations, he emphasises the role of the institutional structure and organisation. Bean (1985) states that beside the factors included in Tinto's model, non-institutional factors, e.g. finances and friends outside university, also strongly affect dropout decisions. An empirical application shows that the probability of dropout increases with more outside friends and opportunities to transfer. The model of Bean and Metzner (1985) for non-traditional student attrition also stresses the role of environmental variables represented by students' off-campus living conditions. In this respect, social integration plays only a minor role, because non-traditional students are affected more by their non-institutional environment than by their social integration at university (see also Metzner and Bean, 1987). As both models of Tinto and Bean share some common features, Cabrera et al. (1992, 1993) recommend a combination of these models, so as to obtain a more comprehensive understanding of the dropout process.

Another model in line with Tinto's was developed by Ernest T. Pascarella (Pascarella, 1980), devoting explicit attention to the impact of students' informal contact with the faculty on dropout decisions. Furthermore, the model includes factors characterising a faculty's structural and organisational concept (Pascarella and Terenzini, 2005).

A further strand of sociologically motivated dropout models focuses on the role of "institutional habitus", cultural capital and the individual environment fit for student's withdrawal from university. Both concepts, habitus and cultural capital, were defined in works by Pierre Bourdieu (see e.g. Bourdieu, 1977). The concept of institutional habitus was developed by Reay et al. (2001) and is defined "as the impact of a cultural group or social class on an individual's behaviour as it is mediated through an organisation" (Reay et al., 2001, para. 1.3). Applied to student dropout, institutional habitus represents the practices, norms and values of a higher education institution. The authors emphasise that "individuals are differentially positioned in relation to the institutional habitus of their school or college according to the extent to which influences of family and peer group are congruent or discordant with those of the institution" (Reay et al., 2001, para. 1.7). This concept was adopted by Thomas (2002) to analyse the relationship between institutional habitus and student retention. Consistency between values, norms and practices of the university and students, seem to be crucial for study success. It is assumed that educationally alienated or non-traditional students obtain a lower amount of cultural capital that is relevant for integration at university and therefore, do have greater assimilation problems (Thomas, 2002).

Psychologically motivated models

In contrast to the dropout models mentioned above, psychologically motivated theories emphasise the role of students psychological characteristics for the decision to persist in the higher education system. Bean and Eaton's psychological model of student retention (Bean and Eaton, 2000, 2001) is based on four psychological theories, i.e. attitude-behaviour theory, coping behavioral theory, self-efficacy theory, and attribution (locus of control) theory, which are combined to build a model of academic and social integration. They describe the dropout process as follows: students enter university endowed with psychological pre-entry characteristics such as self-efficacy and

attributions. These characteristics affect students' interaction behaviour with the institutional environment. Interactions do not automatically integrate students into the environment, but initiate psychological processes such as self-efficacy assessment and coping processes. If successful, these processes may result in positive self-efficacy, reduced stress, increased confidence, and internal attribution and motivation, and also promote academic and social integration. Integration leads to institutional fit and commitment, which positively affect student determination and persistence (Bean and Eaton, 2001).

The psychologically motivated model of student persistence from Ethington (1990) utilises expectancy-value theory, stating that student performance is affected by expectations of study success and their individual valuation of a higher education degree. The former depends on students' academic self-concept (self-assessment of academic skills), whereas the latter is formed by students' economic and social goals (Ethington, 1990).

Economically motivated models

Economic models of student dropout are grounded on theories of rational choice and associated with human capital theory. Important determinants of the dropout decision process are expected returns to education, monetary costs, opportunity cost and expected educational success (Becker and Hecken, 2007). Students will decide to stay at university instead of leaving to start, for instance, vocational training, if the expected benefits exceed the financial (and mental) costs (Hadjar and Becker, 2004). Expected returns to education depend on perceived career prospects and the amount of human capital accumulated during study, which is more efficiently accumulated by students of high ability. Expected success depends on grade performance (Stinebrickner and Stinebrickner, 2008). The evaluation of costs is based on the financial situation and the ability to invest in education. Student expectations could vary during the study process and depend on students' information status. If costs are higher and study performance lower than expected, external opportunities become more attractive and the probability of dropout from university increases (Hadjar and Becker, 2004).

A combined phase model

A highly complex dropout model which combines several theoretical perspectives was developed by the German Centre for Higher Education Research and Science Studies (see e.g. Heublein, 2014a). In line with Tinto, dropout is described as a process which is further divided into three phases, a pre-university phase, a within-university phase, and a decision-making phase. In each phase, different influencing factors become important.

The first phase covers factors representing parental social and educational background, and students' educational background. The authors also point to student preferences and expectations concerning the study programme and the study field, which determine educational decisions and, therefore, the whole study process. The second phase covers all relevant internal and external factors during the course of study, in which internal factors are directly influenceable by the student, and external factors are set by universities. Important determinants are students' mental and physical resources, study motivation, study conditions, capabilities and academic and social integration. According to the model, external factors outside the university environment affect student dropout decisions. These are represented by the financing of studies, living conditions, alternatives to the current study, and advice from parents and friends. These internal and external determinants affect each other and "In a successful study programme it is crucial that internal and external factors are coherent despite constant transformations and developments. This means that students must be able to react appropriately to external conditions in their study behaviour and motivations" (Heublein, 2014a, p. 505). The decision for or against dropout is made in the third phase of the model, in which dropout constitutes a result of incompatibility between internal and external factors (Heublein, 2014a).

In summary, the theories and dropout models described above suggest that many factors from different areas affect student dropout decisions (see also Table 2.2). Some of them reduce the decision process mainly to one aspect, as for example, economically-oriented models which focus on cost benefit considerations. Other theories integrate various different aspects of the dropout process into one model, with several determinants interacting with each other (see. e.g. Tinto's models). These models form the basis for empirical analyses, recommendations and implications for future research. Firstly,

theories reveal the dropout phenomenon to be a very complex process which rarely depends only on one isolated factor, but is rather the result of an inter-relationship of many determinants. Moreover, these bundles of dropout causes seem to be mainly a combination of factors from different areas (for example psychological and institutional factors), rather than covering factors from only one area. This abundance of dropout determinants and especially their inter-relationships have to be taken into account in empirical research and modelling. Therefore, empirical research needs complex models to evaluate the relevance and importance of isolated factors, as well as to assess their complex inter-relationships. Furthermore, theories on dropout reveal relevant factors to be influenceable to a varying degree by the national system, the institutions, the students themselves, or not to be influenceable at all. Hence, it is important in empirical research to evaluate the relative impact of several factors in the dropout process, so as to identify starting points for reducing dropout. Some factors have been investigated more thoroughly than others. This is, among other reasons, mainly because of the specific focus of many studies on for example, psychological or sociological aspects of dropout or due to data availability. A more detailed description of these factors is provided in section 2.3.

2.3 Determinants of student dropout

2.3.1 Selection of empirical studies

This literature review focuses on studies analysing the causal relationships between student dropout from the higher educational system and determinants from different areas of life identified to be relevant from a theoretical perspective. Although there might be some similarities between, for instance, secondary and higher education (e.g. relevance of parental educational background), there are many aspects specific for higher education (e.g. voluntariness of studying). The included studies focus on quantitative empirical research with a clear methodological approach using representative datasets from a survey or administrative data. Furthermore, we tried to cover almost all relevant areas of determinants (e.g. personal, institutional, national, etc.). To allow for comparison, studies included in our review are primarily from European countries and not older than 20 years as there were some greater reforms of the higher education system (e.g.

Bologna reforms for European countries) at the beginning of the 21st century. Exceptions are contributions, which seem to be standard references in the empirical work on student dropout (e.g. older studies or non-European countries). Moreover, the included studies should be published in a (peer-reviewed) scientific journal. The selection process of records is depicted in Figure 2.1. Table 2.3 in the appendix describes the included empirical studies.

In the next sections, we review the different student dropout determinants and classify them according to their level of impact. Referring to Vossensteyn et al. (2015), we distinguish between national system level factors, institutional level factors and individual level factors. Of course, not all determinants could be strictly allocated to one of the three groups. For previous reviews dealing with the state of dropout research see Sarcletti and Müller (2011), Larsen et al. (2013c), and Ulriksen et al. (2010) focusing on STM (science, technology and mathematics) fields.

2.3.2 National system level factors

In this section, we focus on determinants of dropout which are related to the way each national education system is organised.

School system

One important factor refers to the countries' school system and the associated varying pre-tertiary educational pathways of students. Hence, students with different pre-tertiary educational tracks may perform differently at university, and those accessing higher education via non-standard pathways may face difficulties in completing their degree successfully and are possibly more likely to quit tertiary education before graduation. Müller and Schneider (2013), for instance, examined the effect of pre-tertiary educational pathways on dropout from tertiary education in Germany using a sample of 11,649 individuals from the NEPS (National Educational Panel Study). They reveal that students from the upper track in secondary level have a lower dropout rate than students from the lower or intermediate track. A second observation is that students at university with pre-tertiary pathways different from the academic track have higher dropout rates than students who followed the standard path. Moreover, students whose

pre-tertiary path included vocational training have higher dropout rates at universities. Smith and Naylor (2001) analysed a sample of 94,485 undergraduate students in UK, and find that Local Education Agency (LEA) schools tend to promote low dropout probabilities in tertiary education, in contrast to independent schools. Ghignoni (2017) investigated a dataset from Italian universities including over 50,000 students from two cohorts and finds vocational schools to increase the likelihood of dropout in Italy.

Socio-economic inequality

As a second factor, the socio-economic inequality persisting in many countries is assumed to be highly associated with educational inequality or educational disadvantage, which in turn impacts on student dropout. More specifically, if early tracking placements into secondary schooling, which starts in Germany, for instance, already after the basic primary school (four years), is largely based on the performance students achieved during the last year of the primary school, early tracking will be associated with the socio-economic background of students (Krause and Schüller, 2014). Empirical research conducted by Schnepf (2003) on data from the 1995 TIMSS (Trends in International Mathematics and Science Study) and PISA (Programme for International Students Assessment) 2000 and a study by Dustmann (2004) based on the SOEP (German Socio-Economic Panel), confirm the hypotheses of Krause and Schüller (2014). While Schnepf (2003) identifies that children with a lower socio-economic background are considerably disadvantaged in entering the upper secondary school pathway, Dustmann (2004) observes a strong positive correlation between parental background and children secondary school pathways which further affects higher education success. Müller and Schneider (2013) also reveal that students from higher social classes accessed the standard education pathway more often than students with a lower social background. Another illustration of the effects of the socio-economic inequality on academic performance at university is mentioned by Hansen and Mastekaasa (2006), whose analysis of first-year students in Norwegian universities reveal that students who gain higher grades at universities, are those with higher levels of cultural capital.

Geographic origin

Student performance and dropout rates are also influenced by the geographic origin of students. Glaesser (2006) conducted a study in Germany analysing the pathways from late childhood to early adulthood of about 1,500 participants and observes that university students with an urban origin were more than three times at risk of dropping out than students from rural regions. Using data on 1,158 students from the ECHP (European Community Household Panel), Aina (2013) points out that the geographic area plays an important role. Students from the economically more powerful northern regions of Italy have a higher probability of enrolling in a tertiary education programme. Di Pietro (2006) comes to a similar conclusion. He investigated data from 5,907 Italian students, and states as a central result that poor regional labour market prospects decrease the dropout rate significantly.

Financing policy

Another central determinant of university dropout at the national level is a country's financing policy. According to the report of Vossensteyn et al. (2015) on behalf of the European Commission, there are different kinds of financial support students can benefit from: public or private scholarships, grants and loans, and support for tuition or registration fees for students from low-income families. This financial support generally depends on the parental income or on student performance, and is provided to enable students to concentrate more on their studies despite spending much time on paid work. An empirical study by Glocker (2011), based on data from the SOEP and including 787 individuals, reveals that the amount of support students receive decreases the dropout rate on average by 2.6% per 1,000 EUR per semester. Moreover, a rise in financial support by 200 EUR per month reduce the dropout risk by up to one third indicating that increasing of financial support in a country may result in higher probabilities to graduate.

Reforms of the higher education system

A further interesting aspect affecting the student dropout phenomenon has been observed since the introduction of the Bologna Process. The Process was signed in 1999 by Education Ministers from 29 European countries at the University of Bologna and

nowadays has 47 participating countries. Its main goal has been to create a standardised European Higher-Education Area, in which tertiary education degrees in Europe would be comparable, mobility for students and teachers at international level will be supported, and student grades and exams recognised across the member countries (Reinalda and Kulesza-Mietkowski, 2005). The key aspect of the Bologna Process is the introduction of a two-tier degree system based on an undergraduate cycle (Bachelor) and a graduate cycle (Master). Another innovation is the European Credit Transfer System (ECTS), a new scoring system for examinations. A potential impact of the Bologna reform is that it could enable students, unwilling to study four or five years, to still manage to obtain a degree (i.e. three years to earn a Bachelor degree) rather than dropping out. Di Pietro and Cutillo (2008) examined the differences in the dropout probability at Italian universities for students enrolled in 2001 after the education reform, and students enrolled in 1995 and 1998. They observe that the behaviour of the post-reform students slightly reduces the probability of dropout. However, Horstschräer and Sprietsma (2015) investigating the effects of the introduction of Bachelor programmes on college enrolment and dropout rates using administrative data on all German tertiary-education students for all academic terms (1998-2008), observe no significant change in the number of first year students or in dropout rates in general.

These factors do not seem to be "stand-alone" predictors, but rather interact with other national factors, and those from more institutional and individual levels. For instance, geographical origin and the impact of financing policies probably have different effects on study success, depending on the socio-economic status of a student's parents. Moreover, the impact of the chosen pre-tertiary educational pathway may be more or less relevant, according to a student's (and parent's) attitudes, aspirations and motivation. Therefore, in further research, it seems important to take a closer look at interdependencies between national system determinants and other relevant factors.

2.3.3 Institutional level factors

The functioning and study conditions of higher education institutions, i.e. the way the teaching is organised or the equipment universities put in place for education impact students' success and further influence the reputation of the institution. In this section, we present university related factors and differences between study programmes.

Important factors are, for instance, the type of higher education, teaching quality, the relationship between teachers or tutors and students, the organisation and preparation of exams, learning environment, counselling services and especially the subject of study.

Types of higher education institutions

A difference in dropout rates is observed between types of higher education institutions, for instance, between public and private ones. Sarcletti and Müller (2011) reveal that the dropout rate in private institutions is higher than in public institutions. Different dropout rates are also observed for different kinds of public higher education institutions. In Germany, for example, there is a distinction between universities and universities of applied science. While universities are more theory and research oriented, universities of applied science focus on practical applications, offering more structured study programmes and tend to tune students towards industry needs (Mayer et al., 2007). As mentioned by Sarcletti and Müller (2011), the dropout rates in Bachelor programmes from universities of applied science are significantly lower than those in university Bachelor programmes. The same observations are also made by Heublein et al. (2017), analysing a cohort of 6,029 German exmatriculated students from summer term 2014.

Fields of study

In addition, significant differences in dropout rates between subjects and study fields have been observed. According to Heublein et al. (2017), the highest dropout rates are found in the subjects of Engineering, Mathematics and Natural Sciences. Very few students dropout from Arts, followed by Law, Economics and Social Sciences. Several international studies provide partly similar results, but differ in some details. The investigations by Smith and Naylor (2001) of UK universities reveal that, measured by the withdrawal rate, students perform well in Biology, Literature, Classic Sciences, Humanities and Creative Arts. The performance is worse in Mathematics, Computing, Education and Languages. For Spain, Lassibille and Navarro Gómez (2008) analysed 7,000 students from the University of Malaga and observe high dropout-rates especially in subjects like Engineering, Science and Law. Korhonen and Rautopuro (2018) used data on more than

20,000 students from four Finnish universities and find the highest dropout rates in the fields of Information Sciences, Information Technology, Mathematics and Economics. There are large between-subject differences in the time students spend weekly on studying. Students in Technical and Natural Sciences spend, with 37.2 hours per week, much more time on study than students of Economics (29 hours) and Law (24.3 hours) (Brandstfätter and Farthofer, 2003). The more time students spend studying, the better their study performance and their lower the dropout risk.

Study conditions and environment

Study conditions and study environments of higher education institutions also have a great impact on students' performance and dropout decision. These factors enable students to take the courses in a positive atmosphere. As hypothesised by many researchers including Schröder-Gronostay (1999), there are several determinants on the institutional side, such as low teaching quality, lack of transparency, poor quality counseling, teaching staff with low pedagogical ability, and high achievement requirements, which may negatively affect student performance and increase their risk of dropping out. Georg (2009), using data from the Konstanz Student Survey in Germany including about 10,000 students, reveals that within institutional factors only teaching quality is relevant in explaining the dropout phenomenon. More precisely, he analysed the impact of teaching quality on the relationship between student social origin and dropout rate, and observes that improving teaching quality could reduce the social inequality at universities and therefore decrease dropout rate. Hovdhaugen and Aamodt (2009) come to a similar result for Norwegian students in humanities and social sciences. They explain that beside teaching quality, learning environment also has a significant effect. However, they emphasise that many dropout reasons are not in the hands of the institution, as most students dropout or transfer to other institutions for external reasons, among which "start a new programme" or "be employed" are the most common ones. Suhre et al. (2007), analysing 186 first year law students at the University of Groningen (Netherlands) and Ghignoni (2017) for Italy highlight the importance of the relationship between students and tutors or teachers.

Johnes and McNabb (2004) examined a large cross-sectional dataset of about 100,000 English and Welsh students. Their findings show that a good assessment of teaching

quality, high staff-student ratio, high library expenditure and a large number of undergraduates at the institution, all increase the probability of finishing the degree. A surprising result of their study is that a high staff-student ratio reduces the probability of dropping out voluntarily, but increases the likelihood of dropping out due to academic failure. Furthermore, according to Heublein et al. (2011), study conditions like study requirements, study organisation, study structure and teaching quality are assessed more negatively by students who withdraw from university.

Class size seems also affects student dropout behaviour. Montmarquette et al. (2001) examined a panel study of 3,418 students from 43 different programmes at the university of Montreal. They find that the optimal class size for student persistence is between 60 and 110 students. In smaller courses, there is probably not enough money for a teaching assistant or extra tutorials. In larger courses, there may be rather no real interaction between students and teachers.

As already mentioned at the end of Section 2.3.2, it is rarely only one isolated aspect which increases or reduces student dropout, but more likely an interrelationship of different factors. For instance, one could assume that there is a correlation between type of subject, time students spend on studying and their own motivation, which jointly affect the probability of dropping out. Moreover, the impact of teaching and counselling quality or the pedagogical ability of lecturers may be different for students from academic, compared to non-academic backgrounds. These are only some examples of possible interdependencies between institutional and individual factors, but they seem to be very complex and therefore require complex analytical methods.

2.3.4 Individual level factors

The following section addresses dropout factors related to students themselves. The decision to leave the university without obtaining a degree is driven mainly by student personality and academic self-concept (individual level) and less by external factors (institutional side, national system level). This theory is supported by Georg (2009), who examined the relationship between individual and institutional factors in influencing the dropout phenomenon. He discovered that only 5% of information explaining dropout was found at the institutional level, whereas 95% was associated with the individual level.

We distinguish here between two types of individual dropout factors: pre-study factors and study related factors.

Pre-study factors

Pre-study factors are those specifying the starting conditions of students, i.e. before they enter higher education. While universities, to a certain degree, can control some of those factors of the student body as a whole through admission rules, affecting pre-study factors for individual students are generally not in the scope of universities. Pre-study factors can be categorised into different groups: demographic factors (gender, age), prior education factors (grade point average at secondary school), and socio-economic factors (social background and parental education, migration background). Note that the type of the secondary school visited is also an aspect of prior education, but has already been examined in Section 2.3.2.

Gender

We first focus on the impact of gender on the dropout phenomenon. Aina (2013) and Ghignoni (2017), both analysing Italian universities, find that the likelihood of withdrawing from university is significantly lower for female students. Glaesser (2006), examining German students, observes that women are more than twice as likely as men to dropout of vocational training, whereas men are more than twice as likely as women to withdraw from a university programme. Van Bragt et al. (2011b) conducted a study on 1,176 students from the Netherlands, which shows that besides the lower enrollment rate for male students, the dropout rate for males is three percentage points higher than for females. Smith and Naylor (2001), for UK students, observe that only three percent of the total gender gap can be explained by observed characteristics, for example because men prefer subjects with higher dropout rates. As stated in Sarcletti and Müller (2011), dropout rates for men and women depend partly on the gender composition of a course. If there is gender disparity, members of the minority class are more likely to face integration difficulties. However, Mastekaasa and Smeby (2008) analysing 2,422 students from five Norwegian universities, come to the conclusion that male students' dropout is unrelated to the gender composition of study programmes, while women dropout from female-dominated study courses to a lesser extent.

Severiens and Ten Dam (2012) investigated gender differences at Dutch universities including a sample of 10,000 university leavers. They distinguished between male- and female- dominated study programmes, where at least 75 % are either males or females. Here, men have a very high attrition rate in female-dominated study programmes. According to this study, there are four reasons why men dropout more often from female-dominated study programmes, compared to male-dominated studies. First reason is the home situation: men receive no support from parents and friends, which are often negative about their study choice. Secondly, female-dominated programmes offer more often poorer job opportunities, which leads to a lower salary and lower status of the future job. Thirdly, men receive no support from peers. Fourthly, men often dropout without a tertiary degree, since they find a job outside university. The reason why women dropout of male-dominated programmes is mainly due to poor study choices, lack of motivation and uninteresting courses. Compared to men, women seem to be more motivated, disciplined and have better time management skills, which are important characteristics for study performance.

Brandstätter et al. (2006) also analysed the interaction between gender and subject fields among 948 high school graduates who had participated in a career counselling programme in Austria. They observe that the dropout rate in Technical and Natural Sciences is higher for women and lower for men, compared to other subjects, which is in line with Severiens and Ten Dam (2012). Though, Brandstätter et al. (2006) find no influence of gender on the overall dropout rate. In contrast to the previous results, a study of Belloc et al. (2010), who examined 9,725 Bachelor students of Economics at the Sapienza University in Rome/Italy, reveals that the probability of dropout is lower for male students. This can be explained by the fact that only Economics students at only one university were considered.

Age

Regarding the age of students, there is an evidence that older students are more prone to dropout. This is in line with the findings of Müller and Schneider (2013). In their study, they observe that older students are more likely to drop out. This may also explain the higher dropout rate for students who obtained vocational training before entering higher education. Lassibille and Navarro Gómez (2008) and Montmarquette et al. (2001) obtained similar results. A possible reason is the higher opportunity costs

for older students who already have vocational experience. In contrast to these findings, Smith and Naylor (2001) find that women perform better with increasing age, while the best performing group of male students is between 28 and 33 years. Belloc et al. (2010) observe that the higher the time span between secondary school and university, the lower the dropout probability.

Parental background

The positive impact of high educational levels of the parents on the children's educational results and job careers has been thoroughly investigated in research. Smith and Naylor (2001), Di Pietro and Cutillo (2008), as well as Aina (2013), reveal that the better the parental education, the better the students' performance at university, and the lower the probability of dropout. Furthermore, Aina (2013) points out that highly educated parents have a positive effect on the enrollment rate of students. Other studies come to similar results. For instance, Ghignoni (2017) concludes that a lower social class, and a father without a tertiary degree, increase the dropout probability. Johnes and McNabb (2004) investigated the effect of parental occupation and state that unskilled parents increase students' dropout risk. Similarly, Gury (2011), who examined 5,383 students enrolled at universities in France, finds that students whose fathers have blue-collar jobs are more likely to drop out during the first three years. The field of parental education is strongly related to the parental occupation and therefore, with the family income and the financial support of a student by his/her parents.

Migration background

The effect of migration background on university dropout seems to depend strongly on the secondary education and the financial aid system of a country. The studies of Belloc et al. (2010) and Johnes (1990), conducted at the university of Rome and the Lancaster university respectively, mention a higher dropout probability for students from a foreign country. Reisel and Brekke (2009) compared minority students from Norway and the USA. Black students and Latinos in the USA and non-western second-generation immigrants in Norway are defined as minority students in the respective country. In both countries, the parental income is higher for majority students compared to minority ones. Minority students in Norway do not have a higher withdrawal probability, as

the Norwegian social democratic welfare state reduces the disadvantages of minority students. In the USA, Latinos and black students have a significantly lower probability of graduating than white students. As there exists high tuition fees at US colleges and financial aid depends on the parental income of a student, family income has a high and positive influence on the likelihood of graduating. Furthermore, students with a migration background face the problem of less cultural and social capital, and are mainly unfamiliar with the study culture and language of the country in which they intend to study. Migrant students are often less familiar with the study structure, and are not well equipped concerning self-organisation and self-assessment. In addition, Sarcletti and Müller (2011) find that migrant students are more likely to have a poorer background, have less knowledge about the education system and the prevailing culture, and are less familiar with the language. These various aspects make the migrant student more vulnerable to any difficulty occurring at the university, thereby increasing the risk of withdrawal.

School performance

A further important pre-study factor affecting student dropout is the prior education of students, especially the student' high-school grade-point average (GPA). Sarcletti and Müller (2011) claim this factor to be a particularly important indicator of the student ability to meet the level of performance required by the higher education system, which could also serve as a predictor of future dropout risk. Various international studies, for example Stinebrickner and Stinebrickner (2014) analysing 341 students who entered Berea College (USA) in 2000/2001, as well as Johnes (1990) and Di Pietro and Cutillo (2008), observe a positive correlation between GPA and study performance.

Personal characteristics

Personal characteristics of a student are very important for educational performance. Van Bragt et al. (2011a) examined the effect of students' personal characteristics on the number of credit points and the probability of dropout for 1,471 students from a university of applied sciences in the Netherlands. They find that both the number of credit points and the probability of dropout, can be well predicted by personal characteristics.

"Conscientiousness" has a significant positive effect on the number of credit points as well as on study continuance. Unlike, students with high scores on "ambivalence and lack of regulation" are likely to obtained fewer credits and dropout easily. Brandstätter et al. (2006) show that resilience and self-control have a positive effect on study persistence. One surprising result of their study is that students who are uninformed, unsure and afraid with their study choice, do still perform well. A possible reason for this unexpected result might be that these students are afraid of the transition to the labour market and as a consequence, work harder at university. According to Van Bragt et al. (2011b), successful students mainly attribute their success to their own skills. Students who fail and drop out from university usually attribute their failure to external factors.

Mäkinen et al. (2004) divided 1,600 students from a multi-disciplinary Finnish university into three groups. The first are "study-orientated students", who are very interested in the subject, learn intensively and appreciate student life and social relationships. "Work-life orientated students" are interested in this balance, learn a lot for university, and plan their studies systematically. Social relationships are less important for this group. The third group are "non-committed students", who have no study-related goals and few social relationships at university. Relatively, the last group of students change their subjects most often and have the fewest credit points and lowest grade point average (GPA). Moreover, this group has the highest risk for both abandoning and prolonging their studies.

Similar to the previous sections, we observe some relevant interrelationships between the above mentioned determinants. For instance, the gender effect seems to vary with the group composition of study fields, and migration effects depend on national financing and the secondary school system. Therefore, it is essential to account for such interdependencies in empirical research.

Study related factors

In contrast to the pre-study factors addressed in the previous section, the factors examined here are mainly in the hands of students. These factors include learning motivation and self-confidence, study organisation, learning strategies, social integration at

university, study conditions, effort devoted to studying, intrinsic and extrinsic motivation or preparation for exams. Additionally, some studies analyse the effect of off-study work and the financial situation of students.

Self-confidence

Self-confidence denotes the confidence or assurance in one's own personal judgement, ability, power or capability. Self-confident students set higher goals for their study and are prepared to make greater efforts whenever obstacles arise (Brandstätter et al., 2006). Whereas students with low self-confidence tend to have less faith in their own intellectual abilities and give up soon, whenever difficulty occurs. According to Schiefele et al. (2007), analysing 47 dropout students and matched regular students of the university of Bielefeld (Germany), this negatively impacts on study and learning motivation and promotes the risk of dropping out. Heublein et al. (2017) support this observation and state that a sufficiently strong study motivation is a fundamental prerequisite for successful graduation.

Students study organisation

Relating to the study organisation, researchers point out that poor study organisation and an inadequate learning strategy could negatively influence study success. Schiefele et al. (2007) observe that students with poor organisation and whose learning strategies do not suit their study fields, probably start struggling to perform well and justify this low performance by the fact that the study content is too abstract and that they are overwhelmed. Further, they find a correlation between dropout students and students denigrating the teaching quality. Heublein et al. (2017) observe that around two-fifths of the dropout students quit the university because the programme organisation did not match their expectations.

Off-study work

At first glance, one could assume that off-study work has negative influence on study performance and prolongs time to degree (Behr and Theune, 2016), since there is less time available to spend on exam preparation, tutorial attendance and other study-related

work. However, Hovdhaugen (2015) points out that in Norway, off-study work interferes more with students' free time than with students' study time. One hour of working reduces study time only by about five to ten minutes. He analysed 12,726 students and distinguished between three groups of employed students. The "short part-time" group works for a maximum of 19 hours a week, the "long part-time" group works between 20 and 30 hours and the "full-time" group works more than 30 hours a week. The latter may be forced to work, because they need to finance study and living costs, or they may voluntarily choose to work, intending to improve their employability or to fund a higher living standard. Part-time work seems to have no significant influence, while long part-time and full-time work increases the attrition rate significantly. Similar results are reported by Beerkens et al. (2011) for 2,496 students from higher education institutions in Estonia. They observe that more than 25 hours of off-study work decreases the probability of graduating in regular time. Brandstfätter and Farthofer (2003) observe that the number of exams per semester, study satisfaction and the grade point average decreases with the weekly hours of work, whereas the dropout-rate increases. In this study, even "short part-time" work is found to have negative effects. Moreover, off-study work impacts on female students more strongly than male students. A very recent review on this topic is conducted by Neyt et al. (2019) and reports a mainly negative effect of student work on continuing studies.

Satisfaction and person environment fit

Suhre et al. (2007) investigated the relationship of degree programme satisfaction and dropout probability, as well as academic performance, measured by the number of credit points. Unsatisfied students usually spend less time in study exercises and have a higher chance of withdrawal. They find that "degree programme satisfaction", "study motivation", "regular study behaviour" and "attending tutorials" are strongly positively correlated with the number of credit points, and negatively correlated with the dropout probability. Suhlmann et al. (2018) observe the person environment fit between the higher education institution and personal attitude to be strongly related to students' satisfaction. Their findings are based on the analysis of 367 undergraduate students from a German university. If students feel, that they belong to the university, they are more motivated, more satisfied and have a lower chance to drop out. Nordmann et al. (2019) find that class attendance and the use of recorded lectures at the University of Aberdeen in UK has a positive influence on study performance for the

analysed 347 first-year students. Korhonen and Rautopuro (2018) observe that at-risk students spend less time in the study course and are often precarious about their choice of study.

According to Stinebrickner and Stinebrickner (2014), the perceived utility of a tertiary education programme affects student decisions to enter college and to continue the degree programme. Based on their own expected abilities, students decide to stay at university if college is more enjoyable than a job. Moreover, grade point average and performance in the current semester are significant predictors for continuing a study programme.

Students at risk of dropping out seem to accumulate several negative working characteristics, culminating in the decision to leave university without a degree. For instance, there are correlations between off-study work, study satisfaction and performance, which in turn affect the dropout decision. Furthermore, whether negative performance at university leads directly to dropout seems to depend on students' attribution of failure (external or internal). As there is a variety of possible inter-dependencies between these analysed factors, considering these complex inter-dependencies adequately in empirical studies is challenging. In the next section, we discuss some research gaps and implications of this review for further research on student dropout from higher education.

2.4 Implications for future empirical research

2.4.1 Research gaps

Previous theoretical and empirical research from a wide range of disciplines has identified a number of possible reasons for withdrawing from tertiary education. This literature review reveals that determinants before and right at the beginning of study (like secondary education, field choice motives) and "softer" attitudinal based factors (like study satisfaction, social integration) have a strong impact on dropping out.

Administrative data that is used in many studies (e.g. Belloc et al., 2010, Hovdhaugen, 2015), lack information on these "softer" factors (Larsen et al., 2013c) and also information on pre-study factors is limited. Moreover, Singell and Waddell (2010) emphasise the importance of both fixed and time-varying effects for the dropout process, and Gury

(2011) states that some factors do indeed have a constant effect on withdrawal, but other effects (e.g. study conditions) vary over time.

There are just a few recent studies investigating large survey datasets with such a wide range of variables. Many of the mentioned studies are based on small data sets and restrict their analysis to specific academic fields and/or to one university (e.g. Lassibille and Navarro Gómez, 2008, Stinebrickner and Stinebrickner, 2014). They mostly do not consider all the possible determinants relevant for student dropout and often emphasise "hard" university non-malleable factors, like the social background or gender, but research would in fact benefit from dealing more with study-related and "softer" university malleable factors, as these are mainly within the scope of policy action (Larsen et al., 2013c).

Usually, previous studies investigated student dropout using standard econometric regression models like logit- or probit-regression (e.g. Van Bragt et al., 2011a, Beerkens et al., 2011), or methods from survival analysis (e.g. Lassibille and Navarro Gómez, 2008, Aina, 2013, Hovdhaugen, 2015). In the dropout context, it is very important to obtain results which are transferable to other cohorts of students and do not only identify dataset-specific relationships. The two main reasons for a prediction bias are overfitting and changes in general patterns over time. If the results of empirical analysis are to serve as a basis for installing dropout prevention programmes in universities, it is of the utmost importance to avert overfitting. From this prediction-oriented view, most empirical analysis discussed in this review may be prone to overfitting as standard econometric techniques are not equipped with integrated adjusting strategies.

As the dropout phenomenon seems to be the result of a long process including many interacting factors, empirical research needs models to assess their complex relationships, being able to deal with high dimensional data, high-order interactions and correlations as well as to evaluate the relevance of isolated factors.

Furthermore, we observed that specific groups of students (e.g. males/ females, specific study field) are differently affected by determinants and may, therefore, be responsive to a different degree to specifically implemented dropout prevention programmes. Here, specific methods to find similar groups of at-risk students and to implement more individual supporting strategies are of great importance.

2.4.2 Implications for data

As a consequence of the research gaps mentioned above and also stated in Sarcletti and Müller (2011), large prospective and longitudinal survey data are of considerable importance for assessing the dropout phenomenon in its entirety. Data sets should include determinants before and at the beginning of study, as well as "softer" attitudinal based and university malleable factors. Furthermore, empirical studies may benefit from very recent data sets covering a broad time span to account for the long process of dropping out and for possible time-varying effects of determinants. As various types of dropouts are driven by different motives and causes, which has so far not been adequately addressed, the data should allow to distinguish the dropout phenomenon according to student motives for dropping out and the degree of voluntariness.

For Germany, as an example, the National Educational Panel Study (NEPS) provides very interesting and applicable data for further investigations. The NEPS is a comprehensive German panel study containing six cohorts from all stages of life, whereby the focus lies on the fifth cohort, including students in tertiary education and transition to the labor market. This cohort includes more than 18,000 first-year students, and many variables covering a wide range of possible dropout determinants from different areas (Blossfeld et al., 2011). At the moment, the NEPS contains 11 waves, and is therefore very suitable for analysing dropout processes. But of course, comparable data sets for a large number of countries would be preferable, as there might well be country-specific dropout processes at work.

2.4.3 Implications for methodological approaches

As mentioned above, there is a variety of possible interdependencies between factors, and it is impossible to identify them a priori. Therefore, it would be empirically meaningful to apply methods that are able to search for these interdependencies and patterns in the dropout process without restrictions. Here, data mining techniques are of considerable interest as they are very suitable in the presence of high dimensional and correlated data and may outperform classical (linear) models (James et al., 2013). Rodriguez-Muñiz et al. (2019) emphasise that data mining methods are useful to combine determinants of various areas, e.g. personal features, academic and non-academic features, to a single rule to predict dropout, programme change or study continuation. Educational data

mining is a research field with rising importance, and so far, mining techniques have mainly been applied to educational data in an e-learning context (e.g. Yukselturk et al., 2014).

Data mining procedures address the problem of overfitting explicitly by cross-validation techniques, bootstrapping or sampling methods as well as data file splitting into training and test data (Hastie et al., 2009). One problem might be that, for instance, student behaviour, national education policy, job prospects or teaching quality at universities probably change over time and thus affect the dropout rate of students. The best way to solve this problem is to fit a new model regularly with current data. If new data is not available, the magnitude of this problem can be determined by testing the model with the most current data; the remaining data can be used as training dataset.

One important aim of future educational data mining would be to develop statistical models for predicting future dropouts as precisely as possible and to identify at-risk students as soon as possible. Prevalent data mining methods are, for instance, decision trees and random forests. More recent but less transparent methods are, for example, support vector machines and artificial neuronal networks. For a review on data mining in education, see e.g. Romero and Ventura (2010). There are only a few studies applying data mining techniques for dropout prediction, for example Rodriguez-Muñiz et al. (2019), Aulck et al. (2016), Siri (2015), Jadrić et al. (2010), Dekker et al. (2009) and Vandamme et al. (2007). But these studies are mainly based on small data sets for one university, or even one field of study. But especially large longitudinal data sets covering a wide range of variables (as mentioned above) are very well suited for data mining techniques.

Moreover, identifying different groups of students at risk and, based on this, implementing individual- or group-specific prevention measures is of considerable relevance for decreasing dropout rates. Students with different motives for withdrawing from university are very probably also responsive to a different degree to specific implemented dropout prevention measures (similar to market segmentation for advertising). Students who are at risk due to academic failure, potentially need more help with teaching material, for instance through offering extra tutorials. Other groups of students at risk probably need more assistance in choosing appropriate classes or even in finding the right study field, so as to enhance the likelihood of successful graduation. Here, different cluster techniques seem to be very suitable to identify these specific groups of at-risk

students. Furthermore, there seem to exist complex interactions between predictors and some negative working factors might occur together or consecutively in a student's dropout decision process. Association rule learning is conducive to detecting these jointly occurring factors (similar to a market basket analysis) and therefore facilitates weakening mutually reinforcing factors at an early stage. Moreover, previous research reveals relevant factors to be influenceable to a varying degree by the national system, the institution, students themselves or not to be influenceable at all. Hence, it is important in further empirical research to evaluate the relative impact of several factors in the dropout process, so as to identify starting points for efficiently reducing dropout rates. Here, data mining techniques as, for instance, random forests provide suitable and easy interpretable variable importance rankings.

These are only some examples of how data mining techniques could be applied in the student dropout context and how these techniques, if applied to a data set covering a broad range of determinants, could provide new insights into the dropout process and support decision makers in decreasing dropout rates in higher education. The results of an encompassing analysis of the dropout process in the form of prediction models and analyses of factor interrelationships are a helpful tool for universities wishing to implement early warning systems and to prevent study dropouts, by supporting students at risk at an early stage of study. Based on the findings in this review, one could assume that pre-study courses would prepare students for academic courses, as well as subject-specific consultancy which helps students to obtain clarity on requirements, challenges and possible career prospects, or mentoring programmes to promote social and academic integration.

2.5 Conclusion

This study provides a comprehensive overview on the dropout phenomenon across different, mainly European, countries. It includes a discussion of theoretical models, a structured and detailed overview of different determinants and the current state of research related to higher education dropout as well as a discussion of implications for future research.

First, some basic theoretical models of the dropout phenomenon in higher education were described. Sociologically-oriented concepts highlight the academic and social integration

at university (e.g. Tinto, 1975, Bean, 1985), whereas psychological models concentrate on the role of students' attributes and behaviour in the dropout process (e.g. Ethington, 1990, Bean and Eaton, 2001). Economically motivated models focus on concepts of rational decisions and cost-benefit considerations (e.g. Becker and Hecken, 2007, Hadjar and Becker, 2004). A highly complex dropout model that combines several theoretical perspectives was developed by the German Centre for Higher Education Research and Science Studies (e.g. Heublein, 2014a).

According to these theories and the findings of empirical research, the dropout phenomenon is highly complex and withdrawal from university without a degree is rarely the result of short-term or spontaneous decisions, but rather of a long decision-making process. There are several factors and interrelationships promoting a student's dropout, which cover a variety of determinants like pre-study characteristic, psychological, sociological, as well as economical or institutional aspects (Heublein, 2014a, Heublein et al., 2017). According to the level at which dropout determinants exert their influence and to what extent they are malleable, we categorise these factors into three groups: factors associated with the national education system, the institution of tertiary education and individual student factors (Vossensteyn et al., 2015).

According to the national education system level, the institutional arrangement of secondary education in association with socio-economic inequality (e.g. Smith and Naylor, 2001, Müller and Schneider, 2013), the geographical origin (e.g. Aina, 2013, Glaesser, 2006, Di Pietro, 2006), financing policy in the form of financial support (e.g. Glocker, 2011) as well as higher education reforms (e.g. Di Pietro and Cutillo, 2008, Horstschräer and Sprietsma, 2015) are important predictors of university dropout.

On the institutional level, the type of higher education institution (e.g. Heublein et al., 2017), the study field (e.g. Lassibille and Navarro Gómez, 2008, Korhonen and Rautopuro, 2018), teaching quality and learning environment (e.g. Georg, 2009, Hovdhaugen and Aamodt, 2009), class size (e.g. Montmarquette et al., 2001), and the relationship between students and teachers (e.g. Johnes and McNabb, 2004, Ghignoni, 2017) seem to have an impact on the probability of withdrawal.

Moreover, individual pre-study factors have a strong influence on study performance and dropout. For instance, gender (e.g. Van Bragt et al., 2011b, Ghignoni, 2017), especially in relation to the study field (e.g. Mastekaasa and Smeby, 2008, Severiens and Ten Dam, 2012), age (e.g. Lassibille and Navarro Gómez, 2008, Müller and Schneider, 2013), the

migration background (e.g. Johnes, 1990, Reisel and Brekke, 2009, Belloc et al., 2010), the grade point average at secondary school (e.g. Di Pietro and Cutillo, 2008, Stinebrickner and Stinebrickner, 2014), as well as the parental educational background and status (e.g. Hansen and Mastekaasa, 2006, Gury, 2011, Aina, 2013, Ghignoni, 2017). Moreover, personal characteristics as conscientiousness (e.g. Van Bragt et al., 2011b,a), resilience and self-control (e.g. Brandstätter et al., 2006) or commitment (e.g. Mäkinen et al., 2004) play an important role.

Beside these pre-study determinants, several study-related individual aspects affect students' risk of dropping out. These are, for instance, study motivation (e.g. Heublein et al., 2017), especially intrinsic motivation, and study satisfaction (e.g. Suhre et al., 2007), learning strategy and students' study organisation (e.g. Schiefele et al., 2007), class attendance (e.g. Korhonen and Rautopuro, 2018, Nordmann et al., 2019), as well as off-study work (e.g. Brandstfätter and Farthofer, 2003, Beerkens et al., 2011, Hovdhaugen, 2015).

Based on the findings from this review, we discussed the implications for further research. Especially the application of modern data mining techniques on a comprehensive data set covering many aspects of student life seem useful for providing new insights in the dropout process (e.g. Siri, 2015, Rodriguez-Muñiz et al., 2019). Developing a precise prediction model of student dropout should form the focus of further empirical research. Furthermore, other mining techniques such as cluster analysis to identify groups of students with similar dropout reasons, or association analysis for detecting jointly occurring dropout determinants, may reveal detailed relationships within the dropout process. The results may provide a helpful tool for universities wishing to implement early warning systems and promising individual or group-specific supporting measures for students at risk, so as to prevent dropouts at an early stage of the study process.

2.6 Appendix

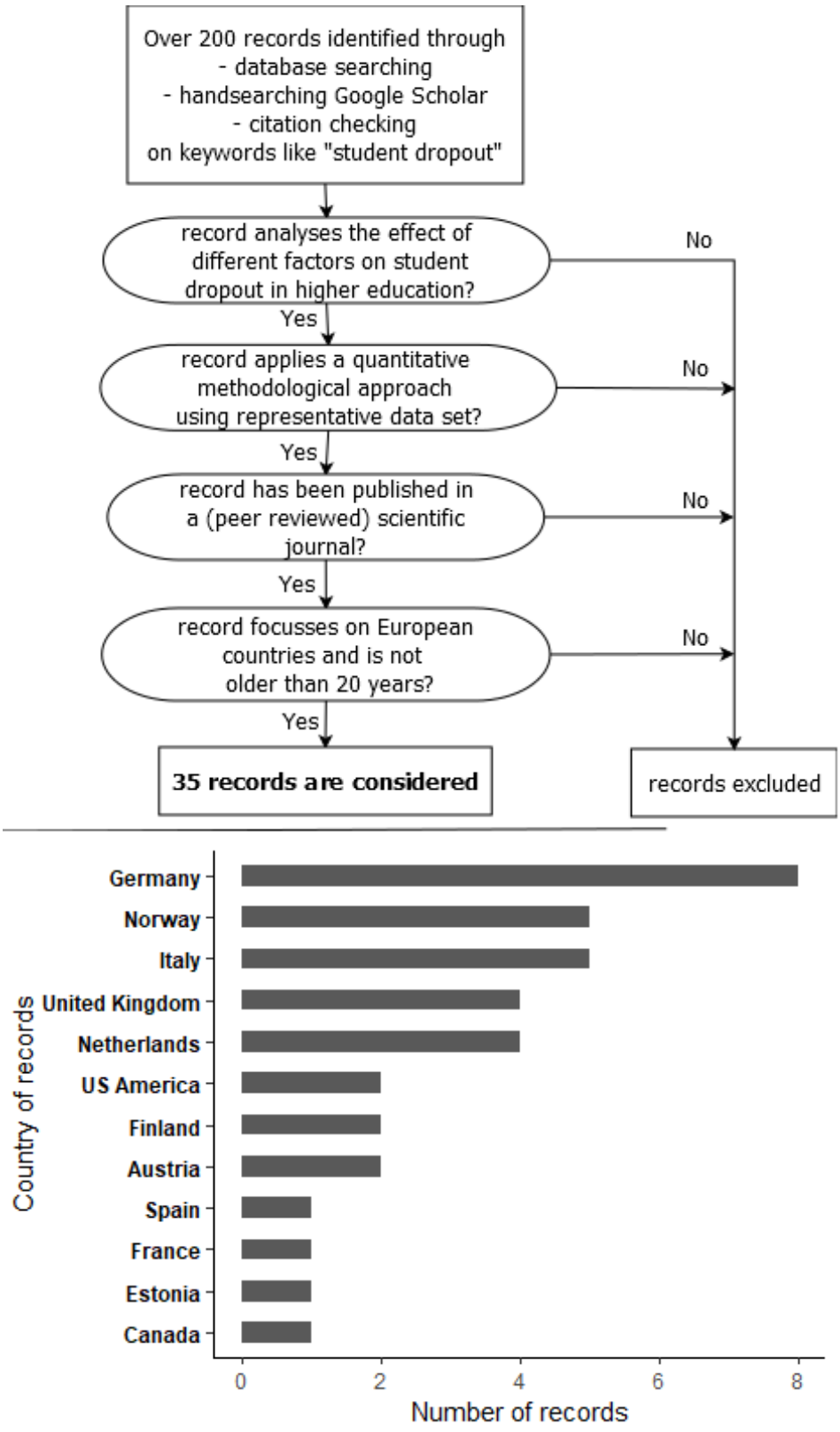

Figure 2.1: Selection of journal articles

Table 2.3: Description of the included empirical records

Nordmann et al. (2019): Turn up, tune in, don't drop out: the relationship between lecture attendance, use of lecture recordings, and achievement at different levels of study	**United Kingdom**: 347 University of Aberdeen students from all four years of the undergraduate degree from the first semester of the academic year 2015/2016	**sociological perspective**: attendance has a strong relationship with final course grade and is a better predictor of academic performance **method**: regression models	demographic data: age, gender, nationality, native English speaker; class attendance data, recording data, final grade for each course's exam, grade point average	class attendance, use of recorded lectures are positive predictors of performance for first-year and slightly for second-year students; no relationship is found for honours-year students
Korhonen & Rautopuro (2018): Identifying problematic study progression and "at-risk" students in higher education in Finland	**Finland**: 20,159 slowly progressing (or non-studying) students of four Finnish universities surveyed from 2009 to 2012	**sociological/ psychological**: reasons inside or outside the institution and personal reasons affect non-completion **method**: logistic regression	age, gender, study field, socio-economic background, life situation, health and well-being resources, study experiences, future expectations	Information Technology, Ma- thematics and natural sciences have a high risk of dropout; first-year students are at the greatest risk of non-completion
Suhlmann et al. (2018): Belonging mediates effects of student-university fit on well-being, motivation, and dropout intention	**Germany**: 367 undergraduate students from different study fields from a German university	**sociological/ psychological**: higher person-environment fit should prevent students from forming dropout intentions **method**: regression analyses	gender, study field, dignity self-construal, perceived university norms, sense of belonging to the university, academic motivation, well-being	students' sense of belonging to the university is related to increased well-being, increased academic motivation and lower intention to drop out

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Ghignoni (2017): Family background and university dropouts during the crisis: the case of Italy	**Italy**: 25,880 students from Italian universities from the cohort of 2007 and 26,588 students from the cohort of 2011	**economic perspective**: financial constraints have a strong impact on dropout rates **method**: probit model	family background, educational background, tuition fees, relationship between students and tutors	strong family economic background and other individual characteristics increases students' probability of succeeding
Heublein et al. (2017): Zwischen Studienerwartungen und Studienwirklichkeiteit	**Germany**: 6,029 exmatriculated students from summer term 2014	**phase model**: university dropout is a multi-dimensional process affected by several determinants **method**: descriptive analysis	pre-uni: educational background; transition into uni: choice of subject; uni phase: academic achievements, study conditions, counseling services	3 motives are largely responsible for dropping out of German unis: high study requirements, lack of study motivation, orientation to practical work
Horstschräer and Sprietsma (2015): The effects of the introduction of Bachelor degrees on college enrollment and dropout rates	**Germany**: administrative data on all German higher-education students of all academic terms from 1998 to 2008	**sociological perspective**: reduction in the time for graduation could encourage students to invest in higher education **method**: fixed effects model	individual study situation (year, subject and place of study), gender, age of students, institutional data as the quality of the universities	the introduction of Bachelor degrees did not affect the dropout rates for most subjects apart from Business Administration and Literature departments

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Hovdhaugen (2015): Working while studying: the impact of term-time employment on dropout rates	**Norway**: register data consisting of 12,726 students enrolled in full-time undergraduate education in the autumn semester of 2003	**economic perspective**: student employment is one of the external obligations affecting persistence in higher education **method**: survival analysis	employment status, field of study, study year, age, gender, parents' education, educational background, geographical background	employment status has an impact on dropout rates: students working full time are less likely to complete their program than those working short part-time
Stinebrickner and Stinebrickner (2014): Academic performance and college dropout: using longitudinal expectation data to estimate a learning model	**USA**: 341 students who entered Berea College in 2000 and 2001 from the longitudinal Berea Panel Study	**phase model**: dropout outcome is best viewed as the end result of a process in which a student learns about a variety of utility-influencing factors **method**: dynamic learning	age, gender, family situation, high school grade point average, academic performance, income expectations, satisfaction	45% of dropout in the first 2 years can be attributed to academic performance; poor performance decreases enjoyability of school and influences beliefs on post-college earnings
Aina (2013): Parental background and university dropout in Italy	**Italy**: 1,158 students from the longitudinal data of the European Community Household Panel from 1995 to 2001	**sociological perspective**: family background and income affect university attainment **method**: probit model	gender, age, demographic background, income, education background, learning behavior, migration, housing	dropout rates are higher for children with low parental education and for students with low family income

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Müller and Schneider (2013): Educational pathways and dropout from higher education in Germany	**Germany**: 11,649 individuals born between 1944 and 1986 from the National Educational Panel Study: Starting Cohort 6 (2009/2010)	**sociological perspective**: social origins and secondary educational pathways have at least initial impact on dropout risks **method:** multivariate analysis	social origin, expenditure of time, educational background	students taking direct pathway via the highest school track and those with a higher social background have lower dropout rates
Severiens and Ten Dam (2012): Leaving college: a gender comparison in male and female-dominated programs	**Netherlands**: Dutch census data on success in higher education from 1995 onwards plus a sample of 10,000 university leavers between 2000 and 2006	**sociological perspective**: gender differences in dropout can be explained by learner, external, institutional factors **method**: multivariate analysis	gender, field of study, graduate and dropout rates after a number of years, number of male and female freshmen in course programs, quality of education	low male retention rates seems to be especially low in female-dominated courses; men leave more often than women due to a perceived negative culture
Beerkens et al. (2011): University studies as a side job: causes and consequences of massive student employment in Estonia	**Estonia**: 2,496 students from 24 higher education institutions from the Survey of Students' Socio-Economic Situation conducted in 2008	**economic perspective**: student employment affects study progress; the effect depends on the work type, number of hours **method**: logistic regression	gender, family background, study situation, living conditions, funding, employment status, academic success	working while studying seems to have a negative effect on academic progress; student working at least 25 hours per week face academic hardships

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Glocker (2011): The effect of student aid on the duration of study	**Germany**: 787 individuals from the German Socio-Economic Panel observed for the years 1984-2007	**economic perspective**: reduction of the student aid leads to students working more, what affects the duration of study **method**: logistic model	age, gender, income per semester, marital status, time management, parental background, duration of study	increase of the student aid has no significant impact on the time to degree but affects the risk to dropout; students with good financial aid dropout less
Gury (2011): Dropping out of higher education in France: a micro-economic approach using survival analysis	**France**: 5,383 students enrolled at university and extracted from the longitudinal national survey of the French Ministry of Education	**sociological/economic**: dropout population shows considerable heterogeneity and differs with individual characteristics **method**: survival analysis	gender, socio-economic background, educational background, parents' level of education, parents' employment, financial situation	men and women do not exhibit the same dropout behavior; socio-economic background and parents' education have an effect only at the start of study
Van Bragt et al. (2011): Why students withdraw or continue their educational careers: a closer look at differences in study approaches and personal reasons	**Netherlands**: 1,176 second year students continuing a full-time bachelor's study in higher education	**sociological/ psychological**: quality of learning processes and study outcome depends on the quality of study approach **method**: t-test, principle component analysis	perception and experience of educational and organisational aspects, loss of interest in the future occupations, perception and experience of learning environment quality	students continuing their education show higher scores on meaningful integrative study approach; scores for withdrawing students are negatively correlated with perception

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Van Bragt et al. (2011): Looking for students' personal characteristics predicting study outcome	**Netherlands**: 1,471 freshmen and full-time students from a university of applied sciences in the Netherlands	**psychological perspective**: study outcome dependents on personal orientations on learning, study approach **method:** logistic regression	personalty traits (extraversion, agreeableness, conscientiousness, emotional stability, autonomy), personal orientations on learning, study approach	students with high scores on conscientiousness obtain higher credits; those with high scores on ambivalence, on lack of regulation dropout more easily
Belloc et al. (2010): University drop-out: an Italian experience	**Italy**: 9,725 undergraduates students from 2001 to 2007 of Economics and Business of the University of Rome	**sociological/economic**: dropout motives may be shaped by doubts on college expectations **method**: linear mixed model	age, gender, residence, type of high school, high school mark, actual performance in degree course, family income	students with high secondary school mark and those performing well in their degree course are less likely to dropout
Reisel and Brekke (2010): Minority dropout in higher education: a comparison of the United States and Norway using competing risk event history analysis	**Norway and USA**: cohort of a US American survey from 1988 to 2000 and a sample of 18-24-year students who started the study in Norway between 1990 and 1998	**sociological perspective**: a reason to expect differences in dropout rates may be that minority student have less educated and less wealthy parents **method:** logistic regression	gender, minority status, generation, socioeconomic background (parents' education, parents' income, family income), academic field, age at entry, year of entry	US higher education system tends to exacerbate initial socioeconomic inequalities bet- ween minority and majority students; in Norway no difference in dropout is observed

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Georg (2009): Individuelle und institutionelle Faktoren der Bereitschaft zum Studienabbruch: eine Mehrebenenanalyse mit Daten des konstanzer Studierendensurveys	**Germany**: data from approx. 10,000 students in the ninth survey of the Konstanz Student Survey from the winter term 2003/04	**sociological/ psychological**: a mono-causal explanation of dropout is not effective, but a bundle of several causes has to be considered **method:** regression models	intrinsic and extrinsic motivation, exam nerves, intermediate examination grade, high school grade, financial situation, time budget course, employment in term-time	willingness to dropout from university can not be explained by the level of ability or social stress, but by a lacking commitment to study in general and study field specifically
Hovdhaugen and Aamodt (2009): Learning environment: Relevant or not to students' decision to leave university?	**Norway**: 3,537 students surveyed in 2005 who started the undergraduate studies in humanities, social science	**sociological perspective**: experience inside the institution is more important for retention than experience before study **method:** factor analysis	gender, field of study, academic progress, reasons for transferring in another university, reason for dropping out	students dropped out because they got a job, or they were lagging behind in study progression due to failed examinations or lost of interest
Di Pietro and Cutillo (2008): Degree flexibility and university dropout: The Italian experience	**Italy**: 16,098-19,996 high school graduates surveyed 3 years after graduation from 3 waves (1998, 2001, 2004)	**sociological perspective**: Bologna reform could enable students struggling with study, to still manage to obtain a first degree rather than withdrawing **method:** probit model	gender, age, family background, academic ability, school related, geographical and university-related characteristics, area of residence	reform has led to a lower dropout risk due to the duration, the structure, the content of the supply of university education and a greater flexibility in the degree program

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Lassibille and Navarro Gómez (2008): Why do higher education students drop out? Evidence from Spain	**Spain**: 7,000 students from the university of Malaga observed over an eight-year period ending in 2004	**sociological/economic**: economic and sociological theories on dropout behavior **method:** proportional odds	age, gender, parental education, secondary educational experience, grade-point average, choice of course of study	academic preparedness is one of the major influences on student completion; older students are more likely to drop out
Mastekaasa and Smeby 2008): Educational choice and persistence in male- and female-dominated fields	**Norway**: 2,422 students from a large longitudinal survey program in five Norwegian university colleges in September 2000	**sociological perspective** the proportion of minority students has consequences for the degree of exposition to discrimination **method:** logistic regression	gender, proportion of men, early decision to study, mother's degree of encouragement, father's and friend's degree of encouragement	women have a much higher dropout probability in relatively balanced and male-dominated programs than in female-dominated ones
Schiefele et al. (2007): Aussteigen oder Durchhalten; Was unterscheidet Studienabbrecher von anderen Studierenden	**Germany**: 47 dropout students and a group of 94 matched regular students of the university of Bielefeld (1996-2002)	**sociological/ psychological**: significant conditions of drop- out: assessed performance, self directed learning, motivation **method**: variance analysis	study field, gender, age, vocational training, number of children, motivational factors, self-concept, learning strategies, social competence	motivation, self-estimated knowledge, learning strategies, social competence indicated the largest differences between dropout and regular students

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Suhre et al. (2007): Impact of degree program satisfaction on the persistence of college students	**Netherlands**: 186 first year law students who started law school in the university of Groningen after the pre-university education	**psychological perspective**: satisfaction with degree program is expected to have positive impact on study motivation **method** regression analysis	enrollment data, student background variables such as gender, grades for subjects in secondary education,	student success not only depends on differences in academic ability but also on degree program satisfaction, whose decrease reduces study motivation
Brandstätter et al. (2006): Prognose des Studienabbruchs	**Austria**: 948 high school gra- duates who had participated in a career counseling program (1991-1998) and started their study at the University of Linz	**sociological/ psychological**: students with similar values are more likely to meet, which facilitates academic success **method**: Cox-regression	grade point average, cognitive abilities, self-confidence, motivation, occupational imagination, satisfaction with study course and study organisation	low grade point average, low scores on cognitive test, on emotional stability, on conscientiousness lead to bad performance, as result to dropout
Di Pietro (2006): Regional labour market conditions and university dropout rates: Evidence from Italy	**Italy**: 5,907 students from the academic years 1987-88 and 1997-98 derived from the Italian National Statistical Center	**economic perspective**: decisions to invest in education are mainly affected by the direct and opportunity costs **method**: regression analysis	gender, age, region of residence, unemployment rate, family background, high school grade point average, marital status	negative relationship between regional unemployment and dropout; students may dropout to benefit from the improved labour market conditions

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Glaesser (2006): Dropping out of further education: a fresh start? Findings from a German longitudinal study	**Germany**: 1500 participants contacted to participate 20 years later in a follow-up study after a prior survey which took place in the years 1979-1983	**sociological/ psychological**: reasons for dropout are related to course, wrong expectations, course disappointment **method**: logistic regression	demographic variables, parent education, school qualification, active in club or church, high school grade, verbal intelligence, learning motivation	individual factors (scholastic achievement, intelligence or motivation) and demographic factors contribute to dropping out and starting again
Hansen and Mastekaasa (2006): Social origins and academic performance at university	**Norway**: 58,000 first-year students and 24,000 higher-level graduates in Norwegian universities in the periods 1997 to 2002 and 1997 to 2003	**sociological perspective**: according to cultural capital theory, students from academical families have great success **method:** logistic regression	gender, university type, and degree of urbanisation, family background, social origin, parental level of education, academic performance	there is an association between class origin and academic performance; students originated from high cultural classes receive the highest grades
Johnes and McNabb (2004): Never give up on the good times: student attrition in the UK	**United Kingdom**: about 100,000 university leavers from 1993 from English and Welsh universities	**sociological perspective:** peer groups and the quality of the match between a university and students are important **method**: logistic regression	academic ability, social integration, gender, date of birth, marital status, high school grade point average, parental occupation, type of school	academically able male students have a higher probability of non-completion of degree programs on which the overall level of ability is relatively low

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Mäkinen et al. (2004): Students at risk: students' general study orientations and abandoning/prolonging the course of studies	**Finland**: 1,600 first year students of a multidisciplinary Finnish university from September 1998 to January 1999	**sociological/ psychological**: problems of commitment at the beginning of studies disturb later engagement in studying **method**: cluster analysis	gender, social class, parents' education level, study orientation, study motivation, field of study, study presence or absence, grade point average	students intending to change their major subject or to abandon their studies altogether belonged most often to the group of non-committed students
Brandstätter and Farthofer (2003): Einfluss von Erwerbstätigkeit auf den Studienerfolg	**Austria**: 361 students of the University of Linz who had participated in the counseling program in the years 1992 to 1998 before entering university	**sociological/economic**: time spent for study and working has an effect on various success criteria of studies **method**: variance analysis	study experience, gainful employment, socio-economic status of parents, personal characteristics, academic success, professional interest	the criteria of academic success (number of exams, grade point average, study satisfaction) are negatively affected by allotting more time to paid work
Montmarquette et al. (2001): The determinants of university dropouts: a bivariate probability model with sample selection	**Canada**: 3,418 students from the University of Montreal surveyed in three semesters (fall 1987, winter 1988 and fall 1988)	**sociological/ psychological**: human capital and experimental models can be used to explain dropout behaviour **method**: probability model	Personal characteristics: age, gender, academic performance, university grade point average; socio-economic factors: mother tongue, region of origin	variables explaining persistence (or early dropouts) are related to a non-traditional class-size effect in the first-year mandatory courses taken by students

Continued on next page

Publication	Country & sample size	Theory & method used	Main factors explored	Main result(s)
Smith and Naylor (2001): Dropping out of university: a statistical analysis of the probability of withdrawal for UK university students	**United Kingdom**: full cohort of 94,485 undergraduate students who had left UK universities in 1993	**sociological perspective**: personal characteristics and A-level scores can explain the variance in degree performance **method**: probit regression	personal information: gender, birth day, marital status; academic history: last time school attended, A-level; annual information: university, subject	degree performance is influenced significantly by age, marital status; positively by A-level score, occupationally-ranked social class background
Johnes (1990): Determinants of student wastage in higher education	**United Kingdom**: sample of 328 students who entered Lancaster University in 1979	**sociological perspective**: voluntary non-graduates find courses unsuitable; involuntary dropouts leave due to failure **method**: logistic regression	personal factors: age, gender, marital status; academic-related factors: A-level score, type of high school; study motivation and commitment	the likelihood of non-comple- tion is determined by various factors: academic ability, work experience prior to university, school background and location

3 Dropping out from higher education in Germany - an empirical evaluation of determinants for Bachelor students

Dropping out from higher education in Germany - an empirical evaluation of determinants for Bachelor students

Andreas Behr, Marco Giese, Herve D. Teguim K., Katja Theune
Chair of Statistics
University of Duisburg-Essen, 45117 Essen, Germany

Abstract

Withdrawing from university is a complex decision-making process, during which several conditions and problems from different areas of life and study accumulate and affect each other. This study is based on the National Educational Panel Study (NEPS), which includes a wide range of information on study course and students' characteristics, and aims at providing an encompassing analysis of determinants influencing students' dropout decision. Determinants can be categorized into demographic and family background, the financial situation of students, their prior education, institutional determinants, as well as motivation and satisfaction with study. Both, a bivariate analysis, as well as a logistic regression model with LASSO regularization identify many important determinants already known before or at the beginning of the study, such as prior education and satisfaction related variables, allowing early identification of at-risk students and the implementation of prevention programs.

Keywords: feature selection, dropout prediction, logistic regression, evaluation methods

3.1 Introduction

Due to the rising number of students in higher education institutions and the social and personal costs related to dropping out of university, analyzing study success and study dropout becomes more and more important. In Germany, the number of students enrolled at institutions of tertiary education increased monotonically over the last years towards 2.9 million in winter term 2019/2020, which is associated with increased educational costs (DESTATIS, 2018). In Germany, 14.7% of Bachelor students do not finish their degree (Schnepf, 2014). Other European countries face an even higher number of students dropping out of higher education, e.g. France (17.9%), Spain (24.2%), Netherlands (28.3%) or Italy (34.1%) (Schnepf, 2014). To minimize the wasting of financial and human resources due to a high number of university dropouts, policy and educational institutions are increasingly interested in detecting determinants that influence the dropout decision.

The current state of empirical research on student dropout carried out within a wide range of disciplines has identified several possible reasons for withdrawing from tertiary education. These include, for instance, demographic and family background, the financial situation of students, prior education, institutional determinants, as well as motivation and satisfaction with study.

This study aims to provide an encompassing analysis of these potential determinants and includes predictors from all of the mentioned categories. Therefore, a bivariate analysis using different effect size measures and a multivariate logit model are used. The database is the National Educational Panel Study (NEPS), a continuously growing and comprehensive German panel study including many variables covering a wide range of possible dropout determinants from different areas (Blossfeld et al., 2011).

Since the number of observations (17,910) and variables (more than 3,000) is large, both steps of the empirical analysis - bivariate and multivariate - are important. Simple bivariate analysis provides an overview of all relevant variables in the dataset and first impressions on their potential usefulness and importance. The main disadvantage is that a large absolute effect size can result in a small partial effect in a multivariate setting due to intercorrelations of the predictors. Therefore, we regard the bivariate analysis as a

prerequisite for the more computing-intensive multivariate models with feature selection (Hastie et al., 2009).

The results of both analyses help to identify many promising starting points for early warning systems for students being at-risk of dropping out.

3.2 Determinants influencing dropouts - a literature review

Higher education dropout is not always defined consistently in the literature. Based on theoretical considerations, many different definitions have been applied and a distinction should be made according to the level at which dropouts occur. Students may change their field of study (within the same subject area or between subject areas), the type of degree, the (type of) university, or students may leave the university system, for instance, due to academic failure, wrong expectations or to favorable job offers (Tinto, 1975, Larsen et al., 2013c). Depending on student's or faculty's perspective, these different types of dropouts could be perceived as transfers (e.g. from one field to another) or as a formal total dropout. The former is sometimes called "re-selection" (Larsen et al., 2013c) or "institutional departure" (Tinto, 1993, p. 36), the latter "de-selection" (Larsen et al., 2013c) or "system departure" (Tinto, 1993, p. 36).

Withdrawing from university is seldom the result of short-term or spontaneous decisions, but rather of a long decision-making process, during which several conditions and problems accumulate and prompt students to leave university without a degree (Heublein, 2014a). Previous studies investigated the dropout of tertiary education in several countries with different focuses and identified several possible reasons for dropping out. Behr et al. (2020a) provide an encompassing and up to date review. These determinants can be categorized into societal aspects which include the demographic and family background, the financial situation of students, and their prior education, into institutional determinants, as well as into motivation and satisfaction with study.

3.2.1 Demographic and family background

Pre-study demographic and background factors seem to have a strong influence on study performance and dropout. Aina (2013) and Ghignoni (2017), both focusing on the re-

lationship between the family background and the dropout decision in Italy, find that the better the parental education and social class, the lower the probability for leaving university without degree. Some studies observe that male students tend to drop out more frequently than female students. Mastekaasa and Smeby (2008) for Norway and Severiens and Ten Dam (2012) for the Netherlands analyze the impact of dropout in male- and female-dominated study fields, and reveal that men have a very high attrition rate in female-dominated fields while women dropout to a lesser extent in those courses. Furthermore, there is evidence that a higher age at enrolment increases the dropout probability (e.g. Müller and Schneider, 2013, Lassibille and Navarro Gómez, 2008), which may also explain the higher dropout rate for students with vocational training before entering higher education (Müller and Schneider, 2013). Reisel and Brekke (2009) investigate the connection between higher education performance and the migration background in Norway and the USA and state that the dropout probability is higher for students from a foreign country, which is also observed by Belloc et al. (2010) for Italy. Similarly to Sarcletti and Müller (2011), they find that students with migration background tend to have less knowledge about the education system and the prevailing culture, and are less familiar with the language which increases the risk of dropping out. According to Aina (2013) and Di Pietro (2006), the latter of whom analyzes the relationship between regional labor market conditions and university dropout rates in Italy, the geographic area plays an important role. Students from economically stronger regions and with good labor market prospects have a higher probability of enrolling in tertiary education and a lower dropout rate.

3.2.2 Financial situation

Another important aspect of study success is the students' financial situation which is related to the possibility of financial support, as well as to their amount of off-study work. A study by Glocker (2011), investigating the effect of financial aid on study success in Germany, reveals that an increased amount of support students receive decreases the dropout rate significantly. According to a Norwegian study by Hovdhaugen (2015), working more than 20 hours a week increases the probability of dropping out, whereas working for a maximum of 19 hours a week seems to have no significant influence on study success. Similar results are reported by Beerkens et al. (2011) for students from Estonia. They observe that more than 25 hours of off-study work decreases the probability of timely graduation.

3.2.3 Prior education

The pre-study education of students seems to be very important for the study success. Müller and Schneider (2013) examine the relationship between pre-tertiary educational pathways and dropout from tertiary education in Germany. They observe that students from the upper secondary school track (e.g. Gymnasium in Germany) and with a standard educational pathway have a lower dropout rate than students from the lower or intermediate track. Especially, students with vocational training before studies tend to have a high dropout rate, which may be associated with increased age at study start (see section 3.2.1). According to Sarcletti and Müller (2011), school performance is of particular importance for study success as it is an indicator of the ability to meet the level of performance required by the higher education system. Various international studies find positive correlations between school (e.g. GPA) and study performance, for instance, Stinebrickner and Stinebrickner (2014), who analyze students at the Berea College in the USA.

3.2.4 Institutional determinants

The type of higher education institution also influences the dropout decision of students. For instance in Germany, Sarcletti and Müller (2011) find the dropout rates in Bachelor courses at universities of applied science to be lower than those at universities. The same observations are made by Heublein et al. (2017), who also reveal the highest dropout rates in Germany to be in Engineering, Mathematics and Natural Sciences. This result is confirmed by Lassibille and Navarro Gómez (2008) for Spain and Korhonen and Rautopuro (2018) for Finland. Moreover, there are some important determinants related to study conditions that affect students' decision to drop out. Hovdhaugen and Aamodt (2009) analyze the impact of the learning environment on leaving university for Norwegian students and find that poor teaching quality and an unfavorable learning environment increase the probability of dropping out. A similar observation is made by Georg (2009) for German students. Suhre et al. (2007) for the Netherlands and Ghignoni (2017) for Italy highlight the importance of the relationship between students and teachers. Furthermore, a good program organization (Heublein et al., 2017) and program flexibility (Di Pietro and Cutillo, 2008) seem to decrease the probability of withdrawal.

3.2.5 Motivation and satisfaction with study

Besides these easily measurable determinants, also students' motivation and satisfaction with study affect their risk of dropping out. The latter determinants are based on the students' subjective self-perception, who have to state a value on a pre-defined scale to measure these variables. Suhre et al. (2007) investigate the association between study satisfaction and dropout probability in the Netherlands and observe unsatisfied students to have a higher risk of withdrawal. A German study by Suhlmann et al. (2018) finds the fit between the higher education institution and personal attitudes to be strongly related to students' satisfaction and motivation which further decreases the probability of dropping out (Schiefele et al., 2007). Nordmann et al. (2019) for the UK and Korhonen and Rautopuro (2018) for Finland find that class attendance and time spent on the study course have a positive influence on study performance. Moreover, according to Van Bragt et al. (2011a) and Van Bragt et al. (2011b), both focusing on the relevance of students' personal characteristics for study success in the Netherlands, aspects such as conscientiousness, ambivalence or attribution are very important for educational performance. Other studies confirm the importance of personal characteristics including, for instance, resilience and self-control (e.g. Brandstätter et al., 2006).

3.2.6 Summary and contribution

To sum up, there are many different aspects of students' life including the pre-study phase, the institutional setting, the financial situation and motivational aspects, which seem to be relevant for the dropout decision. There are also some reviews on dropout research, which group the wide range of predictors in a similar way. For instance, Vossensteyn et al. (2015) categorize them into determinants on the individual level, on the institutional level and those on the level of the higher education system. According to Rodríguez-Gómez et al. (2015), focusing on definitions and common reasons for dropout in America and Europe, dropout is a multi-factor phenomenon which is the result of a complex interaction of determinants from a wide range of reasons including external, institutional, and personal factors among others.

Therefore, to obtain detailed and comprehensive insights into the dropout phenomenon, there are some implications for the data and the methodological approach. First, all of these determinants (categories) found to be important should be considered in the

analysis. Previous research mainly focused only on one or a few aspects of dropout and, as also stated in Larsen et al. (2013c), mainly on pre-study or university "non-malleable" determinants, but research would benefit from dealing more with study-related and university malleable determinants, as these are mainly within the scope of policy action. Singell and Waddell (2010) and Gury (2011) emphasized the importance of both fixed and time-varying effects (e.g. study conditions) on withdrawal, which cannot be analyzed with cross-section data. Administrative data, which have been used in many studies, lack information on pre-study determinants and on determinants based on the subjective self-perception of students. Survey data often contain only too few observations to get representative and reliable results. Therefore, as also claimed in Sarcletti and Müller (2011), large prospective and longitudinal data covering determinants before and at the beginning of the study, as well as students' subjective self-perceptions, are of considerable importance for assessing the dropout phenomenon in its entirety. Moreover, it seems to be important to sort and condense the large number of determinants, to evaluate their degree of impact and to detect the most important ones in the dropout prediction, so as to identify promising and efficient starting points for reducing dropout rates. This study uses a large German survey dataset which covers a wide range of student life and intends to include determinants from all of the mentioned categories. Beside a bivariate analysis of the relevance of these different determinants by measures of effect size, this study aims at identifying the most important ones by applying a LASSO (least absolute shrinkage and selection operator) regression with an internal feature selection. It is hypothesized that from each of the identified determinant categories important features are selected for the final dropout prediction model.

3.3 The National Educational Panel Study

3.3.1 Sample description

The fifth cohort of the National Educational Panel Study (NEPS)[1] is a comprehensive German panel study including students in tertiary education covering a wide range of

[1]This work uses data from the National Education Panel (NEPS): Starting Cohort Students, doi:10.5157/NEPS:SC5:9.0.0. The NEPS data has been collected from 2008 to 2013 as part of the framework program for supporting educational research, funded by the German Federal Ministry of Education and Research (BMBF). Since 2014, NEPS has been continued by the Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) at the Otto-Friedrich-University Bamberg in cooperation with a Germany-wide network.

different aspects of students' background and the course of study (Blossfeld et al., 2011). This study uses nine waves which have been obtained by different survey methods like computer-assisted telephone interviews (CATI), competency tests, as well as computer-assisted web interviews (CAWI). The target population are first-year students (German and non-German) at higher education institutions in Germany in winter term 2010/2011. Interviewed students must be enrolled for the first time at public or state-approved higher education institutions aiming at a Bachelor degree, state examination (medicine, law, pharmacy, teaching), diploma or Master (Roman Catholic or Protestant theology) or specific art and design degrees (Zinn et al., 2017). In the first wave, 17,910 students participated in the NEPS. Table 3.8 in the appendix provides some general information on the dataset.

One limitation of this type of data is the long time-horizon that is necessary to finally evaluate first-semester students. The dataset contains the freshmen cohort of winter-term 2010/11 and represents the most recent study of this sample size and quality in Germany. As we focus on the examination of dropping out at an early stage of study, we therefore use mainly time-invariant variables and determinants from the early study phase. These variables are collected mainly already at the begin of the survey in 2011. Furthermore, since 2010 no major changes have taken place in the German higher education system that can lead to a huge change in the influencing variables as the Bologna process in 1999. A further limitation of the study is caused by panel attrition. This problem has already been analyzed by Behr et al. (2020c) and has no negative consequences on their model.

The sample is drawn as a stratified cluster sample. Clusters are defined by all students enrolled in a certain subject at a particular higher education institution. To oversample teacher education students and students attending private higher education institutions (as little is known about these groups), first-level stratification according to educational institutions was applied. The second level of stratification (within the first-level strata) was conducted according to groups of related subjects. These techniques for composing the NEPS sample, which should represent the entire freshman student population in Germany as closely as possible, are based on data on first-year students from winter term 2008/2009 from the Federal Statistical Office of Germany (Fachserie 11 Reihe 4.1: Bildung und Kultur - Studierende an Hochschulen) (Zinn et al., 2017).

Table 3.1 provides an overview of some relevant characteristics of students participating

in wave 1 (own calculations). There is a substantial overrepresentation of female students (60.46%), but using sample weights (provided in the scientific use file), the proportions of female students, as well as the type of institution, are very similar to the population proportions of beginning students in winter term 2010/2011 in Germany (Statistisches Bundesamt, 2011).

Table 3.1: Student characteristics (wave 1)

Attribute	abs.	%	% weighted
sex			
women	10,828	60.46%	50.71%
men	7,082	39.54%	49.29%
nationality			
German	17,382	97.05%	94.53%
non-German	528	2.95%	5.47%
birth year			
1990-1995	10,360	57.84%	53.05%
1989 and older	7,550	42.16%	46.95%
institution			
appl. sciences	4,259	23.78%	37.15%
university	13,642	76.17%	62.76%
other/n.a.	9	0.05%	0.09%
int. degree			
Bachelor	10,854	60.60%	82.16%
state exam.	1,428	7.97%	7.84%
teacher educ.	5,554	31.01%	9.26%
other	74	0.41%	0.74%

Table 3.2 provides an overview of the distribution of study fields (first field) in the first wave. Again, weighted values of field proportions are very similar to those provided by the Statistical Office. Almost one-third of beginning students start to study in the field of law, economics and social sciences (31.27%), followed by engineering (21.47%), mathematics and physical sciences (18.82%) and linguistics and cultural studies (17.09%).

3.3.2 Predictor variables included in the study

Since the NEPS contains more than 3,000 variables in total, a variable pre-selection is necessary in order not to exceed the scope of the article. To ensure sufficient data quality only variables with less than 20% missing values in the target population are

Table 3.2: Beginning students in Germany in winter term 2010/11 by field of study

study field	abs.	%	% weighted	% stat. office
Linguistics, cultural studies	4,773	26.65%	17.09%	17.43%
Sports	259	1.45%	0.98%	0.97%
Law, economics. social sciences	4,539	25.34%	31.27%	32.30%
Mathematics, physical science	3,910	21.83%	18.82%	17.27%
Human medicine, health sciences	844	4.71%	4.45%	4.22%
Veterinary medicine	49	0.27%	0.29%	0.27%
Agricult., wood, nutrit. sciences	394	2.20%	2.34%	2.21%
Engineering	2,636	14.72%	21.47%	21.70%
Arts, Art science	446	2.49%	3.16%	3.36%
n.a.	60	0.34%	0.14%	0.27%

used, whereby some variables do not apply to every student, e.g. a student is only asked at the beginning of their study if not born in Germany. Since dropout prevention should begin at an early stage of study, we focus on the first waves. This criterion reduces the number of variables to less than 200. The final variable pre-selection is made from a theoretical point of view. Features that have not been found to be important in any previous articles, and also have no considerable influence here in terms of effect-size, are excluded. The final sample includes 52 variables.

These variables were grouped into the identified five thematic fields: Demographic and family background, financial situation, prior education, institutional determinants, and motivation and satisfaction with study.

3.3.3 Identifying dropouts

According to Larsen et al. (2013c), the term "university dropout" can simply be explained as leaving the higher education system without obtaining a degree. This definition is from a macro point of view and mainly important for the whole education system and society. An alternative dropout definition includes students who change their subject field or institution before graduation. This second definition relates to a micro point of view, that of a faculty or institution, for which changes before the first degree could represent a failure in their goal of avoiding dropout from their study program. Here, dropout is defined as leaving the higher education system without a first degree. Changes of the study field, degree or institution are not treated as dropouts, but are considered in the analysis as predictors for dropping out.

The outcome variable for dropping out in this analysis is based on the "status" of a student showing one of the following four categories:

0. Graduate
1. Dropout
2. Still studying
3. Status is not available (NA)

Since the focus of this article lies in the identification/prediction of potential dropout students, the aim is to compare dropouts and graduates. Students who are still studying and those with an unknown status are disregarded in the empirical analysis.

The final sample contains 943 students identified as dropouts and 2,625 graduates ($N = 3,568$). Students' status is a binary variable, where 0 is indicating a graduate and 1 indicating a dropout. The status variable is constructed using relevant variables until wave 9 (summer term 2015). The relative small final sample, compared to the number of participants in wave 1, is a result of right-censored data (many students are still studying), and missing values, since not every student participated in all nine waves.

3.4 Bivariate analysis of dropout determinants

Let Y be the status variable and $\mathbf{X} = \{X_1, \ldots, X_k\}$ a set of k determinants, that are potentially related to the status Y. In section 3.5 (multivariate analysis), the focus lies on $P(Y = 1|X_1 = x_1, \ldots X_p = x_p)$, which denotes the probability that a student drops out, given a subset $p \leq k$ of known determinants (e.g. gender, high school grade etc.). To determine those p variables that might influence the conditional probability of Y, the mean (M) of a specific determinant X_j is compared in the two groups: $M_{0,j} = M(X_j|Y = 0)$ (mean in the group of graduates) and $M_{1,j} = M(X_j|Y = 1)$ (mean in the group of dropouts), $j = 1, \ldots, k$.

The bivariate analysis aims to detect variables differing strongly between dropouts and graduates. Two effect size measures are used to identify differences in the mean of the two groups of dropouts and graduates, which can also be seen as correlation measures: 1) Cohen's d and 2) Point-biserial correlation. The higher the absolute effect size, the larger the mean difference between the two groups (Hartung et al., 2011). In general, one can

expect variables with high absolute effect sizes to have more influence on the probability to drop out. In contrast to tests for statistical significance, effect size measures are not influenced by the sample size in the two populations. [2]

Let $M_{1,j}$ and $M_{0,j}$ be the weighted mean of variable X_j in the group of dropouts and graduates. The absolute point-biserial correlation coefficient r_{pb} is a correlation measure for a dichotomous variable (here the status Y) and a metric variable $X_j, j = 1, \ldots, k$ (Bortz and Schuster, 2010), calculated by

$$|r_{pb}| = \frac{|M_{1,j} - M_{0,j}|}{S_{n_j}} \sqrt{\frac{n_{1,j} n_{0,j}}{n_j^2}} \in [0, 1],$$

with sample sizes $n_{1,j}$ (dropout) and $n_{0,j}$ (graduate) in the two groups, S_{n_j} the overall standard deviation and the overall sample size $n_j = n_{1,j} + n_{0,j}$. The sample sizes $n_{1,j}$ and $n_{0,j}$ vary dependently on the number of missing values of the variable X_j and are given in Table 3.6 in the appendix. Table 3.6 also provides information on variable description, coding and scaling.

Cohen's d (Hartung et al., 2011) is a measure of effect size and defined as

$$\text{Cohen's } d = \frac{M_{1,j} - M_{0,j}}{S},$$

where S is the pooled variance $S^2 = \dfrac{(n_{1,j} - 1)S_{1,j}^2 + (n_{2,j} - 1)S_{2,j}^2}{n_{1,j} + n_{2,j}}$ and $S_{1,j}^2$ and $S_{2,j}^2$ are the weighted variances in the two groups for variable j. The interest is more on the absolute value of Cohen's d, i.e. |Cohen's d| to give a ranking of variables with comparable huge differences in the two groups.

Table 3.3 shows the results of a bivariate analysis of the status variable and the different predictors. In each thematic field, a ranking of variables beginning with the largest absolute Cohen's d is presented.[3]

[2] Statistical tests can be highly accurate for large samples, where even small differences can be detected easily, whereas for small samples they often fail. In regression models or, as in this case, binary classification models, also importance ranking exists, e.g. for random forests (Breiman, 2001). Highly correlated features can influence the importance of a variable in those situations.

[3] For all calculations in this article the statistical software R (R Core Team, 2019) is used.

3.4.1 Demographic and family background

According to Table 3.3, female students tend to outperform their male peers (the weighted mean of male students in the dropout group is 54.4% and only 41.6% among the graduates), and the students in the dropout group are on average somewhat older (year of birth), which is in line with previous literature. Students living in the new eastern federal states of Germany (place of residence) tend to drop out more frequently. According to Aina (2013), students may profit from financial benefits in economically stronger regions (like the old western federal states of Germany). The immigration background has only a minor effect on study success here. There are no consistent results in the literature regarding the effect of immigration background because this effect depends strongly on the country and its national (education) system (Reisel and Brekke, 2009). According to the family background, graduates' mothers and fathers tend to have a higher occupational prestige (coded using the ISEI-08 standard - International Socio-Economic Index of Occupational Status) than parents of dropouts. Mothers of university graduates are on average better educated than those of dropout students. The highest father's diploma seems to play only a minor role. These results are mainly in line with previous studies.

3.4.2 Financial situation

Strongly related to the family background of students is their financial situation and off-study work. Dropouts more often receive BAfoeG (BundesAusbildungsfoerderungs-Gesetz, financial support for students with a poor socio-economic background), which indicates that dropout students more often come from financially weak families. There is just a small difference in the financial income of the two groups, which is in line with Heublein et al. (2008). Graduates work, on average, more than five hours per week more than the dropouts during the term break. During the semester, the difference is not significant and the correlation with the status variable is small. Similarly, previous studies show that working only a few hours has no negative association with study performance. The possibility/willingness to give up other, competing goals to invest in study (study costs) is lower in the dropout group.

3.4.3 Prior education

Educational achievements up to secondary education generally influence the higher education performance (Sarcletti and Müller, 2011). According to Table 3.3, graduates generally seem to be much better prepared and informed than dropouts. The skills acquired before tertiary education (especially mathematical skills) are also of high importance. These aspects have not been in the focus of previous studies. The overall school grade seems to have a large effect size and is highly correlated with the dropout decision. University graduates obtained on a scale from 1 to 4 (1 is the highest grade and 4 the lowest) an average school grade of 2.3 compared to an average grade of 2.7 for university dropouts. Similar results were found in various international studies. Moreover, the number of repeated classes in high school is lower among graduates than among the dropouts and reveals a large effect size. According to the type of high school, students can achieve a general university entrance qualification (the highest one), or a university of applied science entrance qualification (the middle one) or other lower degrees. About 70% of the graduates attended a Gymnasium (highest school track) and only 60% of the dropouts. Related to this, graduates obtained on average a higher school leaving qualification. These results are in line with previous research.

3.4.4 Institutional determinants

Institutional determinants provide information about the structure, organization and study conditions of higher education institutions, which also determine study success. Large differences in dropout rates between the different types of higher education institutions are observed. The majority of individuals in the dropout group withdraw from general university (58.1%) compared to only 35.2% who graduated from a university. General universities are more theory-oriented, while universities of applied science focus on practical applications and offer more structured study programs (Mayer et al., 2007). Note that lower dropout and higher graduation rates may be due to a differing subject profile of universities of applied sciences and the usually shorter time to completion. Figure 3.1 shows the distribution of study fields in the dropout and graduate group. The presented percentages for one specific group over the eight fields sum up to 100%. The highest difference between the dropout and the graduation group is observed for Law, economics and social sciences, which also has the largest effect size of all subject groups.

Comparing the dropout rates within each study field, the highest dropout rates are observed for Engineering and Mathematics and natural sciences. Similar observations are also made by Heublein et al. (2017).

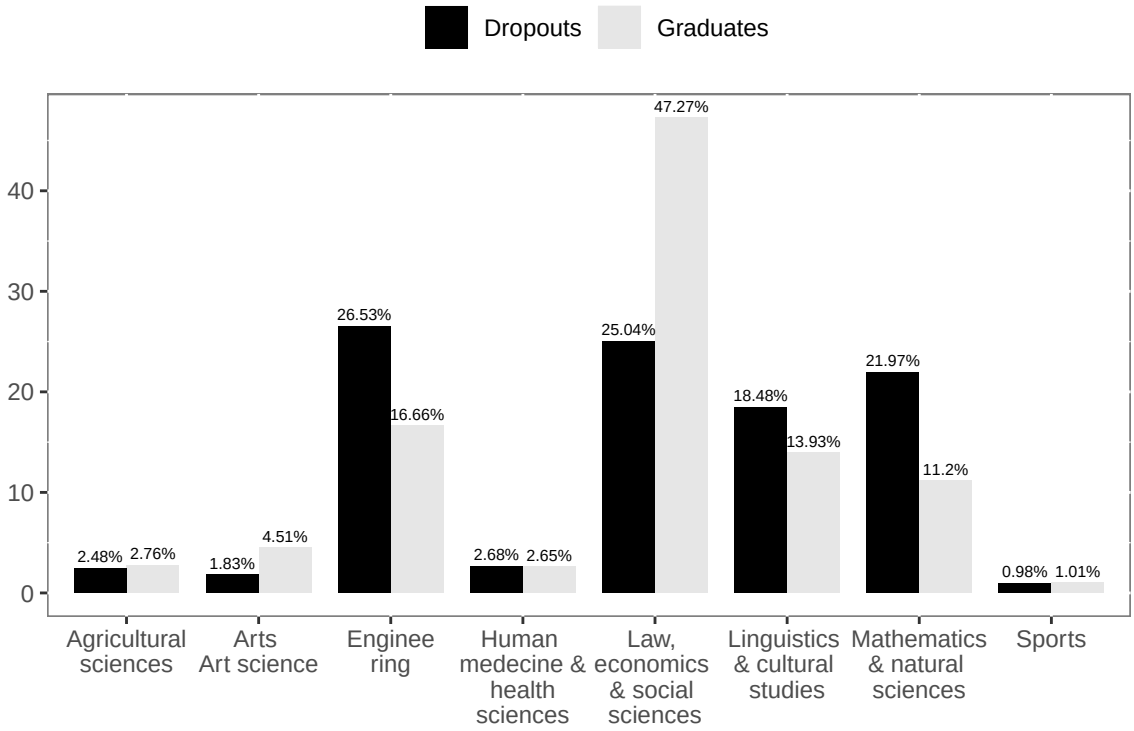

Figure 3.1: Distribution of subjects in the dropout and graduate group

3.4.5 Motivation and satisfaction with study

Determinants based on the subjective self-perception of students such as motivation and satisfaction also may influence academic success. The results indicate that graduates are more extrinsically motivated than dropouts. Related to that, 13.1% of dropouts compared to only 4.4% of graduates had preferred to do something else than studying (alternative to degree). These findings are mainly in line with the sparse previous research on such aspects. The proportion of individuals who feel disappointed concerning their chosen subject (subject of choice satisfied) is significantly higher in the dropout group than among graduates (28% vs. 17%). These effects have not been analyzed in previous research in detail, but Mora (2008) states that students' subject of choice is sometimes pressured by parents, teachers, and peers. Being satisfied with actual studies on the whole, enjoying the degree course, as well as being interested in the degree course

are highly correlated with study success. Additionally, dropouts are more concerned about some frustrating points of the degree course such as “frustrating external circumstances” or “degree course is wearing me down”. Similarly, the sparse previous research on some of these aspects find student satisfaction to have a positive impact on their intention to stay in college.

Table 3.3: Bivariate analysis of determinants and student status

Variable	Dropouts group		Graduates group		Effect size	
	Mean	Std. Err.	Mean	Std. Err.	|Cohen's d|	$\|r_{pb}\|$
Demographic and family background						
Gender (1 for male)	0.544	0.498	0.416	0.493	0.235	0.103
Father's occupation	48.351	21.536	51.106	21.392	0.129	0.056
Mother's occupation	47.047	19.197	49.413	18.392	0.127	0.056
Highest mother's education	4.212	2.248	4.485	2.115	0.091	0.040
Place of residence (West=0, East=1)	0.255	0.436	0.219	0.414	0.084	0.037
Highest father's education	4.553	2.407	4.747	2.325	0.072	0.032
Immigration background	0.235	0.424	0.208	0.406	0.066	0.029
Year of birth	1986.420	6.595	1987.874	4.348	0.014	0.006
Financial situation						
Time budget term break: Employment	12.026	16.117	17.256	16.538	0.318	0.129
Study costs	-0.123	0.901	0.080	0.836	0.239	0.092
BAfoeG	0.914	1.040	0.744	0.970	0.173	0.066
Monthly income	1017.813	893.374	1054.987	823.961	0.044	0.017
Time budget semester: Employment	6.663	11.629	6.409	10.148	0.024	0.010
Prior education						
General preparation	-1.371	3.392	0.603	3.243	0.603	0.244
Overall grade on school-leaving qualification	2.691	0.593	2.318	0.590	0.542	0.235
Number of repeated classes	0.351	0.628	0.171	0.430	0.366	0.160
Maths skills acquired before university	0.181	0.794	0.423	0.784	0.307	0.127
German skills acquired before university	0.483	0.897	0.609	0.831	0.148	0.061
Type of school-leaving qualification	1.537	0.707	1.673	0.597	0.114	0.050
Informed about study	0.973	1.583	1.130	1.399	0.108	0.048
English skills acquired before university	0.393	0.854	0.441	0.831	0.057	0.024
Reading competence	-0.167	0.934	-0.118	0.857	0.056	0.021
Mathematics as Abitur core subject	0.463	0.499	0.501	0.500	0.035	0.015
Type of high school	0.604	0.489	0.689	0.463	0.033	0.015
German as Abitur core subject	0.497	0.500	0.509	0.500	0.010	0.004
Institutional determinants						
General university	0.581	0.494	0.352	0.478	0.463	0.204
Subject (Law, economics and social sciences)	0.250	0.433	0.473	0.499	0.425	0.198
Subject (Mathematics and natural sciences)	0.220	0.414	0.112	0.315	0.321	0.136

Subject (Engineering)	0.265	0.442	0.167	0.373	0.228	0.110
Subject (Linguistics and cultural sciences)	0.185	0.388	0.139	0.346	0.098	0.056
Motivation and satisfaction						
Satisfied with actual studies	6.182	2.502	7.833	1.645	0.870	0.312
Enjoy degree course	6.274	2.208	7.500	1.688	0.666	0.240
Performance related extrinsic motivation	2.957	0.525	3.253	0.469	0.640	0.173
Degree course is interesting	7.161	1.997	7.829	1.652	0.432	0.139
Career related extrinsic motivation	3.281	0.642	3.440	0.503	0.377	0.106
Competition related extrinsic motivation	2.144	0.568	2.369	0.604	0.363	0.105
Alternative to degree	0.131	0.338	0.044	0.206	0.352	0.152
Frustrating external circumstances	4.102	2.694	3.255	2.550	0.325	0.118
Degree course is wearing me down	3.608	2.694	2.848	2.442	0.286	0.109
Subject of choice satisfied	0.717	0.451	0.825	0.380	0.271	0.120
Bridging courses	3.008	0.924	3.142	0.815	0.242	0.074
Degree course obligations hard to match	4.981	2.635	4.419	2.590	0.181	0.078
Concerns of students are not taken into account	5.138	2.635	4.774	2.511	0.181	0.052
Wishing better study conditions	6.231	2.768	5.877	2.825	0.178	0.045
Often tired due to degree course	5.083	2.686	4.779	2.453	0.120	0.044
Intrinsic motivation	3.087	0.589	3.119	0.551	0.091	0.016
Events/forums offered to get to know people	3.218	0.657	3.276	0.664	0.087	0.038
Events/forums concerning study organisation	3.112	0.712	3.146	0.050	0.050	0.022
Higher education institution of choice satisfied	0.848	0.360	0.874	0.332	0.041	0.034

3.5 Methodological considerations

Due to possible intercorrelations among the predictor variables, bivariate effects observed in the previous section may change substantially when all the variables are considered. In this section, partial effects of the variables are analyzed in a multivariate setting using logistic regression.

3.5.1 Logistic regression (logit)

Logit is one of the most popular linear methods utilised for classification problems. Here, the dependent class variable is a binary variable Y containing two possible values: 1 (here for dropouts) and 0 (for graduates). In a logit model, the aim is to estimate the posterior probabilities of both classes via an index function F based on the predictor variables $\mathbf{X} = (X_1, \cdots, X_d)$ (here, $d = 52$).

Probabilities of both events depending on the predictor variables $\mathbf{X}$ are defined as

$$P(Y = 1|\mathbf{X}) = F(\beta_0 + \mathbf{X}^T\boldsymbol{\beta}) = F(\beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_d \cdot X_d), \tag{3.1}$$

$$P(Y = 0|\mathbf{X}) = 1 - P(Y = 1|\mathbf{X}) = 1 - F(\beta_0 + \mathbf{X}^T\boldsymbol{\beta}). \tag{3.2}$$

In the logit model, the logistic distribution function is used for the function F:

$$P(Y = 1|\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_d \cdot X_d)}{1 + \exp(\beta_0 + \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_d \cdot X_d)}. \tag{3.3}$$

The parameters $\beta_0, \beta_1, \cdots, \beta_d$ ($1 + d$ parameters) are calculated using maximum likelihood estimation (Hastie et al., 2009).

Additionally to the main effects $X_1, \cdots, X_d$ considered in Equation (3.3), interaction effects and effects in quadratic order are sometimes included in the logit model. This leads to:

$$P(Y = 1|\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_d \cdot X_d + \sum_{i=1,j=1,i\leq j}^{d} \gamma_{i,j}\mathrm{X}_i\mathrm{X}_j)}{1 + \exp(\beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_d \cdot X_d + \sum_{i=1,j=1,i\leq j}^{d} \gamma_{i,j}\mathrm{X}_i\mathrm{X}_j)}. \tag{3.4}$$

The parameters to be estimated in this case are $\beta_0, \beta_1, \cdots, \beta_d, \gamma_{11}, \cdots, \gamma_{dd}$ (in sum $1 + 2 \cdot d + \binom{d}{2}$ parameters).

3.5.2 Best subset model

As the number of predictor variables can rapidly increase (as in Equation (3.4)), leading to complex models and probably to the presence of irrelevant variables as well as high correlation among the variables, it is therefore of interest to select the best subset of inputs to include in the logit model. For feature selection we use the LASSO

(Least Absolute Shrinkage and Selection Operator) regularization (Tibshirani, 1996). Here, the negative binomial likelihood and a regularization parameter λ is introduced to penalize unimportant or highly correlated features and shrink their coefficients to zero. This leads to the minimization problem (only main effects are considered for simplification):

$$\min_{\beta_0,\beta\in\mathbb{R}^{d+1}} -\left[\frac{1}{n}\sum_{i=1}^{n} y_i \cdot (\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}) - \log(1+\exp[\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}])\right] \\ +\lambda[(1-\alpha)\|\boldsymbol{\beta}\|_2^2/2 + \alpha\|\boldsymbol{\beta}\|_1], \tag{3.5}$$

over a grid of values of the hyperparameter λ, which controls the overall strength of the penalty. The hyperparameter α ($\alpha = 1$ for LASSO regression and $\alpha = 0$ for Ridge regression) controls the "elastic-net" penalty. LASSO regression uses the L_1-norm and leads to a smaller number of relevant coefficients since it picks only one coefficient from two highly correlated variables and shrinks the other coefficient to zero, while Ridge shrinks these coefficients towards each other. For the analysis, models are calculated using the *glmnet* function from the *glmnet* package (Hastie and Qian, 2014) implemented in R. The *glmnet* algorithm applies cyclical coordinate descent for successive optimisation of the cost function over each parameter until convergence. The *cv.glmnet* function computes several models and evaluates the optimal λ for the model with the lowest error via grid search and cross-validation. The higher the λ, the more coefficients are shrink to zero.

3.5.3 Assessment of model performance

The logit model provides the probability that a student drops out of higher education. It can also be seen as a binary classifier, whereby a student with $P(Y = 1|\mathbf{x}) \geq a$ (a is the threshold value defined by the user or automatically calculated depending on the class size), is classified as a dropout and otherwise as a graduate. The performance of the obtained model is evaluated in terms of the mean squared error (MSE), which is the mean of the squares of the errors between the predicted probability for class 1, i.e. $P(Y = 1|\mathbf{X})$, and the observed variable $Y \in \{0, 1\}$; accuracy, which gives the relative number of correctly classified students and the area under the ROC-curve (AUC).

The true positive rate, also called sensitivity or "Recall", is the number of dropouts, truly classified as dropouts, divided by the total number of dropouts. The false positive rate (or 1−specificity) is calculated as the number of graduates classified as dropouts, divided by the total number of graduates. All these measures depend on the threshold. Varying the threshold from 0 to 1 and plotting the true positive rate against the false positive rate outputs the receiver operating characteristic curve (ROC-curve). The area under the (ROC)-curve, named AUC, is a further important measure for binary classification. A value near 0.5 means that the model chooses randomly the class of a new observation, while a value near 1 means that almost all observations are correctly classified. Furthermore, the computations are done by applying 10-fold cross-validation repeated 20 times as suggested by Krstajic et al. (2014) to reduce the variance of the estimations.

3.5.4 Dealing with missing values in the data

The constructed dataset contains 943 dropouts, 2625 graduates and 53 predictor variables (including the intercept), with a considerable number of values missing in the data (about 18%). Since the logit model requires data with complete cases, these missing values should be handled. In general, three approaches are possible: (1) using prediction methods that can handle missing values (instead of logistic regression), (2) using only complete cases which would delete most observations in our dataset (the dataset would be reduced to 36 observations), and (3) imputation techniques which fill the missing values with plausible values. To find the best imputation technique leading to optimal model performance, the 10-fold cross-validated out-of-sample AUC and MSE were computed for several imputation methods including mean or median imputation, regression imputation, stochastic imputation, hot-deck imputation, and multiple imputation (Batista and Monard, 2003, Twala, 2009, Meeyai, 2016).

The median imputation produces the best results in terms of AUC and MSE. Consequently, for further analysis in this study, the complete dataset obtained with this imputation method is used. Garciarena and Santana (2017) also found situations where median imputation outperforms advanced imputation techniques. This dataset has many dichotomous variables where the median imputation reveals good results in terms of model performance. Of course, in many other applications, the median imputation

might not be optimal. Also note, that median imputation has a decreasing effect on the variance of the imputed variables which also affects confidence intervals and p-values of statistical tests (Kleinke et al., 2020).

3.6 Empirical results

Here, the results of the logit model via LASSO regularization are presented. The predictor variables are divided into 5 groups (as presented in section 3.4) and the response variable (status of student) has value 1 for dropouts and 0 for graduates. After computing the LASSO regularization to select the most prominent variables out of all the variables, we ended up with 22 variables, which have absolutely nonzero coefficients (see the appendix on how the number of selected variables is defined). A logit model based on the 22 selected variables is fitted and the standardized regression coefficients are shown in Table 3.4 along with the pseudo R^2, the cross-validated MSE and AUC. Significance concerning this logit model cannot be interpreted in a conventional way (at face value) due to the prior selection process. Nevertheless, the z-values, in addition to the standardized coefficients, provide important information on the partial effects of the variables. A positive value of a coefficient indicates that a higher value of the corresponding variable increases the probability to dropout. The confusion matrix is also reported in Table 3.5 along with accuracy (proportion of correctly identified students), recall (proportion of dropouts correctly identified) and the average threshold (minimum probability for a student to be classified as dropout), which is automatically calculated by the model. The ROC-curve is plotted in Figure 3.2.

Table 3.4: Standardized regression coefficients of the logit model

Variable	**Estimate**	**Std. Error**	**z-value**	**p-value**
Intercept	−1.363	0.050	−27.417	0.000***
Demographic and family background				
Gender	**0.223**	0.047	**4.742**	0.000***
Mother's occupation	−0.108	0.045	−2.412	0.016**
Place of residence	0.198	0.044	4.471	0.000***
Year of birth	−0.190	0.049	−3.870	0.000***
Financial situation				
Time bud. term break: Emply.	**−0.227**	0.050	**−4.583**	0.000***
Study costs	−0.091	0.044	−2.074	0.038**
BAfoeG	**−0.296**	0.048	**−6.162**	0.000***
Monthly income	−0.128	0.052	−2.475	0.013**
Prior education				
General preparation	**−0.250**	0.048	**−5.193**	0.000***
Overall grade on school	**0.506**	0.048	**10.522**	0.000***

Number of rep. classes	0.177	0.042	4.210	0.000***
Type of school leav. qual.	−0.125	0.059	−2.135	0.033**
Type of high school	−0.081	0.052	−1.555	0.119
Institutional determinants				
General university	**0.603**	0.057	**10.641**	0.000***
Subject (Law)	−0.099	0.057	−1.752	0.079*
Subject (Mathematik)	0.165	0.046	3.562	0.000***
Subject (Engineering)	**0.268**	0.052	**5.128**	0.000***
Motivation and satisfaction				
Satisf. with actual. stu.	**−0.387**	0.047	**−8.209**	0.000***
Compet. rel. extr. mot.	−0.124	0.047	−2.652	0.007***
Degree course wear. down	−0.175	0.049	−3.532	0.000***
Alternative to a degree	**0.241**	0.040	**6.081**	0.000***
Subject of choice satisf.	−0.188	0.041	−4.586	0.000***
$n = 3,568$, Pseudo-$R^2 = 0.318$, MSE= 0.301, AUC= 0.796				

*** p-value<0.01, ** p-value<0.05, * p-value<0.1

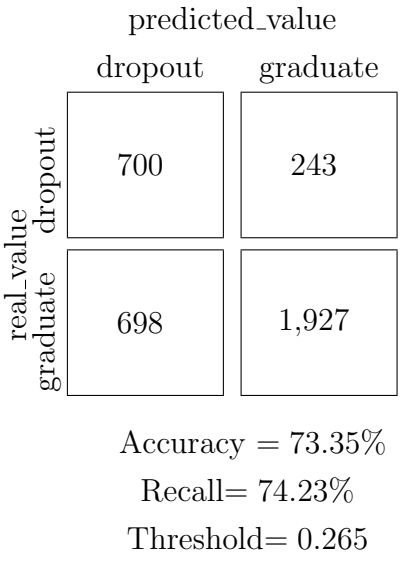

Table 3.5: Confusion matrix

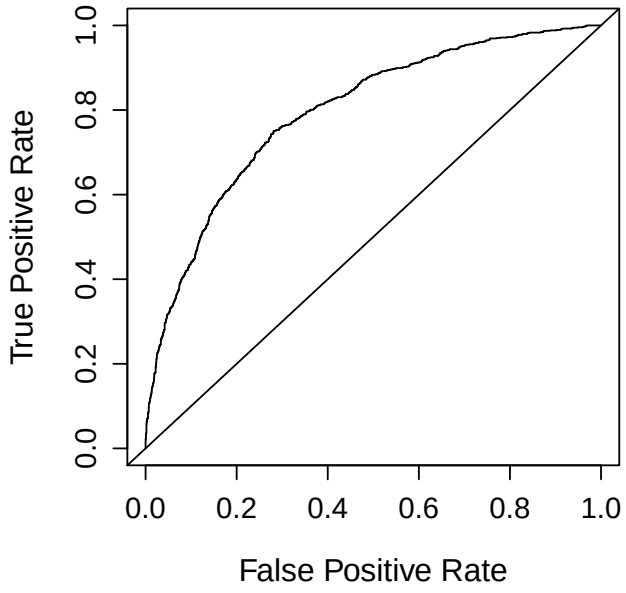

Figure 3.2: ROC-curve

As noted in Table 3.4, many determinants contribute to lower the risk of dropping out. Assuming the non-presence of high colinearity in the data (removed by means of Lasso), values of the standardized regression coefficients outline the relative importance of the predictors (Fox, 2015, Darlington and Hayes, 2016). The largest coefficients and z-values (in magnitude) are in bold, which shows that from each determinant area there are important predictors for student dropout. For instance, being a female student, with good

prior preparation and school grades, studying at a university of applied sciences, being satisfied with the studies, not preferring to do something else instead of studying, and receiving financial aid (BAfoeG) carry a lower risk of dropping out than their counterparts. Moreover, students from Mathematics and Engineering have a higher risk of dropping out, whereas students from Law/Economics sciences are less risky compared to Linguistics and Cultural Sciences. The model achieves a cross-validated MSE of 0.301 and AUC of 0.796. Three-fourths of students are correctly classified and the proportion of correctly identified dropout students amounts to about 75%.[4] The directions of relationships between the covariates and study dropout are mainly in line with the descriptive analysis, theoretical considerations as well as with findings from previous studies (if already analyzed) which are discussed in detail in earlier sections. A counterintuitive result is the direction of the effect of the predictor "Degree course is wearing me down".

Of considerable interest is that besides well known determinants such as school performance, aspects related to students' satisfaction with study are of great importance for academic success. Satisfaction further depends on a student's information and preparation status (Weerasinghe et al., 2017) which become relevant already before or at the beginning of study and lie, up to a certain degree, in universities' (and also secondary schools') scope of action. Therefore, there are many promising starting points for early warning systems for preventing students at risk of dropping out.

3.7 Discussion and conclusion

The current state of empirical research on student dropout from several disciplines has identified numerous possible reasons why students withdraw from tertiary education. This study aims at providing an encompassing evaluation of these determinants and aims at identifying the most important ones by applying bivariate measures of effect size and a multivariate LASSO regression with an internal feature selection to predict the probability of a student to graduating or to dropping out. The analysis is based on a dataset including freshman students, who have started in the winter term 2010/2011 at German institutions of higher education and covering a wide range of different aspects

[4] The predictive performance obtained by the logit model based on the selected variables is acceptable. However, it would be interesting to examine whether potentially quadratic and interaction effects improve the predictive quality of the model. Results of this extended analysis are provided in the appendix.

of students' background and the course of study. In the following, the findings and their possible implications for universities to prevent students from dropping out at an early stage of study are discussed.

From each of the determinant categories there remain important variables in the final prediction model after feature selection (AUC=0.789), which confirms that dropout is a result of several conditions and underlines the complexity of the dropout phenomenon.

3.7.1 Demographic and family background, and prior education

The impact of students' pre-study determinants, such as their prior education and other background determinants, implies that higher education institutions should take into account the increased heterogeneity of students and their specific needs. For instance, a lower educational pathway of students (e.g. type of school leaving qualification; also found by Müller and Schneider, 2013) or a poorer school performance (preparation, school grade, repeated classes; also stated by Sarcletti and Müller, 2011) increase the risk of dropping out. A preferable strategy may be to implement background-specific remediation programs or field-specific bridging courses preparing students for university requirements.

3.7.2 Institutional determinants

Relevant predictors on the institutional level are the type of higher education institution and the field of study. Studying at a general university instead of a university of applied sciences and studying subjects like Mathematics/Natural Sciences and Engineering (also found e.g. by Sarcletti and Müller, 2011, Heublein et al., 2017) seem to increase the risk of dropping out. This observation provides no direct starting point for reducing dropout rates but may point to more structured or practice-oriented study courses (as at universities of applied sciences) to be a relevant determinant of study success. Moreover, the results indicate the usefulness of field-specific intervention measures especially in fields with a high dropout rate.

3.7.3 Financial situation

The financial situation of students, for instance in form of financial aid (BAfoeG in Germany; also found by Glocker, 2011) and the ability to cover living costs (study costs, income), seems to be an important aspect of study success. Here, an improvement of the financial aid system, for example, a higher amount of subsidies, probably decreases the dropout risk for students, especially for those from low-income families.

3.7.4 Motivation and satisfaction with study

Several determinants identified as important for study success are related to student satisfaction (e.g. satisfied with actual studies; also found by Suhre et al., 2007). Regular student surveys to get information on student satisfaction, their wishes, and needs are probably an appropriate first step towards providing a supportive and encouraging environment and thereby increasing satisfaction with studies. Satisfaction highly depends on the gap between students' expectations concerning study content, organization and required qualifications and the real study situation induced by insufficient information and preparation status of students (Suhre et al., 2007, Weerasinghe et al., 2017). Therefore, possible starting points are, for instance, student information days and workshops helping students to get an overview of the different study alternatives early and to find study fields matching their skills and interests. In addition, the implementation of online self-assessment programs for a first overview and evaluation of interests and opportunities may also be useful (Heublein, 2014a). Here, cooperation with secondary schools seems to be of considerable importance (Hetze, 2011). To be able to study the subject of choice, which also seems to have a great impact on study success, early information on formal and content-related requirements may encourage students to obtain these qualifications already at school (e.g. to choose maths as core subject). Moreover, as the fact that students would have preferred to do something else rather than studying (alternative to a degree) is a predictor of dropping out, students should also ponder their non-academic alternatives before starting a study. Here, special offers helping to decide if a study or, for instance, vocational training would better match their aspirations and wishes may prevent student dropout due to discontent and unfulfilled expectations.

Some determinants identified as relevant for the dropout decision are influenceable to a varying degree by institutions or by students themselves whereas others are not. There are many aspects that become relevant already before or at the beginning of the study, such as prior education and also satisfaction, so there are promising starting points for early warning systems.

The findings provide valuable starting points to tackle the dropout phenomenon. However, in the discussions on dropout prevention, it should be kept in mind that a dropout from university may not necessarily be interpreted as a negative event in the educational career. A voluntary dropout may be a sensible revision of a disadvantageous decision allowing students to take a chance with new opportunities and possibilities to find a more appropriate and interesting job instead of persevering in a non-satisfying study program.

3.8 Appendix

How to choose the best lambda and the number of selected variables

A sequence of models for 200 different values of λ ($log(\lambda) \in [-8, -2]$) is fitted and displayed in Figure 3.3.[5] Computation stops if the fraction of (null) deviance explained does not change sufficiently from one lambda to the next (end of the path).[6] Each curve corresponds to a predictor variable and shows the path of its coefficient as λ varies. The number of nonzero coefficients at the current λ, also known as the effective degrees of freedom, is indicated in the axis above. The higher the value of λ, the more the coefficients shrunk to zero.

To select the model that best fits the data, the optimal value of λ should be chosen. This is done by evaluating and comparing the out-of-sample MSE and AUC of each model using the method of cross-validation (number of folds is set to 10).[7] Figure 3.4 shows the results.

[5] Models are fitted using the function *glmnet*.
[6] The deviance is defined as 2*(loglike_sat − loglike), where loglike_sat is the log-likelihood for the saturated model (Friedman et al., 2010).
[7] The function *cv.glmnet* is used for that.

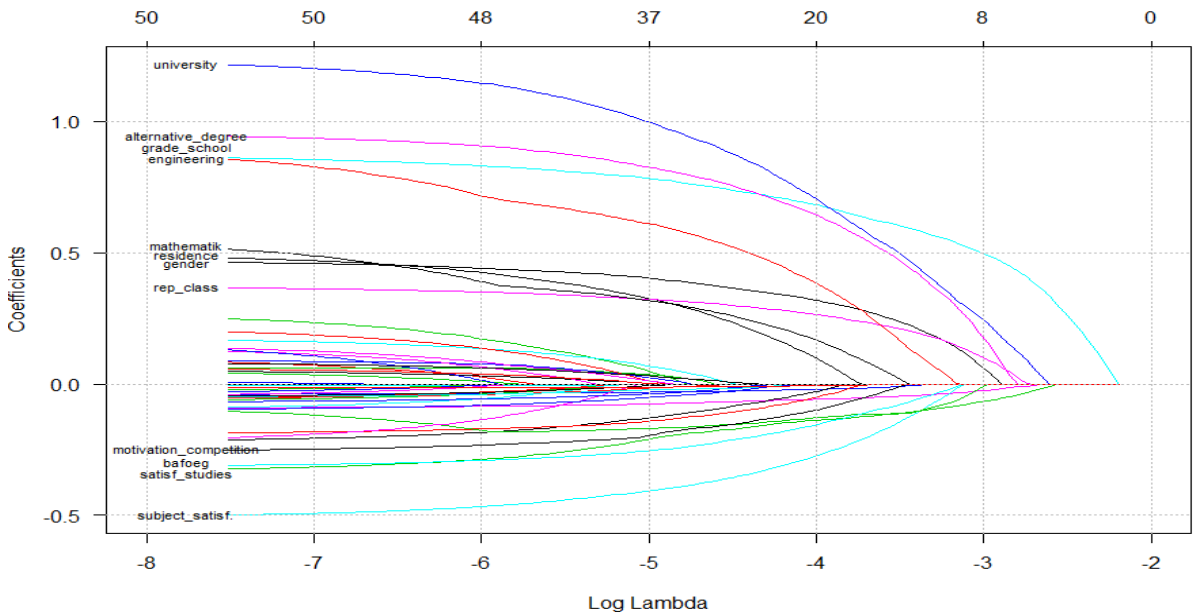

Figure 3.3: Variation of the coefficient of each predictor variable (represented by a curve) as λ varies. Number of nonzero coefficients are indicated in the axis above. Labels of some variables are added.

The graph includes the cross-validation curve (black dotted line in both figures), the upper and lower standard deviation curves along the λ sequence (error bars). These vertical dotted lines indicate the two selected λ's, which correspond to some coefficients, respectively. Left figure: The first line from the left panel of Figure 3.4 provides $\lambda_{min} = 0.0017$, which is the value of λ that gives the minimum mean cross-validated error (here 0.298) and 49 predictors have nonzero coefficients. The second line outputs $\lambda_{1se} = 0.0152$, which gives the most regularized model such that the error is within one standard error of the minimum. This error amounts to 0.307 and 22 predictors have nonzero

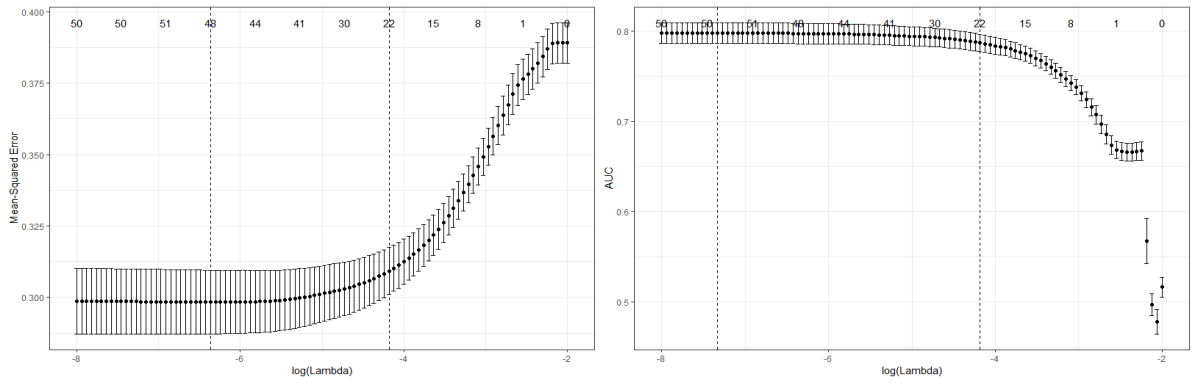

Figure 3.4: Selection of the best λ parameter based on MSE and AUC.

coefficients. The figure on the right provides the best λ based on the AUC. The second vertical line outputs the λ_{1se} also with a value of 0.0152, an AUC value of about 0.789 and 22 nonzero coefficients.

Model improvement

To improve the predictive performance, interaction terms (between the predictor variables) and curvilinear (quadratic) effects are included in the model. Using the selected predictors, the model is (re)-computed and, additionally to main effect terms, terms of quadratic order and interactions within the predictors are considered. This leads to an overall number of 275 variables (22 first order variables + 22 quadratic forms + $\binom{22}{2} = 231$ interactions of the second order). As λ varies, values of MSE and AUC and the number of nonzero coefficients are recorded (Figure 3.5).

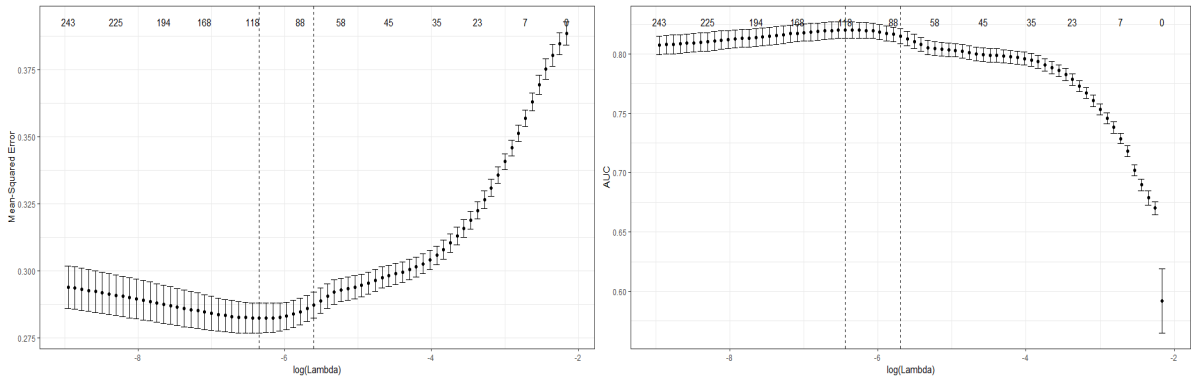

Figure 3.5: Models including main effects, interactions between the predictors and curvilinear effects.

As displayed in Figure 3.5, a slight improvement of the model is noted in both evaluation measures. The MSE improves from 0.309 to 0.282 and the AUC value from 0.789 to 0.821 when the best subset with 78 variables is used. The Accuracy and Recall values also improve, from 73.35% to 76.28% and from 74.23% to 76,35%, respectively. The threshold value, i.e. the minimal probability to be classified as dropout, is $a = 0.268$. These results confirm that considering interactions among the predictor variables and terms of quadratic order in addition to main terms generally improves the predictive performance of the models.

Additionally to the predictive performance, regression coefficients of the predictors are shown in Table 3.7. For convenience, 22 variables are selected as in the prior computed model. It shall be noted that some main effects could be discarded due to the presence of quadratic and interaction effects, which restricts meaningful interpretations of the model. However, this table is primarily to provide an indication on which quadratic and interaction effects are included in the model.

As shown in Table 3.7, only one main effect and one quadratic effect are included in the model, namely alternative to a degree and square of the grade at secondary school. Interaction effects of the grade at secondary school and the institutional determinants are indicated as important predictors. Interaction effects between the satisfaction variables (satisfaction with the studies and satisfaction with the chosen subject) and the financial variables are also important.

Table 3.6: Attribute description

Attribute	Description (Data type)
Demographic and family background	
Gender	Gender of the person (binary: 1 = Male or 0 = Female) Number_Dropouts = 943, Number_Graduates = 2625, wave 1
Father's occupation	Father (stepfather|this person] occupation (ISEI-08) (numeric: from 11.74 to 88.96) Number_Dropouts = 925, Number_Graduates = 2587, wave 1
Mother's occupation	Mother (stepmother|this person] occupation (ISEI-08) (numeric: from 11.74 to 88.70) Number_Dropouts = 925, Number_Graduates = 2587, wave 1
Highest mother's education	Mother's (stepmother's|this person's) highest general school-leaving qualification (numeric: from 0 = No school leaving qualification to 8 = Highest tertiary education) Number_Dropouts = 934, Number_Graduates = 2608, wave 1
Place of residence	In which region of Germany does the student live (binary: West = 0, new eastern federal states = 1 Number_Dropouts = 942, Number_Graduates = 2622, wave 1
Highest father's education	Father's (stepfather's|this person's) highest general school-leaving qualification (numeric: from 0 = No school leaving qualification to 8 = Highest tertiary education) Number_Dropouts = 906, Number_Graduates = 2561, wave 1
Immigration background	(binary: 1, if a person lives in Germany up to the third generation, else 0) Number_Dropouts = 943, Number_Graduates = 2625, wave 1
Year of birth	Year of birth of the person (numeric: from 1946 to 1994) Number_Dropouts = 943, Number_Graduates = 2625, wave 1
Financial situation	
Time budget term break: Employment	Weekly hours spent on employment during term break (numeric: from 0 to 99) Number_Dropouts = 550, Number_Graduates = 2056, wave 2
Study costs	Costs of study: Giving up other, competing goals (numeric: from -1.5 = not apply at all to +1.5 = applies completely) Number_Dropouts = 460, Number_Graduates = 2028, wave 1
Financial aid	Does the student receive financial aid (BAföG)? (numeric: from 0 = never applied for BAföG to 3 = Receive BAföG independent from parental income) Number_Dropouts = 426, Number_Graduates = 2019, wave 2

Monthly income	Sum of monthly income (numeric: from 1 to 11399) Number_Dropouts = 420, Number_Graduates = 1997, wave 2
Time budget semester: Employment	Weekly hours spend on employment (numeric: from 0 to 60) Number_Dropouts = 549, Number_Graduates = 2059, wave 2
Prior education	
General preparation	General preparation for study (numeric: from -9 = bad to +9 = excellent) Number_Dropouts = 601, Number_Graduates = 2074, wave 2
Overall grade on school-leaving qualification	approximate overall grade awarded in the school-leaving certificate (numeric: from 1 to 4.2) Number_Dropouts = 917, Number_Graduates = 2565, wave 1
Number of reapeated classes	Number of repeated classes at secondary school (numeric: from 0 to 4) Number_Dropouts = 942, Number_Graduates = 2625, wave 1
Maths skills acquired before university	(numeric: from -1.5 = not at all to 1.5 = a lot) Number_Dropouts = 493, Number_Graduates = 1704, wave 2
German skills acquired for my studies	(numeric: from -1.5 = not at all to 1.5 = a lot) Number_Dropouts = 534, Number_Graduates = 1945, wave 2
Type of school-leaving qualification	School-leaving qualification obtained (numeric: 2 = general university entrance qualification, 1 = university of applied science entrance qualification, 0 = other degrees) Number_Dropouts = 943, Number_Graduates = 2624, wave 1
Informed about study	(numeric: from -4 = not informed at all to +4 = greatly informed) Number_Dropouts = 926, Number_Graduates = 2621, wave 1
English skills acquired for my studies	(numeric: from -1.5 = not at all to 1.5 = a lot) Number_Dropouts = 599, Number_Graduates = 2062, wave 2
Reading competence	An estimator for the reading competence (numeric: from -3.63 to 4.22) Number_Dropouts = 445, Number_Graduates = 1770, wave 1
Mathematics as core subject	Maths as first examination subject for school-leaving qualification (numeric: 1 = yes, 0 = no) Number_Dropouts = 830, Number_Graduates = 2418, wave 1
Type of high school	Type of school attended (numeric: 1 = upper secondary education, 0 = other types) Number_Dropouts = 913, Number_Graduates = 2552, wave 1
German as core subject	German as first examination subject for school-leaving qualification (numeric: 1 = yes, 0 = no) Number_Dropouts = 830, Number_Graduates = 2417, wave 1
Institutional determinants	
General university	General university attended (numeric: 1 = yes, 0 = no) Number_Dropouts = 943, Number_Graduates = 2622, wave 1
Subject group (Law)	Law, economics and social sciences as chosen subject group (numeric: 1 = yes, 0 = no) Number_Dropouts = 938, Number_Graduates = 2613, wave 1
Subject group (Mathematics)	Mathematics and natural sciences as chosen subject group (numeric: 1 = yes, 0 = no) Number_Dropouts = 938, Number_Graduates = 2613, wave 1
Subject group (Engineering)	Engineering as chosen subject group (numeric: 1 = yes, 0 = no) Number_Dropouts = 938, Number_Graduates = 2613, wave 1
Subject group (Linguistics)	Linguistics and cultural sciences (numeric: 1 = yes, 0 = no) Number_Dropouts = 938, Number_Graduates = 2613, wave 1
Motivation and satisfaction with study	
Satisfied with actual studies	On the whole, satisfied with the current degree course (numeric: from 0 = does not apply at all to 10 = applies completely) Number_Dropouts = 466, Number_Graduates = 2530, wave 3

Enjoy the degree course	Really enjoy the studied subject (numeric: from 0 = does not apply at all to 10 = applies completely) Number_Dropouts = 466, Number_Graduates = 2530, wave 3
Performance related extrinsic motivation	Studying degree course because of completing degree course successfully (numeric: from 1 to 4) Number_Dropouts = 225, Number_Graduates = 2037, wave 5
Degree course is interesting	Find degree course really insteresting (numeric: from 0 = does not apply at all to 10 = applies completely) Number_Dropouts = 466, Number_Graduates = 2530, wave 3
Career related extrinsic motivation	Studying degree course in order to have good career opportunities (numeric: from 1 to 4) Number_Dropouts = 225, Number_Graduates = 2037, wave 5
Competition related extrinsic motivation	Studying degree course in order to be one of the best (numeric: from 1 to 4) Number_Dropouts = 225, Number_Graduates = 2037, wave 5
Alternative to a degree	(binary: 1, if yes, 0, otherwise) Number_Dropouts = 800, Number_Graduates = 2268
Frustrating external circumstances	External circumstances under which degree course is conducted are frustrating (numeric: from 0 = does not apply at all to 10 = applies completely) Number_Dropouts = 466, Number_Graduates = 2530, wave 3
Degree course is wearing me down	Degree course kills me (numeric: from 0 = does not apply at all to 10 = applies completely) Number_Dropouts = 466, Number_Graduates = 2530, wave 3
Subject of choice satisfied	Enrolled in the subject of choice (numeric: 1 = yes, 0 = no) Number_Dropouts = 772, Number_Graduates = 2097, wave 1
Bridging courses	Assessment of participation at bridging courses (numeric: from 1 = not at all helpful to 4 = very helpful) Number_Dropouts = 709, Number_Graduates = 2061, wave 1
Degree course obligations hard to macth	Difficult to reconcile degree course with other obligations (numeric: from 0 = does not apply at all to 10 = applies completely) Number_Dropouts = 466, Number_Graduates = 2530, wave 3
Concerns of students are not taken into account	Not enough attention paid to the concerns of students (numeric: from 0 = does not apply at all to 10 = applies completely) Number_Dropouts = 466, Number_Graduates = 2530, wave 3
Wishing better study conditions	Wish study conditions at university were better (numeric: from 0 = does not apply at all to 10 = applies completely) Number_Dropouts = 466, Number_Graduates = 2530, wave 3
Often tired due to degree course	Degree course often makes feel tired and exhausted (numeric: from 0 = does not apply at all to 10 = applies completely) Number_Dropouts = 466, Number_Graduates = 2530, wave 3
Intrinsic motivation	Studying degree course because of the satisfaction of working with the content (numeric: from 1 to 4) Number_Dropouts = 225, Number_Graduates = 2037, wave 5
Events/forums offered to get to know people	Assessment of participation at events to get to know people (numeric: from 1 = not at all helpful to 4 = very helpful) Number_Dropouts = 709, Number_Graduates = 2061, wave 1
Events/forums concerning study organisation	Assessment of participation at events on study organisation (numeric: from 1 = not at all helpful to 4 = very helpful) Number_Dropouts = 709, Number_Graduates = 2061, wave 1
Higher education institution of choice satisfied	Take up the degree at the university of choice (numeric: 1 = yes, 0 = no) Number_Dropouts = 806, Number_Graduates = 2272, wave 1

Table 3.7: Standardized regression coefficients of the logit model including main effects, interaction and curvilinear effects.

Variable	**Estimate**	**Std. Error**	**z-value**	**p-value**
Intercept	−1.326	0.049	−27.132	0.000***
Alternative to a degree	0.323	0.182	1.778	0.075*
(Overall grade at school)^2	0.385	0.058	6.544	0.000***
Place of residence : General university	0.209	0.043	4.864	0.000***
Gender : Overall grade at school	0.037	0.073	0.502	0.616
Gender : General university	0.204	0.067	3.050	0.002***
Gender : Alternative to a degree	0.065	0.050	1.303	0.193
General preparation : Overall grade at school	−0.214	0.045	−4.730	0.000***
Overall grade at school : Number of rep. classes	0.089	0.065	1.361	0.173
Overall grade at school : General university	0.328	0.071	4.610	0.000***
Overall grade at school : Subject (Mathematics)	0.121	0.050	2.428	0.015**
Overall grade at school : Subject (Engineering)	0.250	0.052	4.846	0.000***
Overall grade at school : Alternative to a degree	−0.124	0.184	−0.673	0.501
Number of rep. classes : General university	0.068	0.063	1.075	0.282
Number of rep. classes : Subject (Mathematics)	0.108	0.054	1.991	0.046**
Type of school leav. qual. : Subject (Law)	−0.122	0.102	−1.195	0.232
Type of school leav. qual. : Subject of choice satisf.	−0.177	0.057	−3.126	0.001***
Type of school leav. qual. : Time bud. term break: Employ	−0.236	0.052	−4.494	0.000***
Subject of choice satisf. : Satisf. with studies	−0.215	0.045	−4.799	0.000***
Subject (Law) : Satisf. with studies	0.149	0.122	1.220	0.223
Subject (Law) : Monthly income	−0.288	0.107	−2.699	0.007***
Compet. rel. extr. mot. : Satisf. with actual. stu.	−0.234	0.055	−4.212	0.000***
Satisf. with actual. stu. : Financial aid	−0.309	0.050	−6.126	0.000***
$n = 3,568$, Pseudo-$R^2 = 0.320$, MSE= 0.300, AUC= 0.800				

*** p-value<0.01, ** p-value<0.05, * p-value<0.1, : means interaction

Table 3.8: Nine waves in the NEPS.

Wave	Method	semester	Time window
1	CATI	1 and 2	10/2010 to 07/2011
2	CAWI	3	10/2011 to 12/2011
3	CATI	4	04/2012 to 07/2012
4	CAWI	5	10/2012 to 12/2012
5	CATI	6	03/2013 to 08/2013
6	CAWI	7	10/2013 to 12/2013
7	CATI	8	04/2014 to 09/2014
8	CAWI	9	10/2014 to 12/2014
9	CATI	10	04/2015 to 08/2015

4 Early prediction of university dropouts - a random forest approach

Early prediction of university dropouts - a random forest approach

Andreas Behr, Marco Giese, Herve D. Teguim K., Katja Theune
Chair of Statistics
University of Duisburg-Essen, 45117 Essen, Germany

Abstract

We predict university dropout using random forests based on conditional inference trees and on a broad German data set covering a wide range of aspects of student life and study courses. We model the dropout decision as a binary classification (graduate or dropout) and focus on very early prediction of student dropout by stepwise modeling students' transition from school (pre-study) over the study-decision phase (decision phase) to the first semesters at university (early study phase). We evaluate how predictive performance changes over the three models, and observe a substantially increased performance when including variables from the first study experiences, resulting in an AUC (area under the curve) of 0.86. Important predictors are the final grade at secondary school, and also determinants associated with student satisfaction and their subjective academic self-concept and self-assessment. A direct outcome of this research is the provision of information to universities wishing to implement early warning systems and more personalized counseling services to support students at risk of dropping out during an early stage of study.

Keywords: student dropout, higher education, dropout prediction, educational data mining, random forest

4.1 Introduction

Study success and failure in tertiary education is an extremely important topic for society, higher education institutions and of course for students themselves. On the one hand, a high percentage of students dropping out exacerbates the lack of highly-qualified individuals on the labor market that is predicted for the next few decades (Vogler-Ludwig et al., 2016). On the other hand, high dropout rates may point to an inefficient use of resources by universities, as well as low-quality teaching, and therefore may damage the reputation of universities. At the individual level, dropping out is often associated with personal failure, an both wasted time and monetary investments (Larsen et al., 2013c).

Dropout rates in tertiary education are very high. In Germany, for instance, 29% of Bachelor students (students beginning in 2010/11) did not finish their degree (Heublein et al., 2017). To reduce dropout rates, universities are increasingly searching for promising measures and programs to identify and help students at risk. The underlying analysis aims at predicting a student's decision to withdraw from university without a degree at an early stage of the study process. The results should serve as a basis for installing more individual dropout-prevention programs, such as expanded information activities, study advice and mentoring programs throughout the degree period

As is evident from previous theoretical and empirical research, dropping out from higher education institutions is a long and complex process during which several determinants accumulate and affect each other reciprocally (Tinto, 1975, 1993). Considering these complex interdependencies adequately in empirical research is challenging and it is hardly possible to identify them all a priori. Hence, it may be more empirically meaningful to apply methods able to search for these interdependencies and patterns in the dropout process without restrictions. Tree-based data mining methods are easily comprehensible and therefore very popular among applied users (Breiman et al., 1984). If the relationship between predictors and the outcome of interest is complex and non-linear, which seems to be the case in the dropout process, decision trees may outperform classical (linear) models (James et al., 2013).

In this analysis, we focus on early prediction of student dropout and observe stepwise the early stages of transition from school to the first semesters at university. First, we

model a student's initial risk before entering university, by including only pre-study determinants, for example, gender, social background, and education. In the next step, we add factors related to the decision phase which are relevant even before the start of their studies (e.g. information sources). In a third step, predictors representing the early study phase of students at the very beginning of study, for instance, participation in specific offers for freshmen students, social integration and commitment to the degree course, are included. This procedure facilitates describing student progress in detail and analyzing how students' starting risk will probably change when interacting with the university environment. The aim is to reveal different starting points for intervention measures and to identify at which point in time a precise dropout prediction is possible.

We observe that adding information from the early study phase to pre-study characteristics substantially increases prediction performance, resulting in an AUC (area under the curve, see section 4.4.2) between 0.83 and 0.88 for all considered study fields. Important predictors include the final grade at secondary school, and also determinants associated with student satisfaction and their subjective academic self-concept and self-assessment. Interestingly, the impact of secondary school performance decreases considerably over the three models, indicating that there are many opportunities to counteract any starting risk and achieve a positive development.

This study is structured as follows. Section 2 provides an overview of previous literature on dropout in higher education with a focus on educational data mining. The data set used for analysis is described in section 3. In section 4, we provide detailed information on the tree-based prediction approach. Section 5 contains the empirical results. Section 6 concludes and discusses starting points for universities to prevent student dropout at an early stage of the study process.

4.2 Literature review

Over the last decades, several theoretical models on higher education non-completion have been developed, originating from different disciplines. They can broadly be divided into psychologically, sociologically, and economically-orientated theories (for more detail, see e.g. Sarcletti and Müller, 2011). Based on these theories, previous empirical studies,

mainly using standard econometric models, identified several possible reasons for dropping out from higher education. They include such demographic determinants as gender (Severiens and Ten Dam, 2012, Johnes and Taylor, 1989, Aina, 2013, Ghignoni, 2017), age (Aina, 2013, Lassibille and Navarro Gómez, 2008, Montmarquette et al., 2001), family background (Smith and Naylor, 2001, Di Pietro and Cutillo, 2008, Aina, 2013), as well as migration background (Belloc et al., 2010, Johnes, 1990). Also important seems to be the pre-study education of students, for instance, the final grade at secondary school (Johnes, 1990, Di Pietro and Cutillo, 2008, Stinebrickner and Stinebrickner, 2014) and the type of secondary school (Müller and Schneider, 2013). Moreover, personal characteristics, like resilience and self-control (Brandstätter et al., 2006, Van Bragt et al., 2011a,b), student motivation (Schiefele et al., 2007) and degree program satisfaction (Suhre et al., 2007) tend to have an influence on study continuation. Important institutional determinants of study success seem to be program organization (Heublein et al., 2017), teaching quality (Georg, 2009), learning environment (Hovdhaugen and Aamodt, 2009), as well as a good relationship between students and teachers (Suhre et al., 2007, Ghignoni, 2017). For more detailed reviews dealing with the international state of dropout research see Sarcletti and Müller (2011), Larsen et al. (2013a), Vossensteyn et al. (2015) and Ulriksen et al. (2010) focusing on STM fields.

Using the same broad German data set as in our analysis, Isphording and Wozny (2018) apply a fixed effects regression and find that studying the field of choice and the final degree at school have the highest importance in explaining the dropout phenomenon. However, they conclude that their standard econometric model seems inadequate for dropout prediction, as all included variables together only explain 12% of the variation (with only 1% explained by observed characteristics). A further and probably more suitable approach would be to apply innovative machine learning algorithms to obtain good predictive power and to address the complexity of the dropout process. Educational data mining is a research field with increasing importance, and several mining techniques have been applied to analyze the university dropout problem. For a review of data mining in education see, for instance, Romero and Ventura (2010, 2013). They point out that educational data mining is not only suitable for turning data into knowledge, but also to further filter and use the knowledge to explain educational phenomena and to derive improvements for student outcomes.

Vandamme et al. (2007) attempt to classify Belgian first-year university students into three risk groups for dropping out: low-risk, medium-risk and high-risk. Most impor-

tant for study success seem to be attendance at courses, the perceived subjective chance of study success, previous academic experience (mainly mathematics), and study skills. They conclude that even though some pre-study factors seem to influence academic careers, there are opportunities to counteract a starting risk and achieve positive development and academic success. No prediction models perform well, with the best result obtained with discriminant analysis (accuracy of 57.35%). Similarly, Kovacic (2010) analyzes dropout from an information systems course at a university in New Zealand, using predictors gathered during the enrollment process (mainly socio-demographic variables). He correctly predicts study outcome with a maximum accuracy of 60.5% for Classification and Regression Trees (CART) and concludes that models based only on enrollment data do not yield good predictions of academic success. Hoffait and Schyns (2017) also limit their analysis on data already available at enrollment to early predict dropout at a Belgian university and increase the prediction accuracy of dropouts by adding an "uncertain" class to the failure and the success class.

For students of an electrical engineering course in the Netherlands, Dekker et al. (2009) observes that classification accuracies using a combined dataset of pre-study and study-related variables are comparable with those achieved on a dataset with university-related data only, and conclude that pre-university data does not add much independent information. Several classification methods such as CART, logit or random forest yield accuracies between 75% and 80%. Yathongchai et al. (2012) come to similar results when comparing determinants before admission, and determinants during study in Thailand.

Siri (2015) addresses dropout prediction in a health care professions degree course at the University of Genoa and focuses on student transition from school to university. Determinants from different areas (demographic, educational, sociological, etc.) seem to be significantly related to academic success. An artificial neural network approach correctly predicts 76% of the dropout cases. Important determinants include family background, educational background, experiences of pre-university guidance, motives for enrollment and interest in the courses.

For Germany, there are a few studies also dealing with the prediction of dropout from higher education institutions. Kemper et al. (2019) apply logistic regression and decision trees to predict student dropout at the Karlsruhe Institute of Technology (KIT), based on examination data and focusing on student progress and performance. They achieve a prediction accuracy of up to 95% after three semesters. The most relevant determinants

are the count/average of passed/failed examinations, the average grade (for models of later semesters) and specific single exams. Also using administrative data - from a private university of applied sciences (polytechnic) and a state university - Berens et al. (2018) apply regression analysis, neural networks, decision trees, and the AdaBoost algorithm to detect students at risk. Similarly to Kemper et al. (2019), they find the model accuracy to improve with increasing semesters of up to 95% (after the fourth semester). Demographic data do not substantially increase model accuracy when also using student performance data.

The abovementioned studies are mainly based on small data sets and restrict their analysis to specific academic fields and/or to one university. Moreover, they mostly do not consider all the possible determinants that are relevant for early prediction of students at risk. But especially large data sets covering a wide range of variables are very well suited for data mining techniques. Currently, there are no broad German studies in this field and the results of international studies are only partly applicable to other countries, as structural differences between national higher education systems also affect student dropout (Heublein, 2014b).

For Germany, the most influential theoretical model describing the dropout process was developed by the DZHW (Deutsches Zentrum für Hochschul- und Wissenschaftsforschung). Their data base is their survey of deregistered students in 87 German higher education institutions (Heublein et al., 2010). In line with Tinto (Tinto, 1975, 1993), the dropout process is divided into three phases, pre-university, within-university, and decision-making. The first phase covers factors representing parental social and educational background, and students' educational background. The authors also point to student preferences and expectations concerning the study program and the study field, which determine educational decisions, and therefore, the whole study process. The second phase covers all relevant internal and external factors during the course of study, in which internal factors can be influenced directly by the student, and external factors are set by universities. Important determinants are students' mental and physical resources, study motivation, study conditions, inherent capabilities, academic and social integration. Moreover, according to the model, external factors outside the university environment affect student dropout decisions, represented by the financing of studies, living conditions, alternatives to the current study, and advice from parents and friends. The decision for or against dropout is made in the third phase of the model, with dropout being a result of incompatibility between internal and external factors (Heublein

et al., 2010, Heublein, 2014b).

Based on this comprehensive theoretical model and the descriptive findings of the studies of the DZHW on dropout determinants, we select a variety of possible predictor variables from all of the different areas for inclusion in our prediction model. We focus on determinants that are already important in the early phase of study, but do not restrict our analysis to enrollment data only, as previous research finds them not to predict academic success very well. The data used for analysis, and more information on the empirical strategy, will be provided in the following sections.

4.3 Data and variables

4.3.1 The National Educational Panel Study

The fifth cohort of the National Educational Panel Study (NEPS) is a comprehensive German panel study including students in tertiary education and covering a wide range of different aspects of student background and the course of study (Blossfeld et al., 2011).[1] Currently, the NEPS cohort 5 contains ten waves which are composed of different survey methods such as a computer-assisted telephone interview (CATI: waves 1, 3, 5, 7, 9 and 10), competency tests, as well as computer assisted web interviews (CAWI: waves 2, 4, 6 and 8). The target population are first-year students (German and non-German) at higher education institutions in Germany in the winter term 2010/2011. Interviewed students must be enrolled for the first time at public or state-approved higher education institutions aiming at a Bachelor degree, state examination (medicine, law, pharmacy, teaching), diploma or Master (roman catholic or protestant theology) or specific art and design degrees (Zinn et al., 2017). The underlying data set contains 17,910 students who participated in the first wave. An important issue in the data set is the variation of the sample size from wave to wave. For an overview of the survey instruments, the number of participants at each wave, temporary panel leavers, the number of observations so far participating for the last time in the respective wave,

[1] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort First-Year Students, doi:10.5157/NEPS:SC5:10.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

as well as final panel leavers (as denoted by the NEPS) of the current scientific use file (SUF) version 10.0.0, see Table 4.11 in the appendix (Zinn et al., 2017, and own calculations).

4.3.2 Student status and predictors

In this section, we describe the dependent variable, the predictor variables, as well as the idea of distinguishing between three relevant (pre-) study phases for analysis.

Let $t = 1, 2, \ldots, 10$ denote the time (in our setting, the wave) and Y_t the status of a student after wave t with three possible outcomes

$$Y_t = \begin{cases} 0, & \text{if the student is still studying and has not graduated yet} \\ 1, & \text{if the student has obtained a first tertiary degree} \\ 2, & \text{if the student has (completely) left the university without a degree.} \end{cases}$$

At the beginning of the survey, each student starts studying, i.e. $Y_0 = 0$. Since we use all the available information up to wave 10 to determine the final status, we write Y without a time index. If a student did not participate in all waves, we use the most recent status information to determine Y.

For the prediction model, student status is addressed using binary classification, in which the binary status of dropout or graduate is considered. Students who are still studying i.e. $Y = 0$, are of no interest in our empirical approach. Some of these students might fall into one of the analyzed categories later in the survey or be defined as panel leavers when information after wave 10 becomes available. The amount of omitted sample persons represents about 44% of the data and might also be caused by panel attrition. To test the robustness of our predictions, we address this issue in sections 4.3.3 and 4.5.5. Furthermore, determinants of graduating or dropping out might change over the course of studies. However, as we focus on early prediction and include features available up to wave 3 (mainly up to wave 2) ($X_{t\leq 3}$) representing the study experiences at the very beginning, this aspect is negligible here. The modeling process can be formalized as follows

$$P(Y = y | X_{t\leq 3} = x), \;\; y \in \{1, 2\}. \tag{4.1}$$

According to Larsen et al. (2013c), the term "university dropout" can simply be defined as leaving a university or a higher education institution without obtaining a degree (early study termination). A distinction should be made according to the level at which dropout occurs. Students may change their field of study (within the same subject area or between subject areas), the type of degree, the (type of) university, or students may leave the university system completely. Depending on the perspective, for instance from a student or faculty perspective, these different types of dropouts may constitute a mere transfer (e.g. from one field to another, "re-selection") or a formal total dropout ("de-selection"). We define dropout as leaving the higher education system completely without a degree. Note that there remains the possibility that a student defined as dropout enters university again later in life. However, we assume this to be very unlikely, as we observed students over ten waves and almost six years and this rarely occurred. Changes of study field, degree or institution are not treated as dropout, but considered in the analysis as predictors.

University graduates are sample persons completing a first degree (e.g. Bachelor, state examination) during the observed period. If there is no graduation during the observed period and the individual did not indicate having abandoned the study, she/he will be considered as "still studying" and is not included in the prediction model.

The dropout predictors are described in detail in Table 4.14 in the appendix, and variable selection is based on findings from previous theoretical and empirical literature as surveyed in section 4.2. Because we aim at predicting university dropout as early as possible, we select information about the predictors up to the third wave and group them into three categories representing the different stages of transition from school to the first 2-3 semesters at university (referring to Siri (2015)). The first category, pre-study, includes "hard" determinants characterizing a student before entering university (e.g. gender, parental background, education) and is observed entirely during the first wave. The second category, the decision phase, includes predictors that become relevant even before the study start, but are related to the chosen program (e.g. subject of choice, information, parent's opinion). Variables of this phase are observed mainly in wave 1, except predictors like study information or preparation variables which are observed in wave 2. The early study phase defines the first three semesters of study, where the first basic courses are attended and students find their way at university (e.g. participation in freshmen programs, commitment to a degree course, social integration, satisfaction). Information on this phase is mainly observed in wave 2. An exception is satisfaction

with study which is observed in wave 3. Based on the three relevant phases, we stepwise build three different prediction models and evaluate which information is important for predicting dropout decisions as precisely as possible. Figure 4.1 illustrates the modeling process with the different stages of analysis.

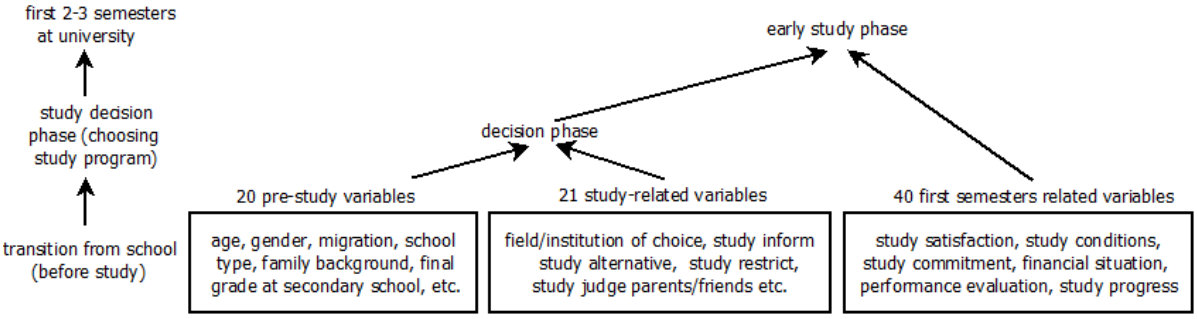

Figure 4.1: Stepwise modeling process of student dropout

4.3.3 Panel attrition

In this section, we present some descriptive findings of panel attrition and retention in our sample.

Panel attrition or panel leaving is a problem in almost every large survey data set (Assaad et al., 2018, Behr et al., 2005). According to NEPS (Prussog-Wagner et al., 2016), students leave the panel by retracting their initial willingness to take part, having not participated in three consecutive CATI interviews (here, participating for the last time in wave 6) or because of other reasons.

Retained individuals are considered as sample members who take part in the interviews at least until wave 7, and have not stated definitely having left the panel for any reason.

Figure 4.2 displays the frequency of these different groups. In total, 5,469 individuals (30.5%) ultimately left the sample and 69.5% remained. The status of some final leavers is observed before they leave the panel. Among the panel leavers, 245 (28% of all dropouts) were identified as university dropouts, 475 (5% of all graduates) as university graduates and 4,749 left the panel before their status was observed.

Panel-leaving can be formalized as follows

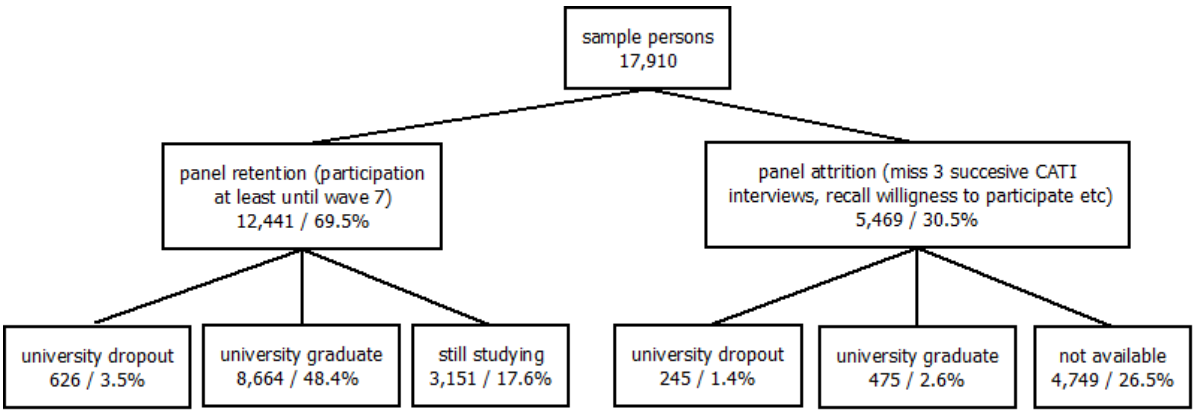

Figure 4.2: Subgroups of sample persons

$$C_t = \begin{cases} 0, & \text{if the student participates in wave } t \\ 1, & \text{if the student finally leaves the panel in wave } t. \end{cases}$$

Similar to the variable Y, we omit the time index, i.e. we distinguish only between final panel leavers in any wave ($C = 1$) and retained persons ($C = 0$). In Figure 4.2, we observe that the sample contains 871 dropout students i.e. 626 who stayed in the panel ($Y = 2 \wedge C = 0$) and 245 who left the survey ($Y = 2 \wedge C = 1$). Furthermore, of the 9,139 graduate students, 8,664 stayed in ($Y = 1 \wedge C = 0$) and 475 left ($Y = 1 \wedge C = 1$) the panel, 3,151 students continued to study ($Y = 0 \wedge C = 0$) and 4,749 individuals left the panel without their status becoming clear (missing $Y \wedge C = 1$).[2]

A comparison of panel leavers and retained persons according to their status, gender and over the four main study fields is displayed in Table 4.1. We observe only a minor difference between the proportion of panel leavers identified as dropout students among males (1.6%) and among females (1.2%) and between the proportion of retained persons identified as dropout students among men (4.1%) and among women (3.1%). Furthermore, only a small difference is noted between male and female graduate students who

[2]The dropout rate in the data amounts to 5% of all sample members and to about 10% of all graduates. This seems to be underestimated, compared to the dropout rate found in Heublein et al. (2017) (29%). Reasons explaining this are twofold. First, the dropout rate is calculated from a panel of first-year students, in which those who abandon the study program are recorded. In the cited literature, the dropout rate is determined on the basis of various data sources, in which the cohort of graduates is compared with the corresponding freshmen-year cohorts using the HIS (Hochschul Information System) procedure. For more details, see Heublein et al. (2012). They further recognize that this procedure is not identical to the immediate tracing of the study courses. Secondly, we assume that panel leaving contributes to this underestimation, since dropouts have a higher probability of leaving the panel than graduates.

leave the sample (2.2% vs. 3%), and additionally, between male and female graduate students who remain in the panel (46.7% vs. 49.5%). In total, the proportion of female students who remained in the panel, is 0.8 percent points higher compared to men.

For a field-specific analysis, we use the four subject areas including the greatest number of students and dropouts. In line with previous research, we find the highest dropout rates for engineering (6.9%) and mathematics/natural sciences (5.7%). The highest rate of panel leavers and retained persons among dropout students is observed in engineering (2.1%, 4.8% respectively). The highest rate of panel leavers for graduate students is observed in law/economics/social sciences (3.2%), and the highest rate of retained persons in mathematics/natural sciences (50.8%). In sum, law/ economics/social sciences students are slightly more prone to leave the panel and mathematics/natural sciences students to remain in the panel. An estimation of the impact of the panel leaving problem on our predictive models is carried out in section 4.5.5.

Table 4.1: Status of panel leavers and retained persons per gender and per study field in % (at the time of the current available wave 10)

	dropouts		**graduates**		**others**		**total**	
	leavers	retained	leavers	retained	leavers	retained	leavers	retained
gender								
male	1.6%	4.1%	2.2%	46.7%	27.2%	18.2%	31%	69%
female	1.2%	3.1%	3%	49.5%	26%	17.2%	30.2%	69.8%
study field								
engineering	2.1%	4.8%	2.5%	49.7%	27.9%	13%	32.5%	67.5%
mathematics, natural science	1.5%	4.2%	2.5%	50.8%	24.8%	16.1%	28.9%	71.1%
law, economics, social sciences	1.2%	3.1%	3.2%	49.5%	29.4%	13.6%	33.9%	66.1%
linguistics, cultural studies	1.2%	3.3%	2.5%	46.6%	25.6%	20.8%	29.3%	70.7%

4.4 Methodology

4.4.1 Conditional inference trees and forests

Decision trees are a top-down binary data-splitting approach, which are popular because of their easy interpretability and low bias. Decision trees can also handle missing values using so-called surrogate splits. If a variable is missing for a specific observation, another predictor variable is used, such that this split is similar to the best split (Twala, 2009). Bootstrap aggregation (bagging) reduces the high variance of a single decision tree by aggregating B trees, fitted with B different bootstrap samples from the data. Calculating the mean value of B single decision trees when using a varying number of features at each split is referred to as random forest (Breiman, 2001, Breiman and Cutler, 2004).

Conditional inference trees, introduced by Hothorn et al. (2006), use a non-parametric permutation test for the binary splits, testing the null hypothesis of independence between the response variable Y and the covariates $\mathbf{X} = \{X_1, \dots X_p\}$. They have the advantage of distinguishing between a significant and an insignificant improvement of the information criterion and further avoid a selection bias towards covariates with many possible splits (numeric or multi-categorical) and many missing values (Hothorn et al., 2006), both of which are present in our data set.

Conditional inference forests can be constructed similarly to Breiman's original random forest approach and also be used to calculate variable importance rankings. Strobl et al. (2007) recommend fitting the forests with subsampling (without replacement) instead of bootstrapping (with replacement) to yield an unbiased variable importance ranking, whereas traditional classification and regression trees (CART) (Breiman et al., 1984) prefer variables with many categories. Moreover, we apply the approach of Hapfelmeier et al. (2014), to construct variable rankings that are unbiased in the presence of missing values.

In contrast to other classification models, such as support vector machines, neural networks, naive Bayes or the well known logistic regression (Aggarwal, 2015, Hastie et al., 2009), tree-based classifiers can handle missing values and various types of variables (metric, ordinal and nominal scaled variables). Furthermore, they are robust against outliers, since only the split point in each node is of interest and not the distance to the split point. Using, for example, a standard logit model, which is not able to

handle missing values, either requires a complete case analysis where all observations with any missing value are deleted or an imputation of the missing data (Aggarwal, 2015). The complete case analysis is not recommended here, because 69% of individuals in the pre-study scenario and 92% in the early study phase have missing values for at least one variable (large-scale imputation). An imputation of missing values can also be problematic, because it stands to reason that some missing values are not missing at random, which leads to biased estimates (Baraldi and Enders, 2009, Twala, 2009).

In addition, tree-based methods have the reputation of imitating the way decisions are made more closely than other regression or classification methods. Furthermore, they are very suitable in the presence of high-dimensional data, high-order interactions and correlations between predictors (Hapfelmeier et al., 2014), in avoiding overfitting, and may outperform classical (linear) models (James et al., 2013).

For our computations, we use the statistic software R (R Core Team, 2019). Conditional inference trees and forests are implemented in the "party" package by Hothorn et al. (2018), which is also used to compute the variable importance.

4.4.2 Measures of predictive performance

A confusion matrix, displayed in Table 4.2, gives an overview of the number of correctly and falsely classified positives and negatives. Positives are the class of interest (Han et al., 2011), which is here the status "dropout". In contrast, negatives are "graduates" in our analysis.

Table 4.2: Confusion matrix

		predicted class	
		positive	negative
true class	positive	true positives (TP)	false negatives (FN)
	negative	false positives (FP)	true negatives (TN)

Based on the values contained in a confusion matrix, several performance measures could be derived (Han et al., 2011). The recall or true positive rate (TPR) denotes the fraction of correctly classified positives:

$$TPR = \frac{TP}{TP + FN}. \tag{4.2}$$

The false positive rate (FPR) is the proportion of negatives that are incorrectly classified as positives:

$$FPR = \frac{FP}{FP + TN}. \tag{4.3}$$

The precision (Pre) is the fraction of correctly classified positives of all classified positives:

$$Pre = \frac{TP}{TP + FP}. \tag{4.4}$$

The accuracy (A) gives the overall fraction of correctly classified observations. Without changing the costs or the threshold (see section 4.4.3), classification methods aim to maximize the accuracy or to minimize the misclassification error ($1 - A$) (Chen et al., 2006):

$$A = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.5}$$

In binary classification settings, receiver operating characteristic (ROC) curves are a useful graphical tool for visualizing the trade-off between TPR, plotted on the y-axis, and FPR plotted on the x-axis. Each point on the ROC-curve belongs to a different threshold (Chen et al., 2006). The area under the ROC-curve (AUC) is a popular performance measure of unbalanced binary classification settings, since the AUC is not influenced by class sizes. An AUC value of 1, where the ROC-curve passes through the point (0,1), represents a perfect classifier, whereby a value of 0.5 (the diagonal) represents a random guess (Han et al., 2011, James et al., 2013).

4.4.3 Model specifications

A graphical instrument for finding an appropriate number of trees (B) is to plot the AUC for different tree numbers. The tuning parameter B should be sufficiently large to reduce the error rate until it stabilizes. A larger B would not further increase predictive power, but would rather increase the computational time. The plot of the overall pre-study model is displayed in Figure 4.5 in the appendix (solid black line). Plots for the two other models and the four subject groups resemble the illustrated plot. The classification performance is poor for less than 20 trees. The curve flattens from about 50 trees. Our choice of 100 trees should be sufficient to ensure the best performance.

Depending on the classification problem, the "cost" of misclassifying observations is probably higher for specific classes than for others. To address this, we introduce a loss matrix $\mathbf{L}$ with $L_{kk'}$ denoting the cost of classifying an observation to class $k^{'}$, although it belongs to class k. A correct classification should not be penalized, i.e. $L_{kk} = 0$ (Hastie et al., 2009). In this setting, we use $\mathbf{L} =$

	predicted dropout	predicted graduate
true dropout	$L_{0,0} = 0$	$L_{0,1} = n_1$
true graduate	$L_{1,0} = n_0$	$L_{1,1} = 0$

Thereby, n_0 is the number of dropouts and n_1 is the number of graduates in the specific scenario. Since the graduates are strongly over-represented ($n_1 \gg n_0$), we assume that a dropout student wrongly classified as graduate is more expensive than a graduate wrongly classified as a dropout. This improves the recall from a value near zero to an acceptable level, but in consequence, the precision decreases. Note that the values in the loss matrix can be adjusted by decision-makers, depending on their opinion on the costs of a study dropout. According to Elkan (2001), in a bivariate classification problem, a threshold τ can be calculated using the 2×2 cost matrix $\mathbf{L}$:

$$\tau = \frac{L_{1,0} - L_{0,0}}{L_{1,0} - L_{0,0} + L_{0,1} - L_{1,1}.} \tag{4.6}$$

In our scenario, we set $\tau = n_0/(n_0+n_1)$, for instance, for engineering students $\tau \approx 0.117$, which means that a student is classified as dropout, if the dropout probability, calculated by the classification rule, is higher than 11.7%. According to Chen et al. (2006), this is an optimal threshold if the class probability is equal to the sample proportion.

Strobl et al. (2008) suggest evaluating different values for the number of variables m used at each split, especially if the variables are correlated. Therefore, a CV (cross-validated) -based search with varying values of m is applied. For small m (approximately until $m \approx \sqrt{p}/2$, whereby p denotes the number of all covariates) the random forest provides poor results in terms of the AUC. As stated by Svetnik et al. (2003), the best results can be found near $m \approx \sqrt{p}$, so that we choose $m = [\sqrt{p}]$. For the pre-study model with all subject groups, the results are displayed in Figure 4.5 (grey dashed line, upper scale).

We pre-selected our variables only from a theoretical point of view and do not use any further statistical variable selection approach. A variable selection based on the full dataset would lead to an underestimation of the CV error rate, because the test set has already "seen" the predictors and is therefore not completely independent (Hastie et al., 2009). Although the number of variables is relatively high, the results of our classification method do not suffer from over-fitting.[3]

4.5 Empirical results

In this section, we present the results of the three different fitted models to describe the transition from school to the first semesters at university. We identify the most relevant dropout predictors and evaluate predictive performance as well as performance improvements when including additional information. For each input setup, we fit a model on the full data and additionally specific models within the main study fields (engineering, mathematics/natural sciences, law/economics/social sciences and linguistics/cultural sciences). The predictive performance of each model is evaluated in terms of AUC, accuracy, recall and precision. Additionally, we provide a ranking of the features for each input setup used in the different models, based on the calculated (relative) variable importance.

An example of a single conditional inference tree is shown in Figure 4.6 for engineering students in the early study phase. Here, to represent the conditional inference tree, we reduce the depth of the tree by setting the significance level of the permutation test, explained in the first step of section 4.4.1, to 0.005 (the default is 0.05).

4.5.1 Dropout prediction with pre-study data

The first step is to predict dropout decisions by using only pre-study variables, that is, variables characterizing a student before he/she enters higher education (e.g. gender, parental background, secondary education, see Table 4.14). Random forests based on

[3] We tested this by applying a variable selection in each CV-loop, and using only the best $6, 7, \ldots$ variables. The results were much poorer when using less than ten variables and no better than the full model when using more than ten variables. Furthermore, this procedure leads to a different set of variables in each CV-loop, resulting in a much more complicated model interpretation.

conditional inference trees with the parameters specified in section 4.4.3 are fitted by computing a 10-fold cross-validation procedure, in which each procedure is repeated 20 times, using different train/test partitions of the data.

Table 4.3: Predictive results for the pre-study model

	dropout	**graduate**	**AUC**	**accuracy**	**recall**	**precision**
full data	871	9,139	0.77	72.00%	71.07%	19.50%
engineering	179	1,352	0.77	75.11%	73.74%	28.32%
mathematics and natural sciences	211	1,962	0.79	73.26%	77.72%	23.50%
law, economics and social sciences	199	2,443	0.75	72.75%	76.38%	18.42%
linguistics and cultural sciences	221	2,398	0.74	70.00%	68.33%	17.32%

Predictive performance: The predictive performances of the different models are provided in Table 4.3. The AUC values obtained are all greater than 0.74, which indicates that the models yield good predictive performance. Using only pre-university data, each model reached an average accuracy of about 73%. In the same scenario, Dekker et al. (2009) obtained predictive results of 69% on average. Prediction worked very well in mathematics/natural sciences with an AUC value of almost 0.80. The recall value (proportion of dropout students who were correctly identified) reached a value over 70% in almost all the different models and also confirms that the fitted classifiers are good at identifying dropouts among the data.

According to the precision values, only about one-fifth of the predicted dropouts are true dropouts. However, since our main aim is to correctly identify dropouts among the data, while maintaining the resulting accuracy as high as possible, we focused on the recall values and rather accept erroneously classifying a graduate as dropout, than a dropout as a graduate.

Variable importance: Table 4.4 presents the relative importance (in percentage) of the most relevant predictors, which sum to 90% - 100%, depending on the study area. As expected and in line with previous research, the final grade at secondary school (grade_school) has by far the highest relative impact (from 46.53% to 65.50%) in the general model as well as in those within each subject field. More precisely, the final grade at secondary school has the greatest relevance for the dropout decision in engineering

and the lowest in law/economics/social sciences. The year of birth is highly relevant in law/economics/social sciences and in linguistics/cultural sciences, with a lower impact in engineering, and mathematics/natural sciences. The number of repeated classes seems to be important in each model, with the highest impact in mathematics/natural sciences. Other relevant determinants in almost all models are the type of secondary school attended and the type of school-leaving qualification, which defines the obtained higher education entrance qualification[4]. Interestingly, parental educational background and gender play only a minor role. This finding is in line, for instance, with Vandamme et al. (2007). On the contrary, previously obtained skills seem not to be as important as in previous research.

Table 4.4: Relative importance of the input variables (pre-study)

	relative importance
full data	grade_school: 58.53%, birthyear: 12.34%, rep_class: 9.72%, school_type: 4.53%, qualif_max: 3.13%, math_prep: 2.78%, gender: 1.94%
engineering	grade_school: 65.50%, school_type: 9.43%, rep_class: 5.78%, qualif_max: 5.65%, birthyear: 5.07%, math_prep: 4.16%, exam_adv_german: 4.12%
mathematics and natural sciences	grade_school: 60.38%, rep_class: 16.00%, birthyear: 8.36%, school_type: 4.98%, exam_german: 1.98%, qualif_max: 1.49%, father_qualif: 1.17%
law, economics and social sciences	grade_school: 46.53%, birthyear: 21.20%, qualif_max: 7.28%, school_type: 5.22%, rep_class: 5.16%, voctrain: 4.19%, father_qualif: 3.85%
linguistics and cultural sciences	grade_school: 51.72%, birthyear: 26.68%, rep_class: 8.37%, qualif_max: 6.54%, gender: 5.80%

4.5.2 Dropout prediction with data related to pre-study and decision phase

In a second step, we expanded the pre-university dataset by adding some decision-phase-related variables representing the phase after school (see Table 4.14). These variables become relevant even before the study start, but are related to the chosen program (e.g. subject of choice, information, parents' opinion).

Predictive performance: The predictive performances of the different models are shown in Table 4.5. We observe a slight improvement in predictive power (AUC) and the recall value, compared to the results of the previous section (pre-study data), which

[4]In Germany, the highest qualification (A-level) enables the student to enter all tertiary education institutions. The restricted A-level (typically after leaving school 1 year earlier) in the first place allows access to universities of applied sciences only. Moreover, a lower school leaving certificate and additional apprenticeship training/schooling also entitles students to enter higher education.

indicates that the variables related to the decision phase only add sparsely independent information for the dropout prediction.

Table 4.5: Predictive results for the model related to pre-study and decision phase

	dropout	**graduate**	**AUC**	**accuracy**	**recall**	**precision**
full data	871	9,139	0.78	72.50%	74.80%	20.00%
engineering	179	1,352	0.77	72.63%	77.65%	26.83%
mathematics and natural sciences	211	1,962	0.81	76.34%	79.00%	25.68%
law, economics and social sciences	199	2,443	0.75	74.30%	71.76%	18.42%
linguistics and cultural sciences	221	2,398	0.76	73.35%	72.00%	20.00%

Variable importance:

The selected variables in Table 4.6 represent an overall impact that ranges from 80% to 90%. Although considerably reduced, the final grade at secondary school always has the highest impact (about 40%) in the general model as well as in the models within each subject field. Among the variables related to the decision phase, only whether students would have preferred to start something else instead of a university study (study_alternative), enrollment in the subject of first choice (fieldofchoice), and study with admission restrictions (study_restrict) contribute to dropout prediction.

Table 4.6: Relative importance of the input variables (pre-study+decision phase)

	relative importance
full data	grade_school: 43.92%, birthyear: 10.38%, rep_class: 9.22%, study_restrict: 6.13%, study_alternative: 5.21%, school_type: 3.55%, qualif_max: 3.50%, fieldofchoice: 1.81%
engineering	grade_school: 45.42%, school_type: 8.61%, birthyear: 5.41%, qualif_max: 4.60% math_prep: 4.58%, study_alternative: 4.32%, exam_adv_german: 4.09%, rep_class: 3.93%
mathematics and natural sciences	grade_school: 43.11%, rep_class: 14.05%, birthyear: 5.33%, school_type: 5.20%, study_alternative: 4.00%, study_restrict: 4.00%, fieldofchoice: 2.96%
law, economics and social sciences	grade_school: 37.75%, birthyear: 19.66%, qualif_max: 6.84%, study_restrict: 4.85%, father_qualif: 4.78%, rep_class: 3.53%, school_type: 3.25%, fieldofchoice: 3.11%
linguistics and cultural sciences	grade_school: 36.40%, birthyear: 15.68%, rep_class: 7.38%, gender: 7.07%, study_restrict: 5.58%, qualif_max: 4.43%, voctrain: 3.78%, study_alternative: 3.73%

4.5.3 Dropout prediction during the early study phase

In the last step, we aim at analyzing our complete data set, which includes the two previous data sets (pre-university data and that related to the decision phase) and new data characterizing the early study phase of students at university. The early study phase represents the beginning of study and how students "get along" with their studies and integrate into the university, for example, through participation in freshmen programs, their commitment to the degree course, social integration, or satisfaction with study (see Table 4.14).

Predictive performance: Table 4.7 reveals that the early study phase variables add independent information that substantially improves the predictive performance. Each model has a very high AUC value, which ranges between 0.83 and 0.88. The recall measures also reach values between 81% and 89%, suggesting that over 80% of dropout students are correctly identified by our fitted models. Particularly in mathematics and the natural sciences, the model correctly identifies almost 90% of dropouts.

Table 4.7: Predictive results for the early study phase model

	dropout	**graduate**	**AUC**	**accuracy**	**recall**	**precision**
full data	871	9,139	0.86	70.62%	84.35%	20.75%
engineering	179	1,352	0.86	78.84%	86.87%	32.94%
mathematics and natural sciences	211	1,962	0.88	78.32%	88.89%	28.90%
law, economics and social sciences	199	2,443	0.83	74.57%	81.33%	22.56%
linguistics and cultural sciences	221	2,398	0.83	75.50%	80.77%	22.56%

Variable importance: The selected variables (about 20) in Table 4.8 represent an overall impact of at least 80%. The final grade at secondary school remains the most important variable in each model, although its impact declines substantially to an average value of 20%. Some pre-university determinants remain relevant in the general model as well as within the study field, for instance, the year of birth, the number of repeated class years (remains very important in mathematics and the natural sciences), the type of secondary school attended, the type of school-leaving qualification obtained, etc.

There are two particularly interesting findings. First, only two of the study-oriented determinants (alternative to study - study_alternative and study restrictions - study_restrict) appear among the best selected variables, confirming that this group of variables makes only a small contribution to dropout prediction. Second, determinants belonging to the early study phase are very informative in each model, since they contribute substantially to the prediction performance. They include determinants describing students' subjective self-assessment of success (e.g. perception of talent for studying - selfconcept, opinion on the probability of graduating - probsuccess, helplessness in obtaining better grades - helplessness), determinants describing satisfaction with studies (e.g. satisfaction with the actual studies - satisf_whole, really enjoy the studied subject - satisf_enjoy, wanting better study conditions - satisf_conditions), determinants describing one's own evaluation of study performance (study progress match to the curriculum plan - workload_match, satisfaction with academic performance - performance_eval), determinants describing commitment to study (not do more than necessary - commit_necessary, high demands on self - commit_demands), and finally determinants describing time spent on employment during semester time (job_semester). Therefore, similarly to Vandamme et al. (2007), we conclude that even though some pre-study factors seem to influence the academic career, there are also opportunities to counteract any adverse starting effects in the direction of positive development and academic success.

4.5.4 Model comparison

To visually compare the predictive performances of the three full data models (see Table 4.9; this serves as an example and is true also within the specific study fields), Figure 4.3 below displays the ROC-curves for each model.

We observe only a slight improvement when only data related to the decision phase are added to the pre-study variables, but a strong improvement in the AUC values when all the determinants are used. This indicates the substantial importance of the early study phase for predicting dropout well as soon as possible. Moreover, an examination of the recall value illustrates the minor increase in the proportion of identified dropouts when decision-phase variables are added into the data, and the large increase of this proportion when early study-phase data are used.

Table 4.8: Relative importance of the input variables (full data)

	relative importance
full data	grade_school: 17.63%, probsuccess: 7.34%, satisf_whole: 5.33%, birthyear: 4.58%, job_semester: 4.49%, performance_eval: 4.11%, workload_match: 3.70%, helplessness: 3.28%, rep_class: 3.00%, satisf_enjoy: 2.87%, commit_demands: 2.76%, study_restrict: 2.69%, satisf_interesting: 2.20%, qualif_max: 2.04%, selfconcept: 1.95%, commit_energy: 1.80%, commit_necessary: 1.70%, school_type: 1.70%, satisf_frustrating: 1.66%, satisf_kill: 1.65%, preparation: 1.63%, satisf_match: 1.60%
engineering	grade_school: 18.64%, probsuccess: 9.22%, helplessness: 9.00%, satis_whole: 6.35% selfconcept: 3.94%, satisf_concerns: 3.84%, performance_eval: 3.50%, qualif_max: 2.85%, school_type: 2.71%, workload_match: 2.65%, costs_direct: 2.56%, preparation: 2.12%, job_semester: 2.11%, birthyear: 2.10%, commit_demands: 1.91%, satisf_enjoy: 1.40%, study_restrict: 1.26%, satisf_frustrating: 1.25%, satisf_kill:1.12% satisf_conditions: 1.10%, commit_identificat: 1.04%, study_alternative: 1.03%
mathematics and natural sciences	grade_school: 24.01%, probsuccess: 7.25%, rep_class: 6.91%, job_semester: 4.20%, workload_match: 4.15%, satisf_whole: 4.02%, satisf_conditions: 3.80%, performance_eval: 3.61%, school_type: 3.35%, birthyear: 3.30%, commit_necessary: 2.84%, study_alternative: 2.57%, satisf_enjoy: 2.11%, helplessness: 2.00%, selfconcept: 1.89%, socint_students: 1.78%, study_restrict: 1.60%, satisf_interesting: 1.34%
law, economics and social sciences	grade_school: 19.30%, probsuccess: 7.25%, satisf_whole: 6.86%, job_semester: 4.68%, performance_eval: 4.65%, birthyear: 4.60%, workload_match: 4.00%, satisf_enjoy: 3.47%, helplessness: 3.43%, rep_class: 2.85%, commit_demands: 2.50%, study_restrict: 2.41%, commit_energy: 1.94, commit_necessary: 1.86%, selfconcept: 1.78%, qualif_max: 1.70%, study_alternative: 1.51%, school_type: 1.46%, satisf_kill: 1.45%, satisf_match: 1.38%, satisf_frustrating: 1.37%, preparation: 1.20%
linguistics and cultural sciences	grade_school: 19.51%, probsuccess: 7.04%, satis_whole: 5.60%, job_semester:4.77% performance_eval: 4.76%, birthyear: 4.69%, workload_match: 3.42%, helplessness: 3.23%, rep_class: 3.00%, satisf_enjoy: 2.89%, commit_demands: 2.50%, study_restrict: 2.28%, study_alternative: 1.87%, qualif_max: 1.85%, selfconcept: 1.80%, preparation: 1.72%, school_type: 1.67%, commit_energy: 1.61%, satisf_frustrating: 1.60%, satisf_match: 1.56%, commit_necessary: 1.51%, commit_identificat: 1.38%

4.5.5 The problem of panel leaving

In this section, we provide an estimation of the impact of the panel-leaving problem on our predictive models. An investigation of the changes in the number of panel leavers over the ten waves, as depicted in Figure 4.4, reveals that dropout students generally leave the panel around waves 4, 5 and 6. Graduate students start leaving the panel from wave 6 onwards, probably after completing their degree program. Among students with unknown status, a fairly high number (about 1,300) leave the sample just after the first wave. From wave 2 up to wave 6, a further 500 students on average in each wave stopped participating in the survey without informing about their status. Unfortunately,

Table 4.9: Predictive results for the three general models

	dropout	**graduate**	**predictor variables**	**AUC**	**recall**
pre-study (pre-university)	871	9,139	21	0.77	71.07%
pre-study+decision phase	871	9,139	41	0.78	74.80%
pre-study+decision phase+early study phase	871	9,139	81	0.86	84.35%

there is no appropriate way to find out the real status for this third category of panel leavers.

A possible approach to addressing this problem could be a wave-by-wave analysis of similarities between panel leavers with unknown status and, firstly, panel leavers who are dropout students, and secondly, panel leavers who are graduate students. This analysis would, however, only be possible for wave 6, since a reasonable amount of panel leavers in each of the three categories is observed during this wave.

Behr (2006) compared various strategies, in order to resolve the panel-leaving problem in the European Community Household Panel and discovered that this issue has only minor effects on estimation, and correcting strategies can reduce the estimation bias on costs of higher variance. For the underlying analysis, the panel-attrition problem could be formalized as follows. Recall equation 4.1 from section 4.3.2; using variable C indicating whether a student leaves the panel (see section 4.3.3), we can write

$$\begin{aligned} P(Y = y|X_{t\leq 3} = x) &= P(Y = y, C = 0|X_{t\leq 3} = x) \\ &\quad + P(Y = y, C = 1|X_{t\leq 3} = x) \\ &= P(Y = y|C = 0, X_{t\leq 3} = x) \cdot P(C = 0|X_{t\leq 3} = x) \\ &\quad + P(Y = y|C = 1, X_{t\leq 3} = x) \cdot P(C = 1|X_{t\leq 3} = x), \end{aligned}$$

with $Y = 1$ if the student has obtained a first tertiary degree and $Y = 2$ if the student has left higher education without a degree (during the observation time).

Three of the four terms can be estimated, but $P(Y = y|C = 1, X_{t\leq 3} = x)$ cannot plausibly be calculated, since Y remains undefined for some sample members when they leave the panel (26.5%). The results in the previous sections were calculated ignoring this subsample population, since we only considered students with a defined Y. However, we assume that there could be a bias in the former predictions due to an underestimation

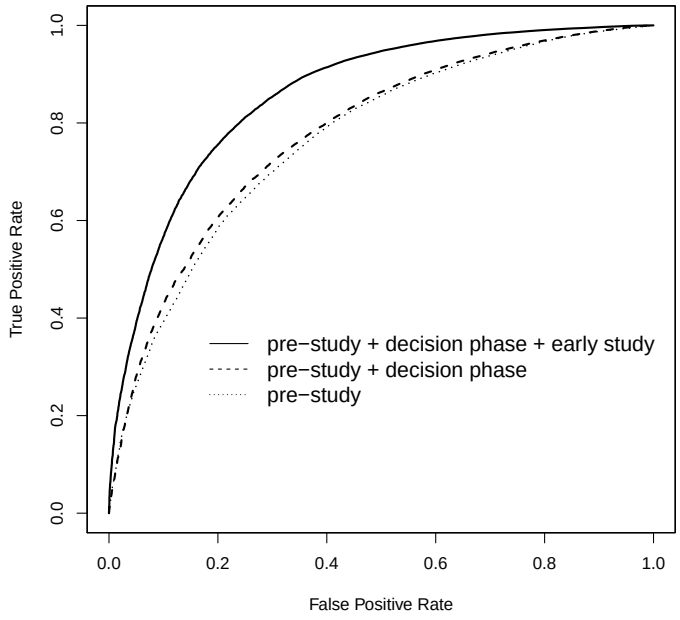

Figure 4.3: ROC-curves of the three general models

of the dropout rate in the data and because it is very likely that the status Y and the random variable C are not independent, since dropouts are more prone to leave the panel than graduates (28% probability vs. 5%, see Figure 4.2). To deal with this issue, we make two assumptions and determine their appropriateness, which may indicate that the results are not seriously biased by panel attrition.

Assumption A1: difference in the distribution of Y, based on covariates, is in both subsample populations ($C = 1$ and $C = 0$) and the overall model ($C = 1 \vee C = 0$) approximately equal.

We investigate the change in the conditional probability

$$P(Y = y|X_{t\leq 3} = x^*) - P(Y = y|X_{t\leq 3} = x^{**}),$$

where x^* and x^{**} are two possible values of a covariate X. According to assumption A1, we have
$P(Y = y|C = 0, X_{t\leq 3} = x^*) - P(Y|C = 0, X_{t\leq 3} = x^{**}) =$
$P(Y = y|C = 1, X_{t\leq 3} = x^*) - P(Y|C = 1, X_{t\leq 3} = x^{**}) =$
$P(Y = y|X_{t\leq 3} = x^*) - P(Y|X_{t\leq 3} = x^{**}).$

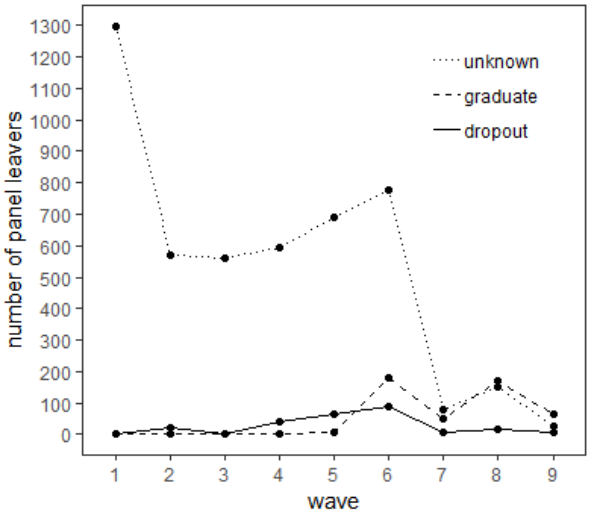

Figure 4.4: Wave-related changes in panel leaver rates according to their final status before leaving the panel

We examine the first assumption using those panel leavers for which the status of dropout or graduate is known. As shown in Table 4.12 in the appendix, equal differences are revealed concerning the distribution of dropout students (1.4%) as well as of graduate students (3%) among leavers and retained persons based on gender. Based on the study fields, more precisely for mathematics/natural sciences and linguistics/cultural sciences, equal differences for the distribution of dropouts is noted (1.2%); for graduates, differences vary slightly. Based on further covariates, approximately equal differences for dropout students are observed.

Assumption A2: panel leavers with available status (720) and panel leavers with unknown status (4,749) reveal similar distributions over the investigated covariates.

To assess this assumption, two statistical hypothesis tests are computed: a two-sample "Student's t-Test" (Rice, 2006) to determine if the means of some determinants in each group are significantly different from each other, and a "Pearson's chi-squared test" (Plackett, 1983) to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories, i.e. whether the occurrence of the outcomes of a given determinant (e.g. gender) and the observed status of a panel leaver (known or unknown) is significantly dependent.

Table 4.13 (in the appendix) provides the p-values of the computed tests. As can be observed, almost all the p-values are greater than 0.05. We find, therefore, neither signif-

icant differences in the means of determinants between both groups (panel leavers with known status and panel leavers with unknown status), nor existing relationships between determinants and the status information of the panel leavers.

Furthermore, we evaluate the panel-leaving bias in our predictions by computing the area under the ROC curve for both models, a model using the subsample of panel leavers with $Y = 1 \vee Y = 2$ and $C = 1$ (720) and a model using the subsample of retained persons with $Y = 1 \vee Y = 2$ and $C = 0$ (9,290). To make the results comparable, the same computational settings reported in section 4.4.3 are used. Since the results for the panel leavers within the study fields, especially engineering and mathematics/natural sciences, suffer under a small number of observations, only the models computed using the overall data are comparable (the same bias is observed when we select a random subsample of the same size from the original model, so we believe the bias at this stage is mainly caused by small subsample size). Absolute differences are calculated between the complete model and that using panel respondents in the different study fields, as shown in the last three rows of Table 4.10. The results indicate strong evidence that there is no major difference between the complete model and the model using only panel respondents. Performance differences between models using panel respondents and those using panel leavers over all study fields are also marginally low. That means, under our assumptions, that there are no noteworthy differences in model performance due to panel leaving.

Moreover, variable importances are calculated for the models using only panel respondents and panel leavers. The results confirm our previous findings. The variable ranking in the model using panel respondents is very similar to that of the complete model. Using only panel leavers, the variable ranking differs slightly, caused by the small subsample sizes within study fields, except in the model over all study fields. These results indicate that the bias caused by panel leaving is very small or even non-existent in our analysis.

4.6 Discussion and conclusion

This study aims at predicting dropout from higher education institutions as soon as possible, using random forests based on conditional inference trees and a broad German data set including freshman students and covering a wide range of aspects of student life

Table 4.10: AUC values of the complete model, the model only with panel respondents, the model only with panel leavers and the absolute difference between the complete model and the model using only panel respondents

	engineering	mathematics and natural sciences	law, economics and social sciences	linguistics and cultural sciences	full data
episode	complete model				
pre-study	0.77	0.79	0.75	0.74	0.77
study-rel.	0.77	0.81	0.75	0.76	0.78
early study	0.86	0.88	0.83	0.83	0.86
	only panel respondents				
pre-study	0.77	0.79	0.74	0.72	0.77
study-rel.	0.78	0.82	0.74	0.75	0.78
early study	0.87	0.87	0.82	0.81	0.86
	only panel leavers				
pre-study	0.68	0.67	0.71	0.71	0.77
study-rel.	0.68	0.70	0.70	0.70	0.78
early study	0.71	0.73	0.80	0.77	0.86
	absolute differences (complete model and model with respondents)				
pre-study	-0.00	-0.00	0.01	0.02	0.00
study-rel.	-0.01	-0.00	0.01	0.00	0.00
early study	-0.00	0.00	0.01	0.02	0.00

and study courses. Dropping out from university is a complex process, and predicting dropout remains challenging, as there are several interacting determinants inducing a student to withdraw from university without a degree. Tree-based data mining methods are well suited in the presence of high-dimensional data, high-order interactions and correlations between predictors, and are therefore very applicable for dropout prediction.

We stepwise model students' transition from school to the first semesters at university. First, we model a student's initial risk before entering university by including only pre-study determinants. In the next step, we add variables covering a students' study-decision phase. In the third step, we include predictors representing the early study phase of students at the very beginning of study. Based on these models, we identify the most relevant dropout predictors and evaluate how accurately the different models predict the risk of dropping out. We observe a strongly increased prediction performance when including variables from the early study phase in the pre-study models, resulting in an AUC between 0.83 and 0.88 for all considered study fields, as well as for the full model.

The high relevance of the secondary school grade is obvious in all three models, but decreases substantially when adding early study-phase determinants. Besides the school grade, also determinants associated with student satisfaction and their subjective academic self-concept and self-assessment play an important role in the dropout process. We conclude that although there might be a starting risk for some students, there are many ways to improve their probability of graduating successfully from higher education.

Our findings further suggest that the final grade at secondary school (grade_- school), the type of secondary school (school_type), the type of school leaving qualification (qualif_max) and the number of repeated classes (rep_class) have an impact on study progress. This is in line with previous literature observing that the pre-study educational career has an impact on academic outcomes (e.g Vandamme et al., 2007, Müller and Schneider, 2013, Stinebrickner and Stinebrickner, 2014, Siri, 2015). To take into account the different backgrounds and increased heterogeneity of students, especially since universities now try to attract more intensively so-called non-traditional students, background-specific remediation programs or bridging courses may help to prepare these students for university requirements and to harmonize the student skill levels.

The most important predictor related to the study decision phase before entering university in our model is whether students would rather have preferred to do something else rather than study at university (study_alternative). Students may benefit from considering their non-academic alternatives before entering a study program. In this respect, special offers to evaluate different opportunities and to decide which educational career would better match their aspirations and wishes may help students to find the best alternative.

Furthermore, our prediction model reveals that there are several relevant determinants that become important in the first semesters at university, but are already related to the study-selection process. Some of them are connected with student satisfaction (also found by Suhre et al., 2007), for instance, satisfied with current studies (satisf_whole) or enjoying the degree course (satisf_enjoy). Satisfaction depends substantially on the gap between student expectations concerning study content, program organization and workload and the real study situation, which is often a consequence of an information paucity (Suhre et al., 2007, Weerasinghe et al., 2017). Therefore, as a starting point, universities may offer general as well as subject-specific information for pupils already in their

qualification phase at school. Here, an expanded cooperation with secondary schools is of considerable importance (Hetze, 2011). A further approach for a first orientation includes besides general student information days, which are offered at most institutions already, probably more personalized and intensive workshops helping students to gain an overview of the different study alternatives and to find study fields matching their skills and interests. In addition, the implementation of online self-assessment programs for an initial evaluation of their own interests and suitable opportunities may also be useful (see also Heublein et al., 2014).

Other important determinants in our model include commitment to study (e.g. not do more than necessary - commit_necessary, high demands on self - commit_demands), academic self-concept (e.g. perception of talent for studying - selfconcept, opinion on the probability of graduating - probsuccess, helplessness in getting better grades - helplessness) as well as subjective satisfaction with performance (e.g. satisfaction with the academic performance - performance_eval). Similar results are found, for instance, by Vandamme et al. (2007) or Siri (2015). These results indicate that in the first semesters of study, there problems often arises concerning one's own performance, negative self-assessments, high demands on themselves and helplessness regarding how to perform better at university. Possible starting points for supporting students would be to provide seminars/work-shops on learning techniques or personalized feedback providing more realistic (rather than subjective) assessments of study progress and revealing concrete shortcomings and suggestions for performing better.

Student satisfaction with the learning environment, for instance, wanting better study conditions (satisf_conditions) or the feeling that student concerns are not taken into account sufficiently (satisf_concerns) also have an impact on dropout in our model (e.g. also found by Hovdhaugen and Aamodt, 2009, Suhre et al., 2007, Ghignoni, 2017). A starting point for improvement may entail regular student surveys to obtain information on student satisfaction and their wishes and needs, so that they are more satisfied and feel more actively encouraged and supported.

In summary, universities wishing to implement promising prevention programs should focus on strategies which help (future) students to obtain detailed information on field-specific content, requirements, study organization and workload, so that they have realistic expectations which correspond to the real study life more closely. Students should then become more satisfied and less concerned with their study choice.

However, it should be kept in mind that dropping out from university may not necessarily be interpreted as a negative event in one's educational career. A voluntary dropout may constitute a sensible revision of a wrong decision, thus allowing students to take advantage of new opportunities and become more satisfied and successful in an interesting alternative to university study, for instance in vocational training.

4.7 Appendix

Table 4.11: Participants, temporary leavers, last participation, and final panel leavers in the current SUF (LIfBi, 2017, and own calculations)

wave	instrument	partic. survey	temp. leavers	last partic.	final panel leavers
1st	CATI (+test)	17,910	0	1,299	1299
2nd	CAWI	12,273	5,591	594	594
3rd	CATI	13,113	4,560	561	561
4th	CAWI	11,202	6,424	638	638
5th	CATI (+test)	12,694	3,444	765	765
6th	CAWI	10,183	7,039	1,041	1,041
7th	CATI (+test)	9,547	7,161	774	138
8th	CAWI	8,629	6,024	1,156	338
9th	CATI	10,096	4,321	1,992	95
10th	CATI	9,090	4,192	9,090	0
sum				17,910	5,469

Table 4.12: Difference in the distribution of Y according to some covariates in the leaver and retained population. Here, categorical variables with low number of missing values are appropriate to be tested.

	retained persons ($C = 0$) / 12,441		**panel leavers** ($C = 1$) / 5,469		**total sample** 17,910	
	dropout ($Y = 2$)	graduate ($Y = 1$)	dropout	graduate	dropout	graduate
gender						
male	5.9%	67.6%	5.3%	7.0%	5.7%	48.9%
female	4.5%	70.9%	4.0%	9.8%	4.3%	52.4%
difference	1.4%	-3.3%	1.3%	-2.8%	1.4%	-3.5%
subject field (*difference for mathematics and linguistics)						
engineering	7.1%	73.6%	6.5%	7.6%	6.9%	52.2%
mathematics	5.9%	71.5%	5.3%	8.7%	5.7%	53.4%
law	4.7%	74.8%	3.5%	9.6%	4.3%	52.7%
linguistics	4.7%	65.9%	4.1%	8.7%	4.5%	49.1%
difference*	1.2%	5.6%	1.2%	0.0%	1.2%	4.3%
immigration						
no	4.9%	71.1%	4.5%	9.1%	4.8%	53.1%
yes	5.7%	64.3%	4.5%	7.5%	5.2%	43.9%
difference	-0.8%	6.8%	0.0%	1.6%	-0.4%	9.2%
family life						
with biol. par.	4.8%	70.4%	4.1%	9.1%	4.6%	52.1%
else	6.5%	64.5%	6.7%	6.1%	6.7%	44.5%
difference	-1.7%	5.9%	-2.6%	3.0%	-2.1%	7.6%
type of school attended						
up. sec. educ.	3.9%	70.4%	3.7%	9.5%	3.9%	53.1%
other types	9.2%	67.1%	6.5%	6.7%	8.2%	45.6%
difference	-5.3%	3.3%	-2.8%	2.8%	-4.3%	7.5%
completed vocational training before study						
yes	8.5%	71.2%	6.7%	8.1%	8.0%	49.8%
no	4.1%	69.2%	3.7%	8.9%	4.0%	51.4%
difference	4.4%	2.0%	3.0%	-0.8%	4.0%	-1.6%
dropout from training before university						
yes	9.5%	62.9%	7.5%	6.7%	8.7%	40.4%
no	4.9%	69.9%	4.3%	8.8%	4.7%	51.4%
difference	4.6%	-7.0%	3.2%	-2.1%	4.0%	-11.0%

Table 4.13: Tests on mean difference and test on independence between some determinants and the status of panel leavers. We test variables with low number of missing values.

	known status		**unknown status**		**p-value**	
	mean	std. err.	mean	std. err.	t-test on mean difference	chisq.test on independence
generation status	3.61	0.84	3.53	0.93	0.012***	0.111
immigration	0.24	0.43	0.27	0.44	0.095	0.114
repeated classes	0.23	0.50	0.22	0.49	0.543	0.899
birth year	1988.16	4.26	1988.42	4.11	0.116	0.402
gender	0.37	0.48	0.41	0.49	0.100	0.112
vocational training	0.28	0.45	0.24	0.43	0.036***	0.033***
dropout from training before study	0.05	0.22	0.05	0.21	0.609	0.663
at least one field change	0.05	0.23	0.05	0.23	0.986	1
at least one uni change	0.03	0.16	0.03	0.16	0.807	0.909
at least one degree change	0.01	0.12	0.02	0.15	0,063	0.157
subject field						0.197
family life	0.85	0.36	0.85	0.36	0.760	0.800
school leaving qualification	1.76	0.54	1.79	0.50	0.310	0.09
direct costs of higher education	3.42	1.03	3.36	1.02	0.124	0.169
informed about study	3.58	0.82	3.58	0.82	0.983	0.986
opportunity costs	2.98	1.05	2.99	1.00	0.728	0.067
mother qualification	4.61	2.21	4.68	2.19	0.430	0.210
father qualification	5.02	2.33	5.01	2.38	0.893	0.197
mother job	50.78	19.86	51.17	19.82	0.684	0.988
father job	53.40	21.94	53.01	22.51	0.698	0.670
grade on school leaving qualification	2.39	0.60	2.37	0.61	0.3389	0.082
type of high school	0.74	0.44	0.75	0.44	0.924	0.961
German as graduation exam	0.77	0.42	0.78	0.41	0.741	0.778
mathematics as graduation exam	0.77	0.42	0.77	0.42	0.695	0.734

*** statistically significant at 5%-level

Table 4.14: Attributes description

Attribute	Description (Data type)
	Pre-study
genstat	Generation status (numeric: from 1 = 1st generation to 4 = no immigration background) Number_Dropouts = 871, Number_Graduates = 9139
immigration	Do you have an immigration background? (binary: 0 = No, 1 = Yes) Number_Dropouts = 871, Number_Graduates = 9139
rep_class	How many class years have you ever repeated? (numeric: from 0 to 4) Number_Dropouts = 871, Number_Graduates = 9138
ger_prep	To what extent had you acquired German knowledge and skills before starting university? (numeric: from 1 = not at all to 4 = very much) Number_Dropouts = 487, Number_Graduates = 6564
math_prep	To what extent had you acquired maths knowledge and skills before starting university? (numeric: from 1 = not at all to 4 = very much) Number_Dropouts = 450, Number_Graduates = 5924
familylife	With whom did you spend most of your childhood up to the age of 14? (binary: 1 = with biological parents, 0 = else) Number_Dropouts = 871, Number_Graduates = 9136
school_type	Type of school attended (binary: 1 = upper secondary education, 0 = other types) Number_Dropouts = 838, Number_Graduates = 8947
qualif_max	School-leaving qualification obtained (numeric: 2 = general university entrance qualification, 1 = university of applied science entrance qualification, 0 = other degrees) Number_Dropouts = 870, Number_Graduates = 9136
grade_school	Approximate overall grade awarded in the school-leaving certificate (numeric: from 1 to 5) Number_Dropouts = 842, Number_Graduates = 8976
exam_german	Was German an examination subject for your school-leaving qualification? (binary: 0 = No, 1 = Yes) Number_Dropouts = 745, Number_Graduates = 8633
exam_adv_german	German as first examination subject for your school-leaving qualification (binary: 0 = No, 1 = Yes) Number_Dropouts = 752, Number_Graduates = 8669
exam_maths	Was maths an examination subject for your school-leaving qualification? (binary: 0 = No, 1 = Yes) Number_Dropouts = 743, Number_Graduates = 8646
exam_adv_maths	Maths as first examination subject for your school-leaving qualification (binary: 0 = No, 1 = Yes) Number_Dropouts = 752, Number_Graduates = 8673
gender	Gender of the person (binary: 1 = Male or 0 = Female) Number_Dropouts = 871, Number_Graduates = 9139
birthyear	Year of birth of the person (numeric: from 1950 to 1994) Number_Dropouts = 871, Number_Graduates = 9139
mother_qualif	Highest mother's general school-leaving qualification (numeric: from 0 = No school leaving qualification to 8 = Highest tertiary education) Number_Dropouts = 862, Number_Graduates = 9088
mother_job	Mother's occupation (ISEI-08) (numeric: from 11.74 to 88.96) Number_Dropouts = 638, Number_Graduates = 6733
father_qualif	Highest father's general school-leaving qualification (numeric: from 0 = No school leaving qualification to 8 = Highest tertiary education) Number_Dropouts = 833, Number_Graduates = 8953
father_job	Father occupation (ISEI-08) (numeric: from 11.74 to 88.96) Number_Dropouts = 677, Number_Graduates = 7184
voctrain	Completed vocational training before university (binary: 0 = No, 1 = Yes) Number_Dropouts = 871, Number_Graduates = 9139
fail_prestudy	Have you ever dropped out from training before university? (binary: 0 = No, 1 = Yes) Number_Dropouts = 871, Number_Graduates = 9139

Decision phase	
fieldofchoice	Enrolled in the subject of first choice (binary: 0 = No, 1 = Yes) Number_Dropouts = 621, Number_Graduates = 7072
institofchoice	Take up the degree at the institute of higher education of choice (binary: 0 = No, 1 = Yes) Number_Dropouts = 649, Number_Graduates = 7438
study_alternative	Would you rather have started something else instead of a degree? (binary: 0 = No, 1 = Yes) Number_Dropouts = 648, Number_Graduates = 7426
study_judge_parent	What do your parents think about the fact that you are studying? (numeric: from 1 = does not apply at all to 5 = applies completely) Number_Dropouts = 649, Number_Graduates = 7446
study_judge_friend	What do your friends think about the fact that you are studying? (numeric: from 1 = does not apply at all to 5 = applies completely) Number_Dropouts = 652, Number_Graduates = 7448
info_useful_...	Usefulness of information received from parents, friends, current university students, school teachers, professionals employed in the field of interest, media, university counseling, literature, school events, sneak peak at university, job agencies, companies etc. (numeric: from 0 = not used to 4 = very helpful) Number_Dropouts = 470, Number_Graduates = 6180
study_restrict	Is the study subject to admission restrictions or a selection procedure? (binary: 0 = No, 1 = Yes) Number_Dropouts = 762, Number_Graduates = 7976
Early study phase	
satisf_enjoy	Really enjoy the studied subject (numeric: from 0 = does not apply to 10 = applies completely) Number_Dropouts = 399, Number_Graduates = 7866
satisf_conditions	Wish better study conditions (numeric: from 0 = does not apply to 10 = applies completely) Number_Dropouts = 398, Number_Graduates = 7865
satisf_match	Degree course and other obligations hard to match (numeric: from 0 = does not apply to 10 = applies completely) Number_Dropouts = 399, Number_Graduates = 7865
satisf_whole	On the whole, satisfied with actual studies (numeric: from 0 = does not apply to 10 = applies completely) Number_Dropouts = 399, Number_Graduates = 7866
satisf_frustrating	External circumstances of study are frustrating (numeric: from 0 = does not apply to 10 = applies completely) Number_Dropouts = 397, Number_Graduates = 7846
satisf_kill	Degree course is killing me (numeric: from 0 = does not apply to 10 = applies completely) Number_Dropouts = 399, Number_Graduates = 7864
satisf_interesting	Degree course is really interesting (numeric: from 0 = does not apply to 10 = applies completely) Number_Dropouts = 399, Number_Graduates = 7864
satisf_concerns	Concerns of students are not taken into account sufficiently (numeric: from 0 = does not apply to 10 = applies completely) Number_Dropouts = 397, Number_Graduates = 7849
satisf_tired	Degree course often makes feel tired and exhausted (numeric: from 0 = does not apply to 10 = applies completely) Number_Dropouts = 399, Number_Graduates = 7865
partic_people	Participation in university events aimed at getting to know people (binary: 0 = No, 1 = Yes) Number_Dropouts = 753, Number_Graduates = 7937
partic_orga	Participation in university events on study organization (binary: 0 = No, 1 = Yes) Number_Dropouts = 746, Number_Graduates = 7865

partic_facil	Participation in university events on the use of central facilities (binary: 0 = No, 1 = Yes) Number_Dropouts = 740, Number_Graduates = 7789
partic_course	Participation in university events on bridging courses (binary: 0 = No, 1 = Yes) Number_Dropouts = 748, Number_Graduates = 7867
partic_acadskills	Participation in university events on academic skills (binary: 0 = No, 1 = Yes) Number_Dropouts = 744, Number_Graduates = 7769
preparation	How can you rate your preparation at the start of the university in work techniques, fundamental academic methods etc.? (numeric: from 0 = bad to 4 = good) Number_Dropouts = 548, Number_Graduates = 7132
skills_prep	Necessary knowledge acquired in maths, German, English and computer science before university (numeric: from 1 = not at all to 4 = very much) Number_Dropouts = 544, Number_Graduates = 7096
workload_match	Study progress (number of courses, credits earned) match to the curriculum plan (numeric: from 1 = much less to 5 = many more) Number_Dropouts = 388, Number_Graduates = 6944
performance_eval	Satisfaction with the academic performances till yet (numeric: from 1 = does not apply at all to 4 = applies completely) Number_Dropouts = 426, Number_Graduates = 6999
probsuccess	Your opinion on the probability that you will graduate (numeric: from 1 = very unlikely to 5 = very likely) Number_Dropouts = 425, Number_Graduates = 6978
selfconcept	Perception of your talent for studying (numeric: from 1 = low to 7 = high) Number_Dropouts = 423, Number_Graduates = 6921
study_informed	How well you are informed about the possibilities, limitations etc for your degree course? (numeric: from 1 = very poor to 1 = very good) Number_Dropouts = 856, Number_Graduates = 9124
socint_instructors	Acceptance by instructors and getting along well with them (numeric: from 1 = does not apply to 4 = applies completely) Number_Dropouts = 426, Number_Graduates = 6998
socint_students	Successful in establishing contacts and getting along well with classmates (numeric: from 1 = does not apply to 4 = applies completely) Number_Dropouts = 425, Number_Graduates = 6986
commit_necessary	Commitment to degree course: Do no more than necessary (numeric: from 1 = does not apply to 5 = applies completely) Number_Dropouts = 424, Number_Graduates = 6977
commit_enjoy	Commitment to degree course: enjoyment of degree program (numeric: from 1 = does not apply to 5 = applies completely) Number_Dropouts = 424, Number_Graduates = 6962
commit_demands	Commitment to degree course: High demands on self (numeric: from 1 = does not apply to 5 = applies completely) Number_Dropouts = 423, Number_Graduates = 6954
commit_identificat	Commitment to degree course: Identification with degree program (numeric: from 1 = does not apply to 5 = applies completely) Number_Dropouts = 421, Number_Graduates = 6938
helplessness	You think you will never get better grades (numeric: from 1 = does not apply to 5 = applies completely) Number_Dropouts = 420, Number_Graduates = 6921
job_semester	Number of hours spent in a week during semester time for employment (numeric: from 0 to 60) Number_Dropouts = 434, Number_Graduates = 7064

study_semester	Number of hours spent in a week during semester time for study-oriented activities (numeric: from 0 to 60) Number_Dropouts = 434, Number_Graduates = 7069
job_break	Number of hours spent in a week during semester break for employment (numeric: from 0 to 60) Number_Dropouts = 434, Number_Graduates = 7061
study_break	Number of hours spent in a week during semester break for study-oriented activities (numeric: from 0 to 60) Number_Dropouts = 434, Number_Graduates = 7059
costs_direct	How difficult is it to pay for direct costs of higher education? (numeric: from 1 = very difficult to 5= very easy) Number_Dropouts = 858, Number_Graduates = 9119
costs_opportunity	Limitation of the possibilities to earn own money and supporting yourself up until graduation (numeric: from 1 = not at all to 1 = a lot) Number_Dropouts = 857, Number_Graduates = 9110
financialaid _bafoeg	Currently receive student financial aid (BAföG)? (binary: 0 = No, 1 = Yes) Number_Dropouts = 201, Number_Graduates = 3093
funding	Amount of money at your disposal on average each month in Euros (numeric: from 0 to 10900) Number_Dropouts = 387, Number_Graduates = 6899
change_field	Have you ever changed the study field at least once in the past? (binary: 0 = No, 1 = Yes) Number_Dropouts = 871, Number_Graduates = 9139
change_uni	Have you ever changed the university type at least once in the past? (binary: 0 = No, 1 = Yes) Number_Dropouts = 871, Number_Graduates = 9139
change_degree	Have you ever changed the type of your degree at least once in the past? (binary: 0 = No, 1 = Yes) Number_Dropouts = 871, Number_Graduates = 9139

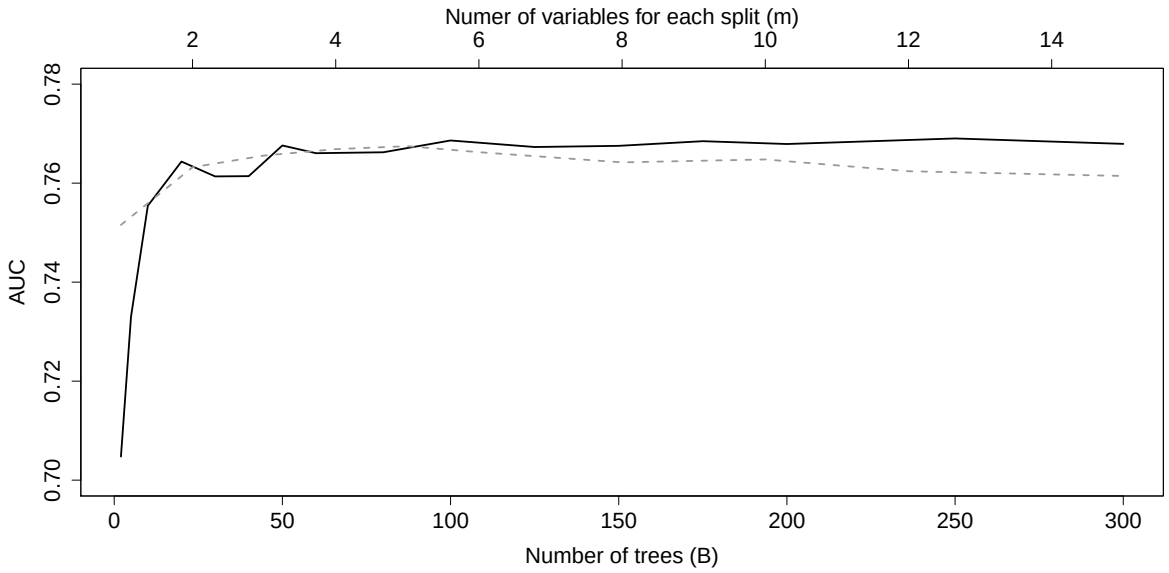

Figure 4.5: Classification performance (AUC) for the pre-study episode using all subject groups dependent on different numbers of trees and fixed $m = [\sqrt{p}]$ (black, solid line, lower scale; $p = 20$) and for different numbers of m and a constant number of trees (grey, dashed line, upper scale; $B = 100$).

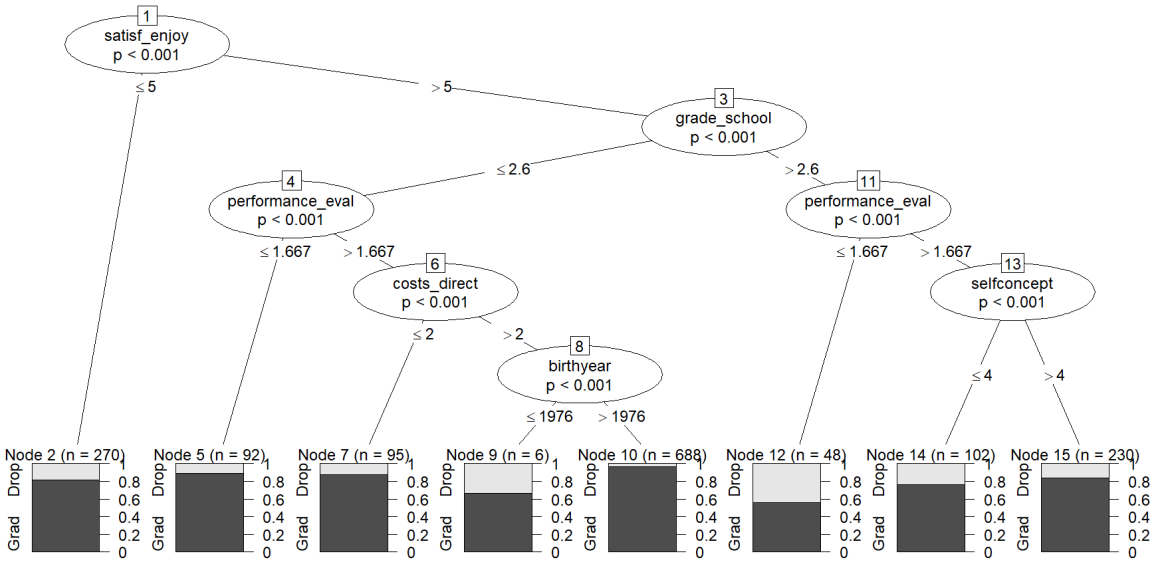

Figure 4.6: Example for one single tree for the engineering students in the early study phase

5 Predicting dropout from higher education - a comparison of machine learning algorithms

Predicting dropout from higher education - a comparison of machine learning algorithms

Andreas Behr, Marco Giese, Herve D. Teguim K., Katja Theune

Chair of Statistics
University of Duisburg-Essen, 45117 Essen, Germany

Abstract

Identifying students at risk of dropping out is a very relevant issue for higher education institutions. Based on a comprehensive German survey data set, we aim at developing an optimal classifier to predict potential dropout students representing a compromise between a good predictive performance, a straightforward interpretation of the results and an easy implementation. We tune and compare different machine learning algorithms. The best predictions are obtained with random forest and AdaBoost, which clearly outperform the benchmark classifiers naive Bayes and logistic regression. Combining different classifiers, called stacking, leads to a further improvement of the predictive models but also becomes more complex. The results are helpful for analysts of higher education institutions interested in implementing efficient early warning systems to identify and support students at risk.

Keywords: student dropout, higher education, dropout prediction, machine learning algorithms, intervention measures

5.1 Introduction

Dropping out of the tertiary education system is a field of increasing interest caused by the rising number of students in higher education institutions and the personal and social costs associated with the dropout phenomenon. For instance, from winter term 15/16 to winter term 17/18 the number of students enrolled in the German higher education system increased from 2.76 million to 2.84 million (DESTATIS, 2018). In the cohort of 25 to 64-year-old persons ever enrolled at a higher education institution in Germany, 14.7% left tertiary education without obtaining any degree. For comparison, in Italy the dropout rate is about 34% (Schnepf, 2014).[1] Therefore, universities are increasingly searching for promising early warning systems and programs to identify and help students at risk.

Dropping out from higher education is a long and complex process. Different determinants accumulate and affect each other so that the final decision to drop out is made as a result of negative circumstances. Advanced machine learning algorithms are particularly suitable to address this complexity and to determine the probability for each student to leave the university without a degree. In this context, it is important to obtain very accurate predictions, and thereby to avoid good students to be misclassified as at-risk or imperiled students to be misclassified as successful. The first-mentioned type of miss-classification may demotivate students who are not really at risk and increase costs for needless programs and waste of resources. The latter miss-classification results in non-identified students at risk which will not be supported by special programs to prevent dropout later in the study. However, higher education institutions mostly do not only aim at obtaining a good prediction model, but also at detecting promising starting points for intervention as early as possible after enrollment. Therefore, besides model performance, we focus on building a prediction model, which will provide such starting points by including determinants from the very early (pre-) study course.

From the pre-study phase, we include demographic variables, parental background and information about secondary education. Furthermore, we consider determinants from the study-decision phase, i.e. the subject or institution of choice or different information sources used for the study decision. Finally, from the early study phase variables like

[1] The proportion of dropout students has to be viewed with caution since it depends on the dropout definition and the data and strongly varies between different studies.

study satisfaction, helplessness or study commitment are used.

For the analysis, we apply different classifiers. After comparison, the classifier with the best performance in terms of the area under the ROC-curve (Receiver Operation Characteristic), named AUC, and the root mean squared error (RMSE) is selected.

Naive Bayes, a so-called "weak learner" which produces results significantly better than a random guess and which is fast in computation, is used as the benchmark model. Further, we apply logistic regression which is a widely used algorithm for binary classification problems. We compare the results of these basic learners to the outcomes of modern machine learning techniques such as support vector machines (SVM), random forest (RF) and AdaBoost. We refer to RF and AdaBoost as tree-based classifiers, since we used classification and regression trees (CART) as a base learner for both ensemble methods. However, bagging (the RF is a bagging algorithm) as well as boosting techniques can be applied to every base learner. For each model, we evaluate the optimal hyperparameter settings. The best performances are obtained by both tree-based classifiers, random forest and AdaBoost (14% better than the benchmark classifier naive Bayes). To further improve prediction performance, we apply feature selection and the stacking method, which is a combination of different classifiers using the strength of each algorithm. We discuss the advantages and disadvantages of the different models and the applicability of the techniques for institutions wishing to implement efficient early warning systems to identify students at risk, to support them with more specific intervention measures and thereby improving the institutions' educational effectiveness.

The study is structured as follows. An overview of previous literature in the field of educational data mining and students' dropout prediction in the higher education system is given in section 2. The data set and used variables are described in section 3. The methodological approach is explained in section 4. Section 5 presents the results of the model comparison with a discussion of the results. Section 6 concludes.

5.2 Literature review

5.2.1 University dropout predictors

Empirical studies, mainly using standard econometric models, identified several possible reasons for dropping out of higher education. On the institutional level, program organization (Heublein et al., 2017), teaching quality and learning environment (Georg, 2009, Hovdhaugen and Aamodt, 2009), as well as the relationship between students and teachers (Ghignoni, 2017) seem to be important predictors for university dropout. On the individual level, demographic determinants such as gender (Gury, 2011), age (Müller and Schneider, 2013), family (educational) background (Aina, 2013, Ghignoni, 2017), as well as migration background (Belloc et al., 2010) and students' pre-study education (Müller and Schneider, 2013) such as the grade point average at secondary school (Stinebrickner and Stinebrickner, 2014) are relevant for the decision to leave university without degree. Moreover, personal characteristics and behavior, including conscientiousness, resilience and self-control (Van Bragt et al., 2011a,b), student motivation, organization and learning strategy (Schiefele et al., 2007), degree program satisfaction (Suhre et al., 2007), person-environment fit (Suhlmann et al., 2018), class attendance (Korhonen and Rautopuro, 2018, Nordmann et al., 2019), as well as off-study work (Beerkens et al., 2011, Hovdhaugen, 2015) tend to have an influence on study continuation. For more detailed reviews dealing with the international state of dropout research see Larsen et al. (2013c) or Vossensteyn et al. (2015).

5.2.2 Machine learning techniques

Luan (2002) was one of the first researchers describing the advantages of the application of modern machine learning techniques on educational data. Subsequently, Baker and Yacef (2009) stated that the field of educational data mining is growing rapidly and will become more important due to increased data availability. And recently, Judith Singer states the need for more advanced techniques beyond traditional methods in educational research, such as machine learning algorithms (Singer, 2019).

Commonly used methods for classification and prediction in the educational domain are, for instance, naive Bayes (NB) (e.g. Rovira et al., 2017), decision trees (DT) (e.g. Kemper et al., 2019), linear (or quadratic) discriminant analysis (LDA/QDA) (e.g. Vandamme

et al., 2007), random forests (RF) (e.g. Hoffait and Schyns, 2017), K-nearest neighbors (KNN) (e.g. Aulck et al., 2016), logistic regression (LR) (e.g. Jadrić et al., 2010), artificial neural networks (NN) (e.g. Hoffait and Schyns, 2017) and support vector machines (SVM) (e.g. Rodriguez-Muñiz et al., 2019). For a review of data mining in education see, for instance, Romero and Ventura (2010, 2013).

5.2.3 Comparing different approaches in dropout prediction

In the following, we focus on studies that compare several classification algorithms for dropout prediction in higher education institutions concerning their performance measured in terms of accuracy, as this value is provided in most of these studies.

Jadrić et al. (2010) use the database of the Faculty of Economics Information System in Split (Croatia) and apply logistic regression, decision trees and neural networks to predict dropout for 715 students. Neural Networks are evaluated as the best models compared to all other ones. Among the included variables such as demographics, educational background, parental background, and attributes referring to the studying process, the number of exam takings or failing, especially first semester marks in Mathematics seem to be important predictors. Similarly, Aulck et al. (2016) analyze about 32,500 students using demographics, pre-college entry information, and transcript records including classes taken and grades received at the University of Washington. Comparing logistic regression, random forests and k-nearest neighbors, they observe accuracies only of around 65% for each model, with the highest performance of the logistic regression (66.59%). The strongest predictors are GPA in math, English, chemistry, and psychology courses.

Hoffait and Schyns (2017) limit their analysis on data already available at enrollment, including demographics, educational and socio-economic background, to early predict dropout of 6,845 students at a Belgian university. They apply three data mining methods, random forest, logistic regression and artificial neural network, all obtaining a correct classification rate around 70% in the failure class. Adding an "uncertain" class to the failure and the success class increases the prediction accuracy of dropouts to about 90%. Important predictors seem to be the mathematics level at school and the age at registration.

In contrast, Rovira et al. (2017) use only grades of each course to predict early dropout of 4,434 students from degree studies in Law, Computer Science and Mathematics at the University of Barcelona to support tutors in providing guidance and advice to their students in specific subjects. They compare several state-of-the-art classifiers and obtain the best results for AdaBoost (AB) and random forests that outperform the base classifiers naive Bayes and logistic regression by far (accuracy of all models above 90%). Also mainly using data on student study progress and performance, Kemper et al. (2019) apply logistic regression and decision trees to predict dropout of over 3,000 students from an Industrial Engineering degree at the Karlsruhe Institute of Technology (KIT, Germany). They achieve a prediction accuracy of up to 95% based on a decision tree with data for three semesters. The most relevant determinants are the count/average of passed/failed examinations, the average grade (for models of later semesters) and specific single exams.

Some studies use a broader data set including also so-called "soft" factors such as student motivation and perceptions. An early work by Vandamme et al. (2007) aims to classify 533 Belgian first-year university students into three risk groups for dropping out: low-risk, medium-risk, and high-risk group. To offer help for students as soon as possible, classification is done before the first university examinations. They apply three different methods, decision trees, neural networks and linear discriminant analysis, and find no prediction model to perform well, with the best result obtained with discriminant analysis (accuracy of 57.35%). Including predictors from students' personal history, their involvement in studies and perceptions, most important for study success seem to be attendance at courses, the perceived subjective chance of study success, previous academic experience (mainly mathematics), and study skills. They conclude that even though some pre-study factors seem to influence academic careers, there are opportunities to counteract a starting risk towards positive development and academic success. Rodriguez-Muñiz et al. (2019) combine information from the university data store, including demographic characteristics, educational background or study-related variables, and information from an additional questionnaire, including motivation, satisfaction or relationship with teachers and peers. They analyze 1,055 students from the University of Oviedo (Spain) and compare the C4.5 algorithm, random forest, classification and regression trees (CART), Bayes nets, and support vector machines. Accuracies are found to be very similar, with the best for random forest (86.6%). Important predictors of academic success are personal and contextual variables, as well as academic performance in the first year.

5.2.4 Summary and contribution

In summary, there is mixed evidence concerning both, the best algorithm for dropout prediction and the most important variables. The results of the studies are not easily comparable as they pursue different research questions (only good prediction vs. identifying relevant determinants) and use different data sets and predictors.

The abovementioned studies are often based on small data sets and/or restrict their analysis to specific academic courses and/or to one university. They mostly do not consider all the determinants identified from theoretical and empirical research to be relevant for student dropout. Usually, higher education institutions not only aim at predicting precisely students as dropouts but also want to know the underlying circumstances and possible intervention strategies. Administrative data, which is used in many studies, often include only "hard" university non-malleable factors, like the social background or gender, or attributes related to study progress, but research would benefit from dealing more with "softer" study-related and university malleable factors, as these are mainly within the scope of policy/institution action (Larsen et al., 2013c). Especially large (survey) data sets covering a wide range of variables are very well suited for data mining techniques and provide starting points for intervention measures. The model performance also strongly depends on the time when the information about the students is gathered. Information raised later in the study course guarantees good performance (Kemper et al., 2019) but usually one is interested in dropout prediction at a very early stage of study to counteract the dropout process (Vandamme et al., 2007).

This study uses a data set which combines administrative and subjective data on a wide range of students' life and aims at early predicting dropout from higher education for a whole student cohort. Moreover, previous studies on evaluating and comparing machine learning algorithms to predict student dropout mainly focus on their predictive power. This study takes a more comprehensive look at the advantages and disadvantages of the different machine learning techniques and discusses their applicability related to the specific problems in the educational context and for practitioners wishing to implement early warning systems to detect students at risk for dropping out. The aim is to develop a prediction model for student dropout which represents a promising compromise between good predictive performance, providing starting points

for intervention, a straightforward interpretation of the results and an easy implementation.

5.3 Data and variables

Our data basis is the fifth cohort of the National Educational Panel Study (NEPS)[2]. This is a comprehensive German panel study with more than 3,000 variables, covering many different issues of student's life and background for a target group of 17,910 freshman students, who were enrolled the first time in a German institution of tertiary education in winter term 2010/2011 (Blossfeld et al., 2011).

The German higher education system is mainly based on two types of institutions: general universities, which are more research-oriented, and universities of applied sciences (or polytechnics, Fachhochschulen), which are more vocationally oriented. Within the Bologna processes, former degrees (Diplom, Magister, Staatsexamen) were substituted by the two-tier structure of Bachelor's and Master's degrees. The Bachelor's degree is awarded after 3 (or 4) years of study and aims to provide academic foundations and qualification for labor market entrance (HRK, 2019). The German educational system is regarded as highly socially selective as there is a strong dependency between parental educational background and children's participation at upper secondary schools and tertiary education (Heublein et al., 2017, Spangenberg and Quast, 2016). Students from non-academic households are less likely to enter higher education and more likely to leave higher education without degree. Almost 50% of the student dropouts in Bachelor programs occur during the first two semesters and about 75% after four semesters. The highest dropout rates are in study fields such as engineering and mathematics/natural sciences (Heublein et al., 2017).

[2]This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort First-Year Students, doi:10.5157/NEPS:SC5:10.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

5.3.1 Student status

Student status is constructed by tracking student's study progress over the ten waves provided by the NEPS and thereby identifying after each wave whether the student is still studying (the initially chosen subject), has successfully graduated or has completely dropped out the studies without degree (in the time span observed).

Relating to Larsen et al. (2013c), the term "dropout" is defined as completely leaving the higher education system without any degree. Field, degree or university changes represent, according to this definition, transfers to another program or university and are therefore not treated as dropouts. This definition derives from a macro point of view, from which dropout is seen as definitely leaving the whole higher education system. Even though we can not exactly assure that a student defined as dropout would never enter university again later in life, we assume this to be very unlikely, as we observed students over ten waves and almost six years after study start and re-entering study rarely occurred. For comparison, we provide results for a further definition of dropout including students changing their study field before graduation in the online appendix . In our main analysis, transfers during the study are considered as predictors. Students who obtained their first higher education degree during the observed period are regarded as "graduates".

Due to data limitations, we only observe four major subject groups (1) Engineering, (2) Mathematics and Natural Sciences, (3) Law, Economics and Social Sciences, and (4) Linguistics and Cultural Sciences. In total, the analysis is based on 8,964 students (810 dropouts, 8,154 graduates) divided into the four subject groups (see Table 5.3).

5.3.2 Predictor variables

As dropping out is a long decision process, we use a wide range of 81 variables from several steps in the course of study based on findings from the previous theoretical and empirical literature. These variables cover 20 determinants from the pre-university phase (e.g. gender, age, migration background, family background, final grade at secondary school, etc.), 21 variables from the study-decision phase (e.g. studying the subject of choice, opinion of parents and friends regarding the study course, information about study, etc.) and 40 determinants from the early study phase (e.g. study satisfaction,

academic and social integration, off-study work, financial situation, commitment to degree course, etc.); see Table E in the online appendix. We use both, time-invariant variables which are fixed at the beginning of the study, and time-variant variables which may change throughout student's progress. Variables of the pre-university and study-decision phases are fixed and have been surveyed at the first and second waves. Variables of the early study phase may vary over the educational career. However, for comparability and since we aim at predicting university dropout as early as possible, we select the early available information on these predictors, i.e. up to the third wave. The first survey wave started in October 2010 at the beginning of the first semester and ended in July 2011. The second wave started at the beginning of the third semester and ended in January 2012. The third wave was surveyed during the lecture period of the fourth semester. In the online appendix, we provide results for an extended model including also university grades and credits as predictor variables. We do not include these determinants in our main classification models for reasons explained in detail in the appendix.

5.4 Methodological approaches

In this section, we describe our methodological approach for hyperparameter optimization and model evaluation as well as the different applied classification algorithms: logistic regression, support vector machine, random forest and AdaBoost. As a baseline comparison of our methods, naive Bayes will be tested.

To build and evaluate the different machine learning models, one is interested in the test error rate, which is the error when predicting the outcome of new unseen observations with the trained model (out-of-sample prediction). To do that, the original sample data is often partitioned into training sets and test sets. The training set is used to fit the model and the test set is used to evaluate the model's goodness of fit on data it has never seen before. A suitable approach is to apply the cross-validation method (Hastie et al., 2009). In the k-fold cross-validation (k-fold CV), the original sample data is randomly partitioned into k folds of approximately equal size. From the k subsamples, $k-1$ subsamples are used as training data to fit the model, and the remaining subsample is used as test data for evaluating the model. The cross-validation process is repeated k times, with each of the k subsamples being used as test data. The average of the resulting error rates is an appropriate estimate of the test error rate. However, we should note that,

since there could be some changes over time within new data, prediction results may not be completely unbiased when the models are applied to more recent data. Sometimes, a third data set, called validation data, can also be generated from original data, which serves to optimize the hyperparameters of each algorithm.

5.4.1 Hyperparameter optimization

There are two kinds of parameters during the training of classification algorithms. Hyperparameters, which must be set by the user before the learning process begins, and other parameters, also called weights, like the β-vector in a regression model, whose values are estimated by the model itself via training. The values we choose for the hyperparameters have an influence on the model performance and always depend on the underlying dataset. Grid search (Ozdemir, 2016) is a traditional technique used to tune or optimize the hyperparameters for good model performance. For each combination (within a manually specified set of values) of the hyperparameters derived from a b-dimensional grid, where b is the number of the different hyperparameters to be set in the model, the grid search trains the classification method and evaluates the model performance. The best hyperparameters are the ones that yield the highest performance. Grid search is generally performed using the cross-validation technique.

During the training process with the $k-1$ parts of the data, the internal grid search is applied using another l-fold cross-validation (training data is split into training data and validation data). This procedure of using an inner cross-validation loop to tune the hyperparameters and embedding this inner cross-validation in each loop of an outer cross-validation, which serves for model evaluation, is called nested cross-validation. Nested cross-validation leads to almost unbiased estimates since the test data has not been used for hyperparameter tuning. Otherwise, the results would be too optimistic (Krstajic et al., 2014). For our analysis, we apply an inner 10-fold cross-validation embedded in an outer 10-fold cross-validation. Furthermore, as suggested by Krstajic et al. (2014), we repeat the nested cross-validation 20 times (using different train/test partitions of the data) to get a distribution of the estimates and to reduce the variance of our estimations. Values of the tuned hyperparameters after a run of the grid search for the different applied algorithms are shown in Table D in the appendix.

5.4.2 Evaluation measures

To assess the predictive performance of the different algorithms[3], we calculate the AUC and the RMSE, and the aggregation of both measures is generally used for the evaluation of student models (Pelánek, 2015).

Dropouts are denoted as positives, and graduates as negatives. Here, the true positive rate is the fraction of correctly classified dropouts among all real dropouts, whereas the false positive rate is the proportion of all graduates misclassified as dropouts among all real graduates. These two measures as well as the accuracy, which is the total proportion of correctly classified observations, are influenced by the class sizes and strongly depend on the selected threshold (Dinov, 2018). The threshold in a binary classification is the dropout-probability which for a new observation must be exceeded to be classified as class 1, i.e. as dropout. AUC provides a robust metric for binary classification evaluation by plotting the true positive rate against the false positive rate for various thresholds. It is not influenced by different class sizes in contrast to the above mentioned evaluation measures. Hence, it is well suited for scenarios with unbalanced classes. The AUC lies in the interval [0, 1], where 0 indicates that always the wrong class was predicted, 0.5 indicates a random guess and 1 means a perfect classification for at least one threshold value.

RMSE measures the squared error between $\hat{p}_{1_i}$, that is the predicted probability for class 1, and the observed value $y_i \in \{0, 1\}$, $i = 1, \ldots, n$:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} (\hat{p}_{1_i} - y_i)^2 / n}. \tag{5.1}$$

The closer the predicted probabilities are to the observed values, the smaller the value of RMSE. A good predictive model should have a high AUC value and a small RMSE value. Therefore, for each model the combined score AUC + (1−RMSE) is computed using a 10-fold cross-validation approach which is also repeated 20 times[4].

[3]For all performance predictions of the five explained machine learning algorithms, we use the statistic software R (R Core Team, 2019).

[4]This score is often used in Data Mining Competitions, see e.g. https://sites.google.com/view/assistmentsdatamining/data-mining-competition-2017

5.4.3 Naive Bayes

Naive Bayes (NB) is one of the Bayesian learning methods, which are based on the well-known Bayes theorem (Tan et al., 2007).

Let $\mathbf{X} = (X_1, \cdots, X_d)$ be a (students') d-dimensional random set of attributes (here e.g. gender, school grades etc.) and Y a random class variable, here $Y \in \{0, 1\}$ ($0 =$ graduate and $1 =$ dropout). The Bayes theorem has the form:

$$P(Y|\mathbf{X}) = \frac{P(\mathbf{X}|Y) \times P(Y)}{P(\mathbf{X})}. \tag{5.2}$$

The Bayesian learning methods consist of modeling the probabilistic relationship between the attribute set $\mathbf{X}$ and the class variable Y by using the a posteriori probability $P(Y|\mathbf{X})$. The most probable value taken by the class variable Y for every combination of $\mathbf{X}$ observed in the data, i.e. the value which maximizes the a posteriori probability, is estimated. Using the maximum a posteriori principle, the most probable output value y_0 for given values of the attribute set $\mathbf{X}$ is obtained as follows:

$$\begin{aligned} y_0 &= \underset{y \in \{0,1\}}{\operatorname{argmax}} \; P(Y = y|\mathbf{X}) \\ &= \underset{y \in \{0,1\}}{\operatorname{argmax}} \; \frac{P(\mathbf{X}|Y = y) \times P(Y = y)}{P(\mathbf{X})} \\ &= \underset{y \in \{0,1\}}{\operatorname{argmax}} \; P(X_1 = x_1, \cdots, X_d = x_d|Y = y) \times P(Y = y), \end{aligned} \tag{5.3}$$

The a priori probability $P(Y = y)$ can easily be estimated by counting the frequency of each class occurring in the n observations. To estimate the joint conditional probability $P(X_1 = x_1, \cdots, X_d = x_d|Y = y)$, Naive Bayes uses the conditional independence assumption, that suggests that the values of the attribute set are conditionally independent given the class variable:

$$P(\mathbf{X}|Y_i) = P(X_1 = x_1, \cdots, X_d = x_d|Y = y) = \prod_{j=1}^{d} P(X_j = x_j|Y = y). \tag{5.4}$$

Substituting Equation (5.4) into Equation (5.3) leads to

$$y_0 = \underset{y \in \{0,1\}}{\operatorname{argmax}} \; P(Y = y) \times \prod_{j=1}^{d} P(X_j|Y = y). \tag{5.5}$$

Therefore, instead of calculating the class-conditional probability for every combination of $\mathbf{X}$, only the conditional probability of each X_j, given Y has to be estimated.

In case of a continuous attribute, discrete intervals are built. The frequencies of the resulting ordinal attributes can be computed.

The NB classifier has one relevant hyperparameter, the Laplace smoothing factor, which specifies the joint conditional probability of a predictor. When it is set to 0, NB predicts a zero probability for any test data point that contains a previously unseen categorical level.[5]

5.4.4 Logistic regression

In the logistic regression model (LR or logit), the dependent class variable Y is coded as 0 or 1, here graduate (0) and dropout (1). The aim is to compute the a posteriori probabilities of both classes depending on the values of the attributes $\mathbf{X} = (X_1, \cdots, X_d)$. These probabilities are defined as:

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = F(\beta_0 + \mathbf{x}^T\boldsymbol{\beta}), \tag{5.6}$$

$$P(Y = 0|\mathbf{X} = \mathbf{x}) = 1 - F(\beta_0 + \mathbf{x}^T\boldsymbol{\beta}) \tag{5.7}$$

and have to be restricted to the interval [0,1]. A natural choice for the index function F is the logistic distribution function with the form

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = F(\beta_0 + \mathbf{x}^T\boldsymbol{\beta}) = \frac{\exp(\beta_0 + \mathbf{x}^T\boldsymbol{\beta})}{1 + \exp(\beta_0 + \mathbf{x}^T\boldsymbol{\beta})}. \tag{5.8}$$

The parameters $(\beta_0, \boldsymbol{\beta})$ are estimated using maximum likelihood method.

In addition to the main effects, we can include interaction effects and effects in quadratic order (see section 5.2.1). This results in:

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp(\beta_0 + \mathbf{x}^T\boldsymbol{\beta} + \sum_{i=1,j=1,i\leq j}^{d} \gamma_{i,j}\mathrm{x}_i\mathrm{x}_j)}{1 + \exp(\beta_0 + \mathbf{x}^T\boldsymbol{\beta} + \sum_{i=1,j=1,i\leq j}^{d} \gamma_{i,j}\mathrm{x}_i\mathrm{x}_j)}. \tag{5.9}$$

[5]The function *naive_bayes* from the R-package *naivebayes* is used for computation (Majka, 2017).

As the number of predictor variables can rapidly increase, Hastie and Qian (2014) proposes the negative binomial likelihood and a regularization parameter λ to penalize unimportant or highly correlated features and shrink their coefficients to zero. This leads to the following minimization problem (for simplification only main effects are considered):

$$\min_{\beta_0,\beta\in\mathbb{R}^{d+1}} -\left[\frac{1}{n}\sum_{i=1}^{n} y_i \cdot (\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}) - \log(1+\exp[\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}])\right] \\ +\lambda[(1-\alpha)\|\boldsymbol{\beta}\|_2^2/2 + \alpha\|\boldsymbol{\beta}\|_1], \tag{5.10}$$

where $\|\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{d}\beta_i^2$ is the squared Euclidean norm and $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{d}|\beta_i|$ is the Manhattan norm. The hyperparameter λ has to be tuned; it controls the overall strength of the penalty, while the hyperparameter α controls the "elastic-net" penalty. We distinguish between $\alpha = 1$ (LASSO regression) and $\alpha = 0$ (Ridge regression). Generally, LASSO regression leads to a smaller number of relevant coefficients since it picks only one coefficient (i.e. one variable) from two highly correlated variables and shrinks the other coefficient to zero. Another main application of LASSO regression is for feature selection.

The *cv.glmnet* function in R from the glmnet package (Hastie and Qian, 2014) evaluates the optimal λ via grid search. The higher the λ, the more coefficients of unimportant features are shrunk to zero and the model becomes less complex. Furthermore, we tune $\alpha \in \{0, 1\}$ via grid search.

5.4.5 Support vector machines

The support vector machine (SVM) is one of the most commonly used classification algorithms due to its promising empirical results in many applications, where data are of very high dimensions such as handwritten digit recognition, text categorization etc. (Tan et al., 2007). The actually most prevalent version of the SVM algorithm, that is used in this article, was introduced by Cortes and Vapnik (1995). For classification, SVM applies the concept of margin hyperplanes, which can be imagined as a surface maximizing the boundaries between the different types of data in order to create subspaces with homogeneous observations with regard to their class membership.[6] When constructing

[6]More precisely, hyperplanes are affine subspaces of dimension $n-1$ in a n-dimensional space.

hyperplanes, three scenarios should be distinguished:

- Data is **linearly separable**: the data can cleanly be separated by a linear hyperplane

  (see Figure 5.1, left panel). No adjustment is needed.

- Data is **linearly non separable**: in practice, complete linearly separable data points occur very rarely. Some data points are misclassified, and the SVM should be adjusted (see Figure 5.1, right panel).

- Data is **non-linear**: in many classification problems, data points are non-linear separable. Transforming the data to higher dimensions using the kernel trick can solve this problem.

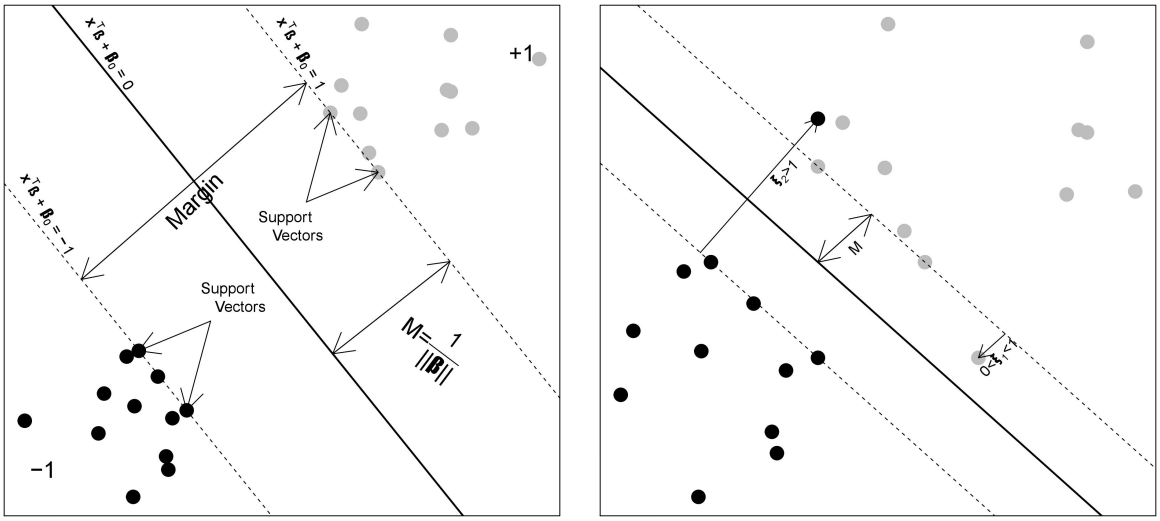

Figure 5.1: Left panel: linearly separable case. There are two support vectors in each class. The solid line is the linear hyperplane and the distance between the two dashed lines indicates the margin. Right panel: linearly non-separable scenario. One data point of each class is on the wrong side of the margin lines.

To find the optimal hyperplane, the margin, i.e. the double of the distance between the hyperplane and the nearest training data points (called support vectors), is maximized (technical details on how the margin is maximised is given in the online appendix). Suppose a set of n observations $\mathbf{x}_i, \cdots, \mathbf{x}_n$ with values of the class label $y_1, \cdots, y_n \in$

$\{-1, +1\}$. Hyperplanes can be expressed in the following form:

$$\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta} + \beta_0 = 0\}. \tag{5.11}$$

Points which lay in the subspace below the hyperplane satisfy the condition $\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0 < 0$ and belong to class -1. Points above the hyperplane belong to the other class $(+1)$ and satistiy $\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0 > 0$.

In the **linearly non-separable case**, slack variables $\xi_i, i = 1, \ldots, n$ with $\xi_i \geq 0\ \forall i$ are introduced to penalise misclassified points. They measure the distance of the misclassified points to their marginal hyperplanes. For simplification, we use the relative distance $\xi_i^* = \xi_i/\|\boldsymbol{\beta}\|$. If $\xi_i^* = 0$, the i^{th} training observation is on the correct side of its marginal hyperplane. If $0 < \xi_i^* \leq 1$, the data point lies inside the margin, but on the wrong side of its marginal hyperplane. Otherwise, if $\xi_i^* > 1$, the data lies on the wrong side of the optimal hyperplane.

For solving the **non-linear separable** problem, a kernel trick is applied. This consists of transforming the data points from the input space into a different space called feature space using a kernel function, so that the hyperplane can be fitted in the feature space. Common choices of the kernels are:

- linear $\quad k(\mathbf{x}, \mathbf{x'}) = \mathbf{x}^T\mathbf{x'}$
- polynomial $\quad k(\mathbf{x}, \mathbf{x'}) = (\gamma\mathbf{x}^T\mathbf{x'} + c)^h,\ \gamma > 0$
- radial basis function $\quad k(\mathbf{x}, \mathbf{x'}) = \exp(-\gamma\|\mathbf{x} - \mathbf{x'}\|^2),\ \gamma > 0.$

During the training of the SVM, we apply the Gaussian radial basis function (RBF) kernel and vary the hyperparameters C in the set of values $\{2^{-5}, 2^{-3},$
$\cdots, 2^{15}\}$ and γ in $\{2^{-15}, 2^{-13}, \cdots, 2^3\}$, as advised by Hsu et al. (2003).[7]

5.4.6 Random forest

Random forest (RF) is a tree-based ensemble method, which has been developed by Breiman (2001). This ensemble method is based on the well-known algorithm "Classification and Regression Trees" (CART), also defined by Breiman et al. (1984). This approach successively divides the set of data points from the top node (root) to some leaf

[7]Computations are done using the function *svm* from the R-package *e1071* (Dimitriadou et al., 2008).

nodes using recursive binary splitting. Breiman (2001) introduces RF to overcome some weaknesses of the CART algorithm, i.e. CART tends to show a high variance, suffers from the overfitting problem and is very sensitive when small changes in the data occur (small robustness). Thus, RF contributes to improving the performance by aggregating a large number of decision trees using the idea of bootstrap aggregation, known as bagging (repeated sampling with replacement) and a random feature subsampling at each node (Breiman, 2001, Breiman and Cutler, 2004).

Suppose a set of n observations $\mathbf{x}_1, \cdots, \mathbf{x}_n$ with values of the class label $y_1, \cdots, y_n$ (here dropout or graduate). The construction of RF as explained by Breiman (2001) is done as follows:

1. At each step k, a single bootstrapped sample Θ_k is generated (the Θ_k's are independent identically distributed),

2. A tree predictor $h(\mathbf{x}, \Theta_k)$ is constructed using the training observations. During the tree construction, a random set of m variables (fixed and much smaller than the total number of variables) at each node is selected. As splitting criterion, the Gini coefficient is used,

3. A large number of trees is generated, $k = 1, \cdots, K$ (usually $K \geq 100$),

4. The RF classifies a new observation $\mathbf{x}_0$ from the test data by proceeding it through each of the K trees and we take the mean of the K class probabilities (relative frequencies of training observations in the specific classes).

   Malley et al. (2012) show that RF is a consistent machine learner in probabilistic estimation, instead of using the majority vote, i.e. the majority class decision of the single trees.

Also for the RF, hyperparameters are tuned: the number of variables randomly sampled as candidates at each split is varied in the range from 1 to 15 and the minimum number of observations required to be at the leaf node in the decision tree is varied in {1,3,5,7}. The number of trees is no classical hyperparameter since the performance should improve with a rising number of trees until a specific level is reached. We chose $K = 1000$.[8]

[8]The function *randomForest* from the R-package *randomForest* has been used for computation (Liaw et al., 2002).

5.4.7 AdaBoost

AdaBoost (AB), introduced by Freund and Schapire (1997), is one of the most commonly used ensemble models. This algorithm belongs to the state-of-the-art boosting models. Boosting is an approach that can be applied in the context of many different base learners (estimators). Decision trees are generally used as base learner (that is why we denote AdaBoost here as tree-based classifier), since they are fast in computation, have a low bias, can handle metric, ordinal and nominal scaled features, as well as missing values (Bishop, 2006).

Contrary to bagging procedure (applied in RF), in which multiple samples generated from the training data are used to construct in parallel individual tree predictors and the results are combined, boosting procedure utilizes a sequential order, in which each tree is computed on a modified version of the original data and the results are added up. Then, AB reweights iteratively the training data, whereby wrongly classified observations receive higher weights than correctly classified observations. In each iteration, weights are automatically updated based on the error rate after fitting the tree classifier on the modified data. The final result is a weighted vote of all the trees predictors. A step-by-step procedure is given in detail in the appendix.

During our classification with AB, 150 boosting iterations appear to be sufficient; the results stabilize already with more than 50 iterations. Furthermore, we tune the learning rate (ν) in the set of values $\{0.01, 0.05, 0.1,\ 0.15, 0.2\}$ and the maximum depth of a tree in the set of values $\{2, 5, 7, 10, 15\}$.[9] As final prediction, we computed the probabilistic estimation as in RF. There also exist other types of this algorithm ('real" or "gentle" AdaBoost), but the results in our application are not better compared to "discrete" AdaBoost.

5.4.8 Model overview

Table 5.1 summarizes the assumptions, advantages and problems of the five classifiers applied in this study. Note that we used random forest and AdaBoost based on decision trees (CART). CART is able to handle missing values, which is often a problem in survey data, using so-called surrogate splits. If a variable is missing for a specific observation, another predictor variable is used, such that this split is similar to the best split (Twala,

[9] In R, we used the function *ada* in the R-package *ada* (Culp et al., 2006).

2009). However, this strategy increases computation time and may bias the variable importance ranking. That is why the RF cannot handle missing data in some software packages by default.

In many applications, the data contains different scaled variables. If the algorithm cannot handle ordinal or nominal features, they have to be converted probably yielding biased results or loosing information.

The publication year in Table 5.1 is based on the algorithm commonly used nowadays. For example, there have been previous versions of the SVM which make not use of the kernel trick and make the SVM only applicable for the linear separable case.

The computation time in Table 5.1 is calculated for the Engineering students and given in minutes. The SVM needs to be tuned to get satisfying results which is computationally expensive. We choose an 11×10 grid to tune the parameters C and γ which means the computer passes 110 times the 5-fold inner cross-validation loop. But even without tuning, the SVM is the slowest algorithm and is not recommended for large datasets without any variable pre-selection.

A more detailed discussion of important advantages and disadvantage of these algorithms for application in the higher education context is provided in the discussion section.

Table 5.1: Short overview of the five machine learning algorithms.

Method	Naive Bayes	Logit	SVM	Random Forest	AdaBoost
Ordinal features	no	no	no	yes	yes
Nominal features	yes	as dummy	as dummy	yes	yes
Missing values	yes	no	no	yes[10]	yes
Computation time	0.73	0.22	1036.68	41.18	18.21
Publication year	1961[11]	1944[12]	1995[13]	2001[14]	1997[15]

5.4.9 Imputation of missing values

Usually, survey data contain many missing values. In the described dataset, 16.5% of all values, except the status variable, are missing. This can be due to the fact that some students do not participate in every wave of the survey or refuse to give plausible answers.

Some of the classification methods presented in section 5.4 are able to handle missing values like tree-based methods (e.g. AB and RF) by using so-called surrogate splits (Twala, 2009). Other methods like NB, LR and SVM require complete data, i.e. without missing values. According to Aggarwal (2015) or Hastie et al. (2009), one solution is to discard all observations with a missing value in at least one variable. This procedure would reduce the dataset from 8,964 to only 780 observations, implying a severe loss of information.

Therefore, we impute the missing values and compare different methods such as mean or median imputation, regression imputation, hot deck imputation and more sophisticated multiple imputation techniques like MICE (Multivariate Imputation by Chained Equations). For more details on the imputation techniques see e.g. Twala (2009) or Batista and Monard (2003). Imputation methods require at least missing at random data (MAR). Not MAR (NMAR) would lead to biased estimates of the missing values (Baraldi and Enders, 2009, Garciarena and Santana, 2017, Twala, 2009). Here, we are interested in finding the imputation method that maximizes the classification performance in terms of AUC and RMSE.

We test the different imputation methods by applying our two benchmark classifiers (NB and LR) on the imputed full data set. Table 5.2 shows the model performance for Engineering students. Findings are very similar for the three other subject groups, except little shifts in the AUC and RMSE values in the same direction. In every subject group, models computed on the dataset generated with the single median imputation produce the best results, in terms of a high AUC and a low RMSE. Consequently, for further analyzes, we choose the complete data set obtained with this imputation method for all five models. Garciarena and Santana (2017) also found some situations where median imputation outperforms advanced imputation techniques like MICE.

Since dropout students are more prone to leave the panel than graduates (28% attrition probability for dropouts vs. 5% for graduates) panel attrition must be assumed to be NMAR. To study the robustness of our models, we split the data into two disjoint subsets containing final panel leavers with available status (810 students) and panel respondents with available status (8,154 students). The results of the models in the two subpopulations and the complete model were almost identical, suggesting that our model is not strongly biased due to panel attrition.

Table 5.2: AUC and RMSE values for NB and LR models computed on the Engineering data set generated with the different imputation methods

Imputation method	Naive Bayes		Logistic Regression	
	AUC	RMSE	AUC	RMSE
Median	**0.81**	**0.38**	**0.83**	**0.29**
MICE	0.80	0.39	0.82	0.31
Hod Deck	0.80	0.39	0.81	0.30
Random Forest	0.79	0.40	0.82	0.31
Mode	0.79	0.38	0.81	0.31
Mean	0.78	0.41	0.80	0.32

5.5 Prediction of university dropout and model comparison

5.5.1 Prediction results

The results of the different algorithms are summarized in Table 5.3. For each algorithm, the model is computed using 81 predictor variables and the optimized hyperparameters (see Table D in the online appendix), which produce the highest predictive performances. University grades and credit points are excluded here because of the poor data quality in these two variables and the relative late time in the study when the variables are raised. These predictive performances are given in terms of AUC, RMSE and the aggregated score. The best model is the one with both low RMSE and high AUC. The average and the standard deviation obtained in 20 repetitions of the 10-fold cross-validation are reported. The best result for each data set (different study fields) is shown in bold.

RF and AB, both tree-based algorithms, achieve the best test results in terms of AUC, RMSE and a combined score higher than 1.60 in each data set. Within the Mathematics and Natural Sciences data, the highest combined score (1.64) is achieved by the RF algorithm, which gives an AUC of 0.89 and a RMSE of 0.25. The Logit and SVM models perform equally well, generally producing an average AUC of 0.82 and an average RMSE of 0.26. The lowest results are obtained by the baseline classifier NB with an average combined score of 1.42 and an average RMSE of 0.38, which is by far the worst produced RMSE. The main reason for the poor results of the NB classifier is that the assumption of independence of the features is not fulfilled (which is mostly the case). The baseline algorithm is outperformed by all the other algorithms, which suggest that

Table 5.3: Predictive results of the classification methods.

Dataset	Model	AUC	RMSE	AUC+(1-RMSE)
Engineering (17.07%) Dropout: 179 Graduate: 1,351	Logit	0.83 ± 0.05	0.29 ± 0.02	1.54 ± 0.06
	NB	0.81 ± 0.06	0.38 ± 0.04	1.43 ± 0.08
	SVM	0.81 ± 0.05	0.29 ± 0.02	1.52 ± 0.06
	RF	**0.88 ± 0.04**	**0.26 ± 0.03**	**1.62 ± 0.06**
	AB	0.88 ± 0.05	0.27 ± 0.03	1.61 ± 0.06
Mathematics and Natural Sciences (24.24%) Dropout: 211 Graduate: 1,962	Logit	0.85 ± 0.04	0.26 ± 0.02	1.59 ± 0.05
	NB	0.82 ± 0.04	0.38 ± 0.03	1.44 ± 0.07
	SVM	0.84 ± 0.04	0.26 ± 0.02	1.58 ± 0.05
	RF	**0.89 ± 0.04**	**0.25 ± 0.02**	**1.64 ± 0.05**
	AB	0.89 ± 0.04	0.26 ± 0.03	1.63 ± 0.05
Law, Economics and Social Sciences (29.47%) Dropout: 199 Graduate: 2,443	Logit	0.80 ± 0.04	0.25 ± 0.02	1.55 ± 0.06
	NB	0.78 ± 0.05	0.38 ± 0.03	1.40 ± 0.06
	SVM	0.80 ± 0.05	0.25 ± 0.02	1.55 ± 0.06
	RF	**0.83 ± 0.05**	**0.24 ± 0.02**	**1.59 ± 0.06**
	AB	**0.83 ± 0.05**	**0.24 ± 0.03**	**1.59 ± 0.06**
Linguistics and Cultural Sciences (29.22%) Dropout: 221 Graduate: 2,398	Logit	0.81 ± 0.04	0.26 ± 0.02	1.55 ± 0.05
	NB	0.79 ± 0.04	0.37 ± 0.03	1.42 ± 0.06
	SVM	0.80 ± 0.04	0.26 ± 0.02	1.54 ± 0.06
	RF	**0.84 ± 0.05**	**0.24 ± 0.02**	**1.60 ± 0.06**
	AB	**0.84 ± 0.04**	**0.24 ± 0.03**	**1.60 ± 0.06**
Model with all 4 subject groups Dropout: 810 Graduate: 8,154	Logit	0.82 ± 0.02	0.26 ± 0.01	1.56 ± 0.03
	NB	0.81 ± 0.02	0.38 ± 0.01	1.43 ± 0.03
	SVM	0.84 ± 0.02	0.26 ± 0.01	1.58 ± 0.03
	RF	**0.87 ± 0.02**	**0.24 ± 0.01**	**1.63 ± 0.03**
	AB	**0.87 ± 0.02**	**0.24 ± 0.01**	**1.63 ± 0.03**

the different tested learning methods achieved good predictive performances. The results also show that the best classification results are obtained for students of Mathematics and Natural Sciences, Engineering and the model with all subject groups. In contrast to that, students of Law, Economics and Social Sciences and Linguistics and Cultural Sciences are harder to classify by any of the machine learning algorithms. As an example of the quality of the predictions, Figure 5.2 shows the ROC-curves for the different algorithms within the Mathematics and Natural Sciences dataset.

5.5.2 Variable importance

The RF method is also commonly used for ranking variables based on the RF variable importance measures. To obtain an unbiased variable ranking, it is recommended to use an AUC-based importance measure (Janitza et al., 2013) instead of the widely used mean decrease of Gini impurity since this approach prefers variables with high variance

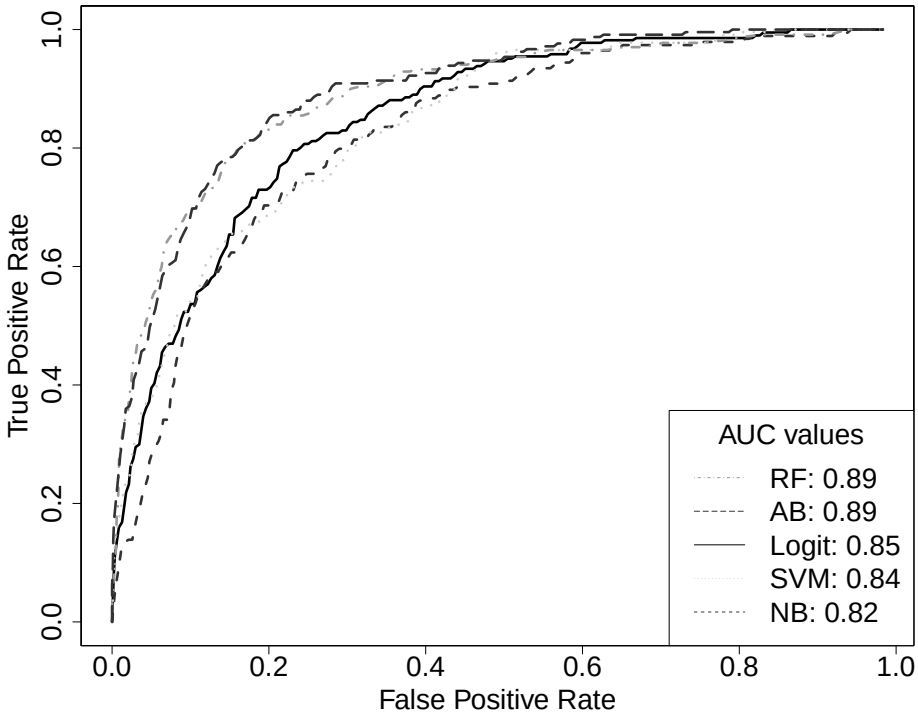

Figure 5.2: ROC-curves of the different algorithms in the Mathematics and Natural Sciences data

(Boulesteix et al., 2011). To calculate an unbiased AUC-based importance measure, we follow the approach of Strobl et al. (2007) who used a subsampling (sampling without replacement) strategy with conditional inference trees introduced by Hothorn et al. (2006). First, the AUC for the original model is calculated. In a second step, the variable of interest is randomly permuted and the AUC for this model is computed. The greater the difference in terms of AUC, the more important is the variable for dropout prediction. For a better interpretation, we provide a relative importance measure by dividing the absolute values by the sum of all absolute values.

In Table 5.4, we provide the importance ranking of the different variables for the overall model and within each subject group (the 10 most important variables are listed). Important pre-university phase predictors are the final grade at secondary school (most important variable), year of birth, as well as the number of repeated school classes. Important predictors from the early study phase are predictors describing one's own evaluation of study performance (satisfaction with academic performance - performance_eval, study progress match to the curriculum plan - workload_match), predictors describing

students' subjective self-assessment of success (e.g. opinion on the probability of graduating - probsuccess, perception of talent for studying - selfconcept), determinants describing satisfaction with studies (e.g. satisfaction with the actual studies - satisf_whole, wanting better study conditions - satisf_conditions), as well as determinants describing commitment to study (not do more than necessary - commit_necessary, high demands on self - commit_demands). We can also remark that no variable from the decision phase belongs to the most important predictors. Especially in the "hard" study fields Engineering, Mathematics and Natural Sciences, there is a large gap of relative importance after the two most important variables.

5.5.3 Model improvement

Although the predictive performances obtained by the algorithms from the first models are acceptable, we aim at examining several methods, which may contribute to improve the performance and appropriability of the models. We suppose that there are uninformative features among the predictor variables. Apart from computational overhead when fitting high dimensional data with complex models, it is also prone to overfitting. Removing these uninformative variables and including only the relevant variables may yield an improvement of some of the models. Additionally, we apply stacking, which is an efficient ensemble method used for improving predictive performance.

Feature selection

The LASSO regularization method, used to reduce overfitting when computing a logistic regression model, can also be applied as a powerful feature selection technique. LASSO regularization does the selection by tossing out less important variables from the model and the coefficients of these variables are shrunken to zero. This process automatically selects significant variables. A further advantage is that only one feature among two or more highly correlated variables is selected and the algorithm is extremely fast (Hastie and Qian, 2014).

We compute the LASSO technique in every single loop of the cross-validation to avoid a bias in the AUC and RMSE estimates (Hastie et al., 2009). On average, 20 relevant features are selected and at each time, grade at secondary school, age, number of repeated

Table 5.4: AUC-based relative importance of the predictors.

	Overall model		Engineering		Mathematics and Natural Sciences		Law, Economics and Social Sciences		Linguistics and Cultural Sciences	
1	grade_school	14.33%	grade_school	16.28%	grade_school	18.79%	grade_school	14.30%	grade_school	18.43%
2	probsuccess	13.39%	probsuccess	16.23%	probsuccess	15.01%	performance_eval	9.02%	probsuccess	12.85%
3	performance_eval	8.05%	performance_eval	5.44%	socint_students	7.23%	birthyear	8.07%	performance_eval	10.99%
4	selfconcept	6.49%	satisf_whole	5.14%	rep_class	5.00%	probsuccess	7.30%	birthyear	9.33%
5	satisf_whole	6.02%	satisf_conditions	4.59%	commit_necessary	4.91%	satisf_whole	6.97%	socint_students	3.90%
6	socint_students	4.89%	selfconcept	4.49%	satisf_conditions	3.97%	workload_match	3.80%	gender	3.43%
7	birthyear	4.04%	satisf_match	3.89%	satisf_whole	3.47%	socint_students	3.14%	commit_demands	2.97%
8	satisf_conditions	3.74%	preparation	3.55%	performance_eval	3.22%	selfconcept	3.07%	selfconcept	2.93%
9	commit_demands	2.88%	satisf_interesting	3.41%	selfconcept	3.01%	school_type	2.74%	qualif_max	2.85%
10	satisf_kill	2.75%	study_break	3.33%	satisf_kill	2.88%	satisf_interesting	2.71%	satisf_whole	2.58%

school classes, study satisfaction, study alternative, social integration at university, study commitment and the time spend on off-study work during term break are among the most important variables. With the selected features, we (re)-compute the different models. For the logit model, we now also consider terms of quadratic order and interactions of second order as explanatory variables (so far, we included only the main effects). For example, in the dataset containing all 4 subject groups 22 relevant variables were selected, which leads to an overall number of 275 variables (22 first order variables + 22 quadratic forms + $\binom{22}{2} = 231$ interactions of second order). Table 5.5 summarizes the results.

Table 5.5: Predictive results of the classification models computed using only the selected variables. Improvement in the results, compared to the results when all the variables are used for modeling, are shown in bold. The best models are underlined.

Dataset	Model	AUC	RMSE	AUC+(1-RMSE)
Engineering	Logit	0.85 ± 0.04	0.28 ± 0.03	1.57 ± 0.06
	NB	0.82 ± 0.05	0.36 ± 0.03	1.46 ± 0.08
	SVM	0.84 ± 0.04	0.29 ± 0.02	1.55 ± 0.06
	RF	**0.88 ± 0.04**	**0.26 ± 0.03**	**1.62 ± 0.06**
	AB	0.88 ± 0.05	0.27 ± 0.03	1.61 ± 0.06
Mathematics and Natural Sciences	Logit	0.86 ± 0.04	0.26 ± 0.02	1.60 ± 0.05
	NB	0.84 ± 0.04	0.35 ± 0.03	1.49 ± 0.06
	SVM	0.85 ± 0.04	0.26 ± 0.02	1.59 ± 0.05
	RF	**0.89 ± 0.04**	**0.25 ± 0.02**	**1.64 ± 0.05**
	AB	0.89 ± 0.04	0.26 ± 0.03	1.63 ± 0.05
Law, Economics and Social Sciences	Logit	0.81 ± 0.04	0.25 ± 0.02	1.56 ± 0.06
	NB	0.80 ± 0.05	0.33 ± 0.03	1.47 ± 0.06
	SVM	0.82 ± 0.05	0.24 ± 0.02	1.58 ± 0.06
	RF	**0.83 ± 0.05**	**0.24 ± 0.02**	**1.59 ± 0.06**
	AB	**0.83 ± 0.05**	**0.24 ± 0.03**	**1.59 ± 0.06**
Linguistics and Cultural Sciences	Logit	0.82 ± 0.04	0.26 ± 0.02	1.56 ± 0.05
	NB	0.79 ± 0.04	0.35 ± 0.03	1.44 ± 0.06
	SVM	0.82 ± 0.04	0.25 ± 0.02	1.57 ± 0.05
	RF	**0.84 ± 0.05**	**0.24 ± 0.02**	**1.60 ± 0.06**
	AB	**0.84 ± 0.04**	**0.24 ± 0.03**	**1.60 ± 0.06**
Model with all 4 subject fields	Logit	0.83 ± 0.02	0.26 ± 0.01	1.57 ± 0.03
	NB	0.81 ± 0.02	0.35 ± 0.01	1.46 ± 0.03
	SVM	0.84 ± 0.02	0.25 ± 0.01	1.59 ± 0.03
	RF	**0.87 ± 0.02**	**0.24 ± 0.01**	**1.63 ± 0.03**
	AB	**0.87 ± 0.02**	**0.24 ± 0.01**	**1.63 ± 0.03**

We observe, that performing the feature selection and using only the important variables improves the performance of some models. Logit, SVM and NB improve their performances in almost all the evaluation measures and within almost each data set.

For example, within the Law, Economics and Social Sciences the AUC and the combined score are boosted from 0.80 to 0.81 and from 1.55 to 1.56 for the Logit model, from 0.78 to 0.80 and from 1.40 to 1.47 for the NB algorithm, from 0.80 to 0.82 and from 1.55 to 1.58 for the SVM algorithm. In addition, the RMSE is reduced from 0.38 to 0.33 for NB and from 0.25 to 0.24 for SVM. This illustrates that for the NB and the SVM algorithms, a prior selection of the important variables before training the models is useful to obtain higher performances. Besides, for the logit algorithm, considering interaction among the selected explanatory variables and also terms of quadratic order improves the predictive performance of the models.

Nonetheless, it is observed that the RF and the AB models still obtain the same good (and best) performance as when using all the variables. This can be explained by the fact that both tree-based algorithms include a type of inner feature selection by searching at each time the variable which maximizes the information gain.

Model stacking for improving predictive performance

Introduced by Wolpert (1992), stacked generalization is an efficient ensemble method used for improving and boosting predictive performance. Also known as model stacking, this technique is sometimes referred to as a "wisdom of crowds" approach and works in two phases. First, multiple learning algorithms are applied to predict the class. Second, a new learner to combine their predictions with the aim of reducing the generalization error is used. Sill et al. (2009) illustrate it as a two-level learning method, in which the predictions generated from different base classifiers in the first-level are used as inputs in a second-level learning algorithm (generally a logistic regression) that is trained to optimally combine the model predictions to form a final set of predictions. An important advantage of model stacking is that it averages out the noise from diverse models and usually overcomes the problem of under-fitting when complex datasets are trained using simple classifiers. On the other hand, over-fitting is minimized by using nested cross-validation.

We apply the stacking technique on the dataset along with our five selected learning algorithms, which are trained in the first level using 10-fold cross-validation repeated 20 times. Starting with the combination of predictions obtained from the Logit and the NB models, stacked ensemble models are gradually created and evaluated using logistic regression as second-level learning method. Table 5.6 reports the results of the analysis.

We observe that within each subject field the AUC value improves when the prediction of logit and NB are stacked together compared to the results in Table 5.3. Adding the predictions generated from the SVM classifier into the stacked ensemble only improves the models in the Mathematics and Natural Sciences and also in the Law, Economics and Social Sciences data sets as well as in the global data set. An interesting finding is that stacking the predictions of the first four algorithms (Logit, NB, SVM and AB) together does not produce better predictive performances compared to training AB uniquely. The fact that this ensemble does not lead to a model improvement highlights a high correlation between the predictions of the different algorithms. Combining the RF predictions with all the other predictions outputs the best results. This stacked ensemble reaches, for example in the Mathematics and Natural Sciences data set, an AUC value of 0.90 and an RMSE value of 0.24 for a combined score of 1.66 (in bold).

Table 5.6: Predictive results of the classification models computed based on the stacking approach and using all the variables.

Dataset	Model	AUC	RMSE	Score
Engineering	Logit(Logit+NB)	0.84 ± 0.04	0.29 ± 0.03	1.55 ± 0.06
	Logit(Logit+NB+SVM)	0.84 ± 0.05	0.29 ± 0.03	1.55 ± 0.07
	Logit(Logit+NB+SVM+AB)	0.88 ± 0.04	0.27 ± 0.03	1.61 ± 0.06
	Logit(Logit+NB+SVM+AB+RF)	0.89 ± 0.04	0.26 ± 0.03	1.63 ± 0.06
Mathematics and Natural Sciences	Logit(Logit+NB)	0.86 ± 0.04	0.26 ± 0.03	1.60 ± 0.05
	Logit(Logit+NB+SVM)	0.87 ± 0.04	0.26 ± 0.02	1.61 ± 0.05
	Logit(Logit+NB+SVM+AB)	0.89 ± 0.03	0.25 ± 0.03	1.64 ± 0.05
	Logit(Logit+NB+SVM+AB+RF)	**0.90 ± 0.04**	**0.24 ± 0.03**	**1.66 ± 0.06**
Law, Economics and Social Sciences	Logit(Logit+NB)	0.80 ± 0.05	0.25 ± 0.02	1.55 ± 0.06
	Logit(Logit+NB+SVM)	0.82 ± 0.04	0.25 ± 0.02	1.57 ± 0.05
	Logit(Logit+NB+SVM+AB)	0.84 ± 0.04	0.24 ± 0.02	1.60 ± 0.05
	Logit(Logit+NB+SVM+AB+RF)	0.84 ± 0.05	0.23 ± 0.02	1.61 ± 0.06
Linguistics and Cultural Sciences	Logit(Logit+NB)	0.82 ± 0.04	0.25 ± 0.02	1.57 ± 0.06
	Logit(Logit+NB+SVM)	0.82 ± 0.04	0.25 ± 0.02	1.57 ± 0.05
	Logit(Logit+NB+SVM+AB)	0.85 ± 0.04	0.24 ± 0.02	1.61 ± 0.06
	Logit(Logit+NB+SVM+AB+RF)	0.85 ± 0.04	0.23 ± 0.02	1.62 ± 0.06
Model with all 4 subject fields	Logit(Logit+NB)	0.83 ± 0.02	0.26 ± 0.01	1.57 ± 0.03
	Logit(Logit+NB+SVM)	0.84 ± 0.02	0.26 ± 0.01	1.58 ± 0.03
	Logit(Logit+NB+SVM+AB)	0.87 ± 0.02	0.24 ± 0.01	1.63 ± 0.03
	Logit(Logit+NB+SVM+AB+RF)	0.88 ± 0.02	0.24 ± 0.01	1.64 ± 0.03

5.5.4 Discussion

The results of the model comparison show the effectiveness of some machine learning techniques in the dropout context and the application of different model improvement approaches prove to be powerful techniques for optimizing predictive performances, which is an important issue when aiming at dropout prediction. Our findings demonstrate that tree-based ensemble methods generally achieve the best prediction performances of up to 89% and should be preferred to benchmark classifiers. Compared to other models, these algorithms also have other important advantages for application in the higher education context. First, they are generally able to deal with missing values, which is a very relevant problem in educational data sets. As stated earlier, there are some software packages where the RF cannot handle missings, but in general, decision trees handle missings with surrogate variables. Many other algorithms (e.g. Logit, SVM) require complete data and therefore the application of (complex) imputation techniques which may induce biases of the results. Furthermore, tree-based models are not based on the conditional independence assumption (as e.g. NB), which suggests that the values of the attributes are conditionally independent, which is mostly not the case when analyzing the complex process of student dropout from university. Determinants of academic success usually are of various types, e.g. nominal, ordinal or metric scaled. Tree-based ensemble models such as RF can handle each type of attribute without transformation procedures, which may further induce inaccuracies. Additionally, when aiming at dropout prediction, the underlying complex process does not allow for determining all relevant features a priori and therefore, one often has to deal with many attributes of which some of them may be unimportant. Here, tree-based models have the advantage of implementing a type of inner feature selection and of yielding good results also when including non-relevant features. In contrast, the results of NB or SVM worsen by using unimportant features and require a prior feature selection which increases complexity and computational overhead when trying to avoid biases of the estimates. A further advantage of tree-based models is that they are very robust against outliers which may also drive the outcome of a prediction model. At least, as tree-based models do not require complex adjustments and are more intuitive than many other algorithms, early warning systems based on such models may be easier to implement and results may be easier to communicate to practitioners.

Usually, higher education institutions not only aim at predicting students as dropouts or graduates but also want to know which circumstances are mainly relevant for students to

be in the at-risk group. In contrast to many other algorithms, tree-based models provide informative and intuitive variable importance rankings which may help institutions to identify very relevant attributes and to implement special programs to support students at risk. Not all variables used in this analysis are available at the beginning of study and only some of them are included in institutions' administrative data collected during enrolment. However, the results based on a broad data set provide useful information for institutions or faculties on how, for instance, voluntary questionnaires for students at the beginning of studies and during the first semesters should be designed to get very important information for dropout prediction.

When using SVM or Logit to develop efficient early warning systems, prior feature selection seems to be a necessary tool for model improvement. Moreover, more complex algorithms like model stacking improve predictive performances. Here, the advantage is that instead of choosing the best model, all models can be considered and incorporated in the system. An early warning system based on a stacked ensemble model would benefit from averaging out the noise from diverse models and usually overcomes the problem of under-fitting when complex datasets are trained using simple classifiers. But apart from the problem of computation time, which can drastically increase when all models have to be calculated, such models are not easy to implement and to interpret.

We further observe that there are differences in the prediction performance across different study fields. In Natural Sciences/Mathematics and Engineering dropout prediction seems to be more precise than, for instance, in Linguistics. This may hint for other driving factors in such more "soft" study fields and indicates to implement field-specific early warning systems.

In general, it should be kept in mind that the results may depend on the research question, the used data set and variables and on the analyzed setting (e.g. online courses vs. traditional classroom). Nonetheless, findings of this study can be utilized as assistance for analysts of higher education institutions to implement or improve early warning systems to accurately detect students at risk for dropping out. This may be relevant in the context of dropping out from traditional studies as well as from online courses or distance learning, which became increasingly important in higher education over the last years.

Moreover, it should be noted that the presented algorithms cannot address specific prob-

lems where not enough data is available. Therefore, human student counseling cannot be replaced completely by prediction models, but such models may be a suitable tool for providing assistance in identifying students at risk and to develop specific programs to prevent students from dropping out and thereby to improve the institutions educational effectiveness.

5.6 Conclusion

In this study, we apply advanced machine learning algorithms to predict a student's probability to leave university without a degree. Dropping out from higher education represents a very important issue and higher education institutions are increasingly searching for efficient early warning systems and programs to identify and help students at risk. Our analysis is based on a broad German data set including first year students and covering a wide range of variables from several steps in the study course.

Correctly identifying students at risk constitutes a challenging task since there are many interacting factors leading a student to withdraw from university without degree. We develop a range of machine learning models, which are sophisticated and enhanced enough to address this complexity and very precise predictions are achieved.

After searching for the optimal hyperparameter settings for each model, the best performances with AUC values up to 0.89 are obtained by the two tree-based classifiers, random forest and AdaBoost, which are 14% better than the benchmark classifier naive Bayes. To further improve prediction performance, we apply feature selection and stacking, which is the combination of different classifiers and makes optimal use of the strength of each individual algorithm. An improvement of the models is observed and the highest predictive performance (AUC = 0.90) is achieved by combining the RF predictions with all the other prediction outputs. The most important dropout predictor is the final grade at secondary school.

After some discussions of advantages and disadvantages of the different models, we conclude that using random forests for developing prediction models for student dropout represents a promising compromise between a good predictive performance, a straightforward interpretation of the results and an easy implementation since the RF al-

gorithm is already included in many software environments for statistical computing (e.g. R).

The findings confirm that student dropout could be predicted very precisely and results could serve as assistance for practitioners and student counselors in higher education institutions for implementing efficient early warning systems and to develop promising programs to support students at risk.

5.7 Appendix

Support Vector Machines

Three cases are distinguished when computing hyperplanes: data are linear separable, data are linear non separable or data are non-linear.

The linearly separable case: In this case, the data can cleanly be separated by a linear hyperplane, so data points belonging to class -1 are all on the same side of the hyperplane, data points belonging to class $+1$ are on the other side. To find the optimal hyperplane, the vertical distance between the hyperplane and the nearest training observations is maximized. Thereby, the position of the linear hyperplane is only influenced by the training data points closest to the hyperplane, the so called support vectors. Figure (5.3) illustrates the linear separable case in the left panel.

Given n training observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, a d-dimensional vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ indicating the normal direction to the hyperplane and a bias β_0, the hyperplane can be expressed in the following form:

$$\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta} + \beta_0 = 0\}. \tag{5.12}$$

The margin, which is the double of the distance between the hyperplane and the nearest data points has the form:

$$\text{Margin} = 2M = \min_{i=1,\ldots,n} 2\frac{|\mathbf{x}_i^T\boldsymbol{\beta} + \beta_0|}{\|\boldsymbol{\beta}\|}. \tag{5.13}$$

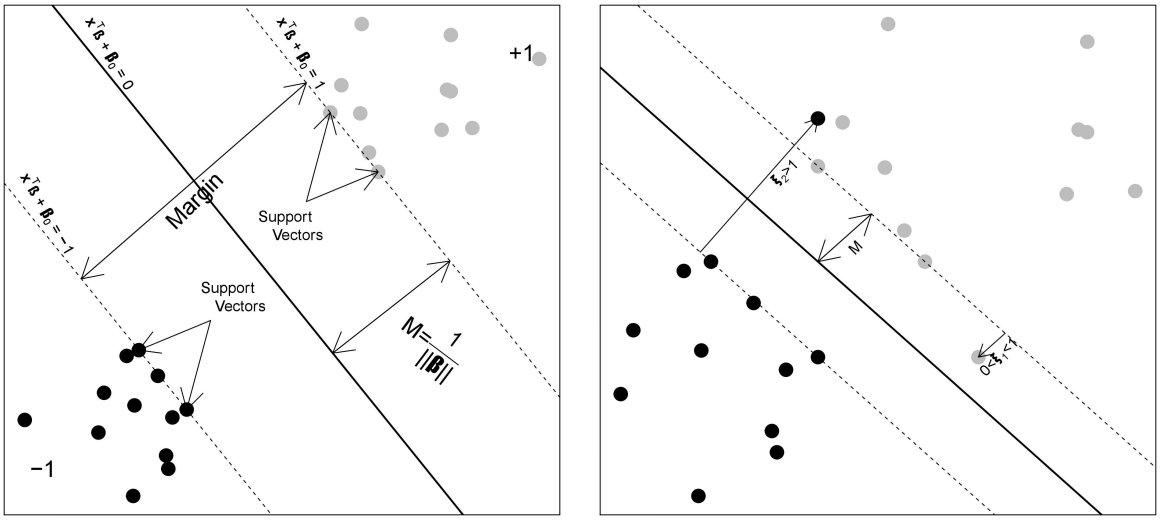

Figure 5.3: Left panel: linearly separable case. There are two support vectors in each class. The solid line is the linear hyperplane and the distance between the two dashed lines indicates the margin. Right panel: linearly non-separable scenario. One data point of each class is on the wrong side of the margin lines.

A new observation $\mathbf{x}_{new}$ is classified to class -1, if $\mathbf{x}_{new}^T\boldsymbol{\beta}+\beta_0<0$ and to class $+1$, if $\mathbf{x}_{new}^T\boldsymbol{\beta}+\beta_0>0$. This results in $\left|\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0\right|=y_i(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0)>0$ for all data points. In order to simplify the form of the margin, Bishop (2006) proposes to rescale the parameter $\boldsymbol{\beta}\longrightarrow k\boldsymbol{\beta}$ and $\beta_0\longrightarrow k\beta_0$ such that for the margin points, also called support vectors (data points closest to the hyperplane), we have $y_i(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0)=1$, and for the data points lying outside the hyperplane $y_i(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0)>1$. Due to the rescaling, the margin has the form

$$\text{Margin}=2M=\min_{i=1,\ldots,n}2\frac{|\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0|}{\|\boldsymbol{\beta}\|}=\frac{2}{\|\boldsymbol{\beta}\|}. \tag{5.14}$$

We want to maximize the margin $2/\|\boldsymbol{\beta}\|$, or equivalently minimize $\|\boldsymbol{\beta}\|^2/2$ (we take the power of two for later convenience), under the assumption that every training data point is correctly classified, i.e. $y_i(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0)\geq 1$. This leads to the minimization problem

$$\underset{\boldsymbol{\beta},\beta_0}{\text{argmin}}\ \frac{1}{2}\|\boldsymbol{\beta}\|^2,\quad \text{subjected to } y_i(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0)\geq 1, i=1,\ldots,n. \tag{5.15}$$

Equation 5.15 can be solved by introducing Lagrange multiplier $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_n)^T\geq\mathbf{0}$,

with the Lagrange function

$$\mathbb{L}_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) - 1]. \tag{5.16}$$

The derivatives of the Lagrangian function with respect to $\boldsymbol{\beta}$ and β_0 set to zero gives $\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. Subsituting these results in equation (5.16) leads to the so called Lagrangian dual optimization problem

$$\underset{(\alpha_1,\cdots,\alpha_n)}{\operatorname{argmax}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j. \tag{5.17}$$

under the constraints $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. The Lagrangian dual problem can be solved with standard statistic software, more details about solving the Lagrangian dual are in Aggarwal (2015). Once the Lagrange multiplier $\boldsymbol{\alpha}$ is determined, values of the parameter vector $\boldsymbol{\beta}$ and the bias β_0 can be computed.

The linearly non-separable case: In practice, data with complete linearly separable data points occurs very rarely since there generally exists misclassified points among the data. Therefore, we need to adjust the support vector machine in order to allow training points to be misclassified. To do that, we introduce slack variables $\xi_i, i = 1, \ldots, n$ with $\xi_i \geq 0 \ \forall i$ and $\sum_{i=1}^{n} \xi_i \leq \textit{constant}$, which measures the distance of the misclassified point to its marginal hyperplane. For simplification, we use the relative distance $\xi_i^* = \xi_i / \|\boldsymbol{\beta}\|$. This modification leads to

$$y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i^*, i = 1, \ldots, n. \tag{5.18}$$

If $\xi_i^* = 0$, the ith training observation is on the correct side of its marginal hyperplane. Else if $0 < \xi_i^* \leq 1$, the data point lies inside the margin, but on the wrong side of its marginal hyperplane. Otherwise $\xi_i^* > 1$, the data lies on the wrong side of the optimal hyperplane. In the right panel of figure (5.3) one can see that $0 < \xi_1^* < 1$ and $\xi_2^* > 1$. To consider a combined effect, i.e. maximization of the margin and soft penalization of errors of some magnitude, we use a variable C as penalty strength, which quantifies how much we treat the training points lying on the wrong side. So the minimization problem

in equation (5.15) results in

$$\underset{\boldsymbol{\beta},\beta_0}{\operatorname{argmin}} \ \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}\xi_i^*, \text{ subjected to } y_i(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0) \geq 1-\xi_i^*,\ \xi_i^* \geq 0\ \forall i. \tag{5.19}$$

Large values of the tuning parameter C result in small margins, whereas small values of C result in wide margins. For the Lagrange function we get

$$\mathbb{L}_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}\xi_i^* - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0) - 1 + \xi_i^*)] - \sum_{i=1}^{n}\mu_i\xi_i^*, \tag{5.20}$$

with Lagrange multipliers $\alpha_i \geq 0, \mu_i \geq 0$. Derivatives of the Lagrange function set to zero give $\boldsymbol{\beta} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i$, $\sum_{i=1}^{n}\alpha_i y_i = 0$ and $\alpha_i = C - \mu_i$, $\forall i$. Substituting these results in equation (5.19) leads to the Lagrangian dual optimization problem

$$\underset{(\alpha_1,\cdots,\alpha_n)}{\operatorname{argmax}} \ \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j, \tag{5.21}$$

under the constraints $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{n}\alpha_i y_i = 0$. Solution of this problem is the estimate of the vector $\boldsymbol{\beta}$, that defines the hyperplane:

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n}\hat{\alpha}_i y_i \mathbf{x}_i. \tag{5.22}$$

The Lagrangian dual problem in this case looks equal to the linear separable case in equation (5.17). Only the data points on the wrong side of the margin or on the edge of the margin influence the location of the hyperplane: these data points are the support vectors and satisfiy $y_i(\mathbf{x}_i^T\boldsymbol{\beta}+\beta_0) = 1-\xi_i$.

The non-linear case: In many classification problems, the data points are non-linear separable. Using a linear hyperplane as decision boundary would lead to poor results. For solving a non-linear problem, a kernel trick can be applied. This consists of transforming the data points from the input space into a different space called **feature space** using a kernel function so that the linear classifier can be fitted in the feature space. Common choices of the kernel function are:

- linear $k(\mathbf{x}, \mathbf{x'}) = \mathbf{x}^T\mathbf{x'}$

- polynomial kernel $k(\mathbf{x}, \mathbf{x'}) = (\mathbf{x}^T\mathbf{x'} + c)^h$
- Gaussian radial basis kernel $k(\mathbf{x}, \mathbf{x'}) = \exp(-\|\mathbf{x} - \mathbf{x'}\|^2/2\sigma^2)$.

A kernel function behaves like a scalar product in $\mathbb{R}$. Its parameters have to be tuned. For example, a high value of the polynomial degree h or a small value of σ in the radial basis kernel can lead to overfitting of the training data. On the other hand, a too small value of h can also lead to poor classification results, if the real data is of degree $h + k$ with $0 < k \in \mathbb{N}$. The Lagrangian function with the transformed points is given by:

$$\underset{(\alpha_1,\cdots,\alpha_n)}{\operatorname{argmax}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j). \tag{5.23}$$

AdaBoost

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the $d-$dimensional data points and $y_1, \ldots, y_n$ with $y_i \in \{-1, 1\}$, $i = 1, \ldots, n$, value of a binary target variable . The discrete AdaBoost algorithm, described in Hastie et al. (2009), starts with equal weights for all training observations ($w_i = 1/n$, $i = 1, \ldots, n$) in the first iteration. In the iteration steps $m = 1, \ldots, M$, the algorithm repeats the following steps:

1. A weak learner $f_m(\mathbf{x})$ with the actual weights w_i is fitted on the training data. For $f_m(\mathbf{x})$, we use a single decision tree (CART) (Breiman et al., 1984). Note that any other learner can be used for $f_m(\mathbf{x})$, but generally weak learners (in the sense of fast computation time, which usually do not perform well as single classifiers) are used.
2. Compute the weighted error rate $\varepsilon_m = \frac{\sum_{i=1}^{n} w_i I_{[y_i \neq f_m(\mathbf{x}_i)]}}{\sum_{i=1}^{n} w_i}$, where I is the indicator function, and also compute the estimator coefficient $c_m = \nu \log(\frac{1-\varepsilon_m}{\varepsilon_m})$, where $0 < \nu \leq 1$ is the learning rate,
3. Adjust the weights $w_{i,new} = w_{i,old} \exp(c_m I_{[y_i \neq f_m(\mathbf{x}_i)]})$, for $i = 1, \ldots, n$ and normalize the new weights so that $\sum_{i=1}^{n} w_{i,new} = 1$.

Prediction Result of an Extended Model using an Alternative Dropout Definition

Higher education dropout is not always defined consistently in the literature. The definition used in the article is from a macro point of view and defines dropout as students who leave the higher education system without a degree. This definition is best suited to our data and our research question. For comparison, we also use an alternative dropout definition including students who change their subject field before graduation. This second definition comes from a micro point of view, that of a faculty, for which changes in the study field before the first degree could represent a failure in their goal of avoiding dropout from the study program. Due to data limitations, we only observe four major aggregated subject groups (1) Engineering, (2) Mathematics and Natural Sciences, (3) Law, Economics and Social Sciences, and (4) Linguistics and Cultural Sciences, what makes it difficult to identify dropouts from a specific faculty. For the second definition, the sample size amounts to 9,673 students (2,186 dropouts, 7,487 graduates). Moreover, two new variables which are often used in previous studies, the number of credit points and the average grades until the end of the second semester, are included as predictive variables.

We do not include these variables in our main classification models for some reasons. First, as grades and credits also could be regarded as an academic output, it seems to be tautological to predict academic success with another information on academic success. Obviously, using these output variables as predictors, model performance would increase and grades or credits would be the most important determinants probably masking other important aspects. As long as only model performance do play a role in analysis, this procedure may be the best approach. However, additionally to finding the best prediction model, the aim of universities is to detect starting points for early intervention measures, preferably before students obtain bad grades or do not pass some tests. A second reason for not including grades or credit points is the fact that many dropouts do not participate at the first test phase and an imputation of missing values may yield biased results.

Table C provides the results of the five algorithms for this extended model. Although, as expected, the two new variables have high importance and are two of the best predictors in the importance ranking, the data quality suffers under the high amount of missing values in the obtained grades (51.91% missings) and credit points (66.24% missings).

Note that these variables may be included if administrative data is used to improve the predictive performance of the models. In our model, these two variables improve each model by approximately 1-2% in terms of AUC and RMSE (no matter which status variable is used). The two reasons for the relatively small model improvement are the data quality (many missings) and the high correlation to other prediction variables, e.g. the correlation between school grades and the grades after the second semester is 0.26. Furthermore, Table C shows worse results compared to the first dropout definition. The reason for this is that the data predominately predicts the performance of the student and performance has a higher correlation with the first dropout definition. Field changes are often not caused by poor study performance but rather by the wish for a new orientation. The dropout definition from the faculties' point of view may be applied if administrative data is used since this information can easily be provided by the faculties examination offices.

Table C: Predictive results of the extended model.

Dataset	Model	AUC	RMSE	AUC+(1-RMSE)
Engineering (16.74%) Dropout: 308 Graduate: 1,311	Logit	0.78 ± 0.05	0.35 ± 0.03	1.43 ± 0.07
	NB	0.78 ± 0.05	0.44 ± 0.04	1.34 ± 0.08
	SVM	0.77 ± 0.05	0.35 ± 0.03	1.42 ± 0.07
	RF	0.83 ± 0.05	**0.33 ± 0.03**	1.50 ± 0.07
	AB	**0.84 ± 0.04**	**0.33 ± 0.03**	**1.51 ± 0.07**
Mathematics and Natural Sciences (25.87%) Dropout: 668 Graduate: 1,835	Logit	0.81 ± 0.04	0.37 ± 0.03	1.44 ± 0.05
	NB	0.76 ± 0.04	0.47 ± 0.03	1.29 ± 0.06
	SVM	0.80 ± 0.04	0.37 ± 0.03	1.43 ± 0.05
	RF	0.82 ± 0.03	0.37 ± 0.02	1.45 ± 0.05
	AB	**0.84 ± 0.03**	**0.36 ± 0.03**	**1.48 ± 0.05**
Law, Economics and Social Sciences (27.84%) Dropout: 448 Graduate: 2,245	Logit	0.75 ± 0.04	0.34 ± 0.02	1.41 ± 0.06
	NB	0.73 ± 0.05	0.44 ± 0.03	1.29 ± 0.07
	SVM	0.75 ± 0.04	0.34 ± 0.02	1.41 ± 0.07
	RF	0.77 ± 0.05	**0.33 ± 0.02**	1.44 ± 0.06
	AB	**0.79 ± 0.04**	**0.33 ± 0.02**	**1.46 ± 0.06**
Linguistics and Cultural Sciences (29.55%) Dropout: 762 Graduate: 2,096	Logit	0.75 ± 0.04	0.40 ± 0.02	1.35 ± 0.06
	NB	0.72 ± 0.05	0.48 ± 0.03	1.24 ± 0.07
	SVM	0.75 ± 0.04	0.40 ± 0.03	1.35 ± 0.07
	RF	0.76 ± 0.04	**0.39 ± 0.02**	1.37 ± 0.06
	AB	**0.77 ± 0.04**	**0.38 ± 0.02**	**1.38 ± 0.06**
Model with all 4 subject groups Dropout: 2,186 Graduate: 7,487	Logit	0.76 ± 0.03	0.37 ± 0.01	1.39 ± 0.03
	NB	0.73 ± 0.03	0.45 ± 0.01	1.28 ± 0.04
	SVM	0.76 ± 0.03	0.37 ± 0.01	1.39 ± 0.03
	RF	0.78 ± 0.02	**0.36 ± 0.01**	1.42 ± 0.03
	AB	**0.79 ± 0.02**	**0.36 ± 0.01**	**1.43 ± 0.03**

Table D: One example for the hyperparameter settings of each model reported in Table 4

Model	Dataset	Hyperparameter
NB	Engineering	laplace = 0
	Mathematics and Natural Sciences	laplace = 0
	Law, Economics and Social Sciences	laplace = 0
	Linguistics and Cultural Sciences	laplace = 0
	Model with all 4 subject fields	laplace = 0
Logit	Engineering	LASSO, alpha = 1, lambda = 0.008
	Mathematics and Natural Sciences	LASSO, alpha = 1, lambda = 0.009
	Law, Economics and Social Sciences	LASSO, alpha = 1, lambda = 0.008
	Linguistics and Cultural Sciences	LASSO, alpha = 1, lambda = 0.0078
	Model with all 4 subject fields	LASSO, alpha = 1, lambda = 0.0081
SVM	Engineering	gaussian kernel, cost = 4, gamma = 0.0039
	Mathematics and Natural Sciences	gaussian kernel, cost = 4096, gamma = 3.05e-05
	Law, Economics and Social Sciences	gaussian kernel, cost = 2, gamma = 0.0078
	Linguistics and Cultural Sciences	gaussian kernel, cost = 4, gamma = 0.0156
	Model with all 4 subject fields	gaussiam kernel, cost = 8, gamma = 0.0084
RF	Engineering	mtry = 13, nodesize = 7
	Mathematics and Natural Sciences	mtry = 7, nodesize = 5
	Law, Economics and Social Sciences	mtry = 6, nodesize = 3
	Linguistics and Cultural Sciences	mtry = 14, nodesize = 3
	Model with all 4 subject fields	mtry = 9, nodesize = 7
AB	Engineering	nu = 0.05, maxdepth = 7
	Mathematics and Natural Sciences	nu = 0.05, maxdepth = 10
	Law, Economics and Social Sciences	nu = 0.05, maxdepth = 15
	Linguistics and Cultural Sciences	nu = 0.02, maxdepth = 2
	Model with all 4 subject fields	nu = 0.1, maxdepth = 5

Table E: Attributes description

Attribute	Description (Data type)
Pre-study phase	
genstat	Generation status (numeric: from 1 = 1st generation to 4 = no immigration background)
immigration	Do you have an immigration background? (binary: 0 = No, 1 = Yes)
rep_class	How many class years have you ever repeated? (numeric: from 0 to 4)
ger_prep	To what extent had you acquired German knowledge and skills before starting university? (numeric: from 1 = not at all to 4 = very much)
math_prep	To what extent had you acquired maths knowledge and skills before starting university? (numeric: from 1 = not at all to 4 = very much)
familylife	With whom did you spend most of your childhood up to the age of 14? (binary: 1 = with biological parents, 0 = else)
school_type	Type of school attended (binary: 1 = upper secondary education, 0 = other types)
qualif_max	School-leaving qualification obtained (numeric: 2 = general university entrance qualification, 1 = university of applied science entrance qualification, 0 = other degrees)
grade_school	Approximate overall grade awarded in the school-leaving certificate (numeric: from 1 to 5)
exam_german	Was German an examination subject for your school-leaving qualification? (binary: 0 = No, 1 = Yes)
exam_adv_german	German as first examination subject for your school-leaving qualification (binary: 0 = No, 1 = Yes)
exam_maths	Was maths an examination subject for your school-leaving qualification? (binary: 0 = No, 1 = Yes)
exam_adv_maths	Maths as first examination subject for your school-leaving qualification (binary: 0 = No, 1 = Yes)
gender	Gender of the person (binary: 1 = Male or 0 = Female)
birthyear	Year of birth of the person (numeric: from 1950 to 1994)
mother_qualif	Highest mother's general school-leaving qualification (numeric: from 0 = No school leaving qualification to 8 = Highest tertiary education)
mother_job	Mother's occupation (ISEI-08) (numeric: from 11.74 to 88.96)
father_qualif	Highest father's general school-leaving qualification (numeric: from 0 = No school leaving qualification to 8 = Highest tertiary education)
father_job	Father occupation (ISEI-08) (numeric: from 11.74 to 88.96)
voctrain	Completed vocational training before university (binary: 0 = No, 1 = Yes)
fail_prestudy	Have you ever dropped out from training before university? (binary: 0 = No, 1 = Yes)
Decision phase	
fieldofchoice	Enrolled in the subject of first choice (binary: 0 = No, 1 = Yes)
institofchoice	Take up the degree at the institute of higher education of choice (binary: 0 = No, 1 = Yes)
study_alternative	Would you rather have started something else instead of a degree? (binary: 0 = No, 1 = Yes)
study_judge_parent	What do your parents think about the fact that you are studying? (numeric: from 1 = does not apply at all to 5 = applies completely)
study_judge_friend	What do your friends think about the fact that you are studying? (numeric: from 1 = does not apply at all to 5 = applies completely)

info_useful_...	Usefulness of information received from parents, friends, current university students, school teachers, professionals employed in the field of interest, media, university counseling, literature, school events, sneak peak at university, job agencies, companies etc. (numeric: from 0 = not used to 4 = very helpful)
study_restrict	Is the study subject to admission restrictions or a selection procedure? (binary: 0 = No, 1 = Yes)
Early study phase	
satisf_enjoy	Really enjoy the studied subject (numeric: from 0 = does not apply to 10 = applies completely)
satisf_conditions	Wish better study conditions (numeric: from 0 = does not apply to 10 = applies completely)
satisf_match	Degree course and other obligations hard to match (numeric: from 0 = does not apply to 10 = applies completely)
satisf_whole	On the whole, satisfied with actual studies (numeric: from 0 = does not apply to 10 = applies completely)
satisf_frustrating	External circumstances of study are frustrating (numeric: from 0 = does not apply to 10 = applies completely)
satisf_kill	Degree course is killing me (numeric: from 0 = does not apply to 10 = applies completely)
satisf_interesting	Degree course is really interesting (numeric: from 0 = does not apply to 10 = applies completely)
satisf_concerns	Concerns of students are not taken into account sufficiently (numeric: from 0 = does not apply to 10 = applies completely)
satisf_tired	Degree course often makes feel tired and exhausted (numeric: from 0 = does not apply to 10 = applies completely)
partic_people	Participation in university events aimed at getting to know people (binary: 0 = No, 1 = Yes)
partic_orga	Participation in university events on study organization (binary: 0 = No, 1 = Yes)
partic_facil	Participation in university events on the use of central facilities (binary: 0 = No, 1 = Yes)
partic_course	Participation in university events on bridging courses (binary: 0 = No, 1 = Yes)
partic_acadskills	Participation in university events on academic skills (binary: 0 = No, 1 = Yes)
preparation	How can you rate your preparation at the start of the university in work techniques, fundamental academic methods etc.? (numeric: from 0 = bad to 4 = good)
skills_prep	Necessary knowledge acquired in maths, German, English and computer science before university (numeric: from 1 = not at all to 4 = very much)
workload_match	Study progress (number of courses, credits earned) match to the curriculum plan (numeric: from 1 = much less to 5 = many more)
performance_eval	Satisfaction with the academic performances till yet (numeric: from 1 = does not apply at all to 4 = applies completely)
probsuccess	Your opinion on the probability that you will graduate (numeric: from 1 = very unlikely to 5 = very likely)
selfconcept	Perception of your talent for studying (numeric: from 1 = low to 7 = high)
study_informed	How well you are informed about the possibilities, limitations etc for your degree course? (numeric: from 1 = very poor to 1 = very good)
socint_instructors	Acceptance by instructors and getting along well with them (numeric: from 1 = does not apply to 4 = applies completely)
socint_students	Successful in establishing contacts and getting along well with classmates (numeric: from 1 = does not apply to 4 = applies completely)

commit_necessary	Commitment to degree course: Do no more than necessary (numeric: from 1 = does not apply to 5 = applies completely)
commit_enjoy	Commitment to degree course: enjoyment of degree program (numeric: from 1 = does not apply to 5 = applies completely)
commit_demands	Commitment to degree course: High demands on self (numeric: from 1 = does not apply to 5 = applies completely)
commit_identificat	Commitment to degree course: Identification with degree program (numeric: from 1 = does not apply to 5 = applies completely)
helplessness	You think you will never get better grades (numeric: from 1 = does not apply to 5 = applies completely)
job_semester	Number of hours spent in a week during semester time for employment (numeric: from 0 to 60)
study_semester	Number of hours spent in a week during semester time for study-oriented activities (numeric: from 0 to 60)
job_break	Number of hours spent in a week during semester break for employment (numeric: from 0 to 60)
study_break	Number of hours spent in a week during semester break for study-oriented activities (numeric: from 0 to 60)
costs_direct	How difficult is it to pay for direct costs of higher education? (numeric: from 1 = very difficult to 5= very easy)
costs_opportunity	Limitation of the possibilities to earn own money and supporting yourself up until graduation (numeric: from 1 = not at all to 1 = a lot)
financialaid _bafoeg	Currently receive student financial aid (bafoeg)? (binary: 0 = No, 1 = Yes)
funding	Amount of money at your disposal on average each month in Euros (numeric: from 0 to 10900)
grades_earned	What average grade for your academic achievements in your current degree program so far? (numeric: from 1 to 5)
credits_earned	How many ECTS credits have you earned in your current degree program? (numeric: from 0 to 160)
change_field	Have you ever changed the study field at least once in the past? (binary: 0 = No, 1 = Yes)
change_uni	Have you ever changed the university type at least once in the past? (binary: 0 = No, 1 = Yes)
change_degree	Have you ever changed the type of your degree at least once in the past? (binary: 0 = No, 1 = Yes)

6 Motives for dropping out from higher education - an analysis for Bachelor students in Germany

Motives for dropping out from higher education - an analysis for Bachelor students in Germany

Andreas Behr, Marco Giese, Herve D. Teguim K., Katja Theune
Chair of Statistics
University of Duisburg-Essen, 45117 Essen, Germany

Abstract

The increasing number of students enrolled in higher education institutions and the growing demand in the labour market for university graduates make the analysis of study success and study dropout become more and more important. Dropping out from higher education is a complex process and students have very diverse motives for leaving university without obtaining a degree. We provide a detailed analysis of the different dropout reasons and aim at identifying distinctive types of dropout students using cluster analysis. The most important reasons for leaving university without a degree are mainly related to interest and expectations concerning the study as well as performance aspects. Using hierarchical cluster analysis, we further find that the dropout decision is rather based on different reasons than just caused by a single motive. Our results provide higher education institutions insights into the process of dropping out and thereby a basis for suitable and more specific countermeasures.

Keywords: student dropout, higher education, dropout motive, cluster analysis, intervention measures

6.1 Introduction

The phenomenon of student dropout in tertiary education is a very important topic, specifically for higher education institutions. It points to an inefficient use of resources and might result in students' dissatisfaction, as they do not achieve their educational goals. Dropout rates in tertiary education are very high. According to Heublein et al. (2017), dropout rates in Germany are about 29%. For an international comparison, we also provide dropout rates reported by Schnepf (2014) according to which Germany has a dropout rate of 14.7%, France of 17.9%, Spain of 24.2%, the Netherlands of 28.3%, and Italy of 34.1%. The proportion of dropout students varies between different studies since it depends on the dropout definition, the data source and the calculation methods. We define dropouts as students who leave the higher education system without obtaining a first degree (Larsen et al., 2013c). This definition generally leads to much lower dropout rates than the definition from a micro perspective, where field and institution changes are considered as dropouts. To minimise the wasting of financial and human resources, higher education institutions are increasingly searching for promising measures and programmes to identify and help students at risk. Hence, gaining insights into students' individual reasons for leaving university without a degree is of considerable importance. We define dropout from a macro perspective as leaving the higher education system completely without a degree. University or study field changes are not considered as dropouts since they just represent a transfer to another programme or another university. This study aims to identify important motives and motive bundles for dropping out, as well as differences between specific students and student types. The sparse prior research dealing with dropout motives uses mainly only descriptive approaches or focuses only on a single university/faculty with small sample sizes and older data. As these findings may not be generalized, we use a nationwide, recent, and large German data set covering 24 different motives for dropping out to provide new insights into the dropout phenomenon. We evaluate the importance of dropout motives in detail and aim at identifying different types of dropout students by applying cluster analysis. The descriptive analysis reveals that the most important reasons for leaving university without a degree are mainly related to interest and expectations concerning the study as well as performance aspects. Using cluster analysis in a first step, we find six central groups of dropout reasons: study conditions, performance and requirements, interest and expectations, job alternative and career, personal and family aspects, as well as financial aspects. Results of a further cluster analysis based on these six variables reveal highly relevant motive areas

and correlations among them. We observe, for instance, that students lacking interest in the study field also often show poor academic performance. Both analyses, descriptive and clustering, show that the dropout decision is based on a bundle of different reasons than just caused by a single motive. Identifying inter-related dropout motives is a helpful basis for the implementation of more effective individual- or group-specific prevention measures, as students with different (bundles of) problems are rather heterogeneous in their responsiveness to implemented dropout prevention measures. For instance, wrong expectations resulting in poor performances might be avoided by providing more detailed programmes helping students with an overview of the different study fields concerning study content and structure. Moreover, students, despite being interested in their study field may require help with learning strategies, and study organisation might benefit from specific workshops on these topics. This study is structured as follows. The following section gives an overview of previous literature on student dropout motives and types of dropouts. The third section contains the description of the dataset, a brief description of the German higher education system, and a short discussion of the limitations due to the problem of panel attrition and right censoring. In the fourth section, we describe the procedure of our analysis and the main statistical methods. The results are presented in the fifth section. The last section provides a discussion of the practical relevance of our findings.

6.2 Literature review

Theoretical considerations: Several theoretical research revealed the dropout phenomenon to be a very complex process, which rarely depends only on one isolated factor, but is rather the result of a bundle of reasons. The most influential dropout models are sociologically motivated, for instance, the well-known student attrition model developed by Vincent Tinto (e.g. Tinto, 1975, 1988). In his interactionist model, social (e.g. peer groups) and academic (e.g. performance) integration are the most important determinants of dropping out. The level of academic and social integration modifies students' initial institutional commitment, goals and intentions, which affect students' decision to stay or to leave university. Events external to the university affect dropout decisions mainly indirectly, due to their impact on student goals and institutional commitments. Further sociologically motivated dropout models focus on the role of institutional habitus and cultural capital. Consistency between values, norms and practices of the university

and students affects study success positively. It is assumed that students from non-academic households have a lower amount of cultural capital and therefore greater assimilation problems (Thomas, 2002). Psychologically motivated theories emphasize the role of student's behaviour, expectations and attitudes towards study. Especially the interaction of personal characteristics and the learning behaviour is of importance for the dropout decision. Aspects such as self-efficacy, coping strategies and attribution play an important role here (Bean and Eaton, 2000, 2001). Student's behaviour is also the key concept of the student engagement approach, which assumes that the amount of time invested in learning activities, determines the risk of dropping out (Müller and Braun, 2018). Economic models of student dropout are grounded on theories of rational choice and associated with human capital theory. According to these theories, students compare expected returns to education with monetary and opportunity costs, as well as expectations about their educational success. Expected returns to education depend, for instance, on perceived career prospects (Becker and Hecken, 2007). Behr et al. (2020a) provide an extensive overview of the current literature regarding higher education dropout. In sum, we observe that there are several interactions and relations between dropout reasons.

Dropout motives: According to Tinto (1975), it is important to break down the dropout phenomenon by student motives for dropping out and, thereby, to differentiate between the degrees of voluntariness. For example, from a student perspective, a dropout caused by academic failure would be perceived as non-voluntary. Despite that, a more or less voluntary dropout may probably occur due to financial distress or other personal problems. Students may drop out entirely voluntarily because of more favourable job options outside university. These various types of dropouts are driven by different motives. Involuntary dropout, for instance, is rather a result of insufficiently academic integration, such as in the form of bad grades, whereas voluntary dropouts are mainly consequences of social isolation at university or a bad match of study content and students' preferences (Tinto, 1975). Heublein et al. (2017) consider 33 motives to leave university without a degree. They cover several aspects of the study course which were identified to be important in previous research such as study organisation, own suitability, performance, expectations as well as favourable outside options. These motives were aggregated by a factor analysis into nine motive groups: performance problems, lack of study motivation, financial hardships, wishing practical work, vocational alternative, study organisation, study conditions, personal reasons, and family reasons. Motives relating to performance problems are the most frequently stated ones, followed

by a lack of study motivation and the wish to do practical work. Less important seem to be family reasons or study conditions and organisation.

Types of dropout students: There are only very few studies trying to identify different types of dropouts based on bundles of students' motives. In a very early approach, dropouts are grouped according to their timing. Two types are distinguished, namely "early dropouts" and "late dropouts" (see e.g. Gold, 1988). Extending this timing approach, Griesbach et al. (1998) analysed 3.400 dropout students from 1993/94 in Germany and found seven dropout types. Early student dropouts without vocational reorientation are mainly characterized by wrong expectations concerning their study field. Early student dropouts with vocational reorientation have doubts concerning studying in general. Late student dropouts without vocational reorientation experience a rising gap between their study field and their intentions but defer their decision to drop out. Late student dropouts with vocational reorientation additionally fail due to their study conditions in general and question the benefit of studying at a university. The remaining three groups are dropouts due to family reasons, failing examinations, and financial reasons. The authors concluded that these different dropout types need specific and more individual prevention programmes. For instance, for the first group of early student dropouts without vocational reorientation, they recommend more programmes in school to evaluate their skills and ambitions and more information to find the right study field (e.g. student information days or "try-out courses"). According to Heublein et al. (2010, 2017), dropout motives change over time and especially between students before and after the Bologna reforms (in 1999). Therefore, previous findings are not simply applicable to the new tertiary education degrees. In a more recent study, Blüthmann et al. (2012) analyse 375 Bachelor students of the University of Berlin (Freie Universitaet Berlin) exmatriculated in 2007. Using a cluster analysis based on 34 different dropout motives, they find the following four dropout types: "wrong choice", "over-challenged", "disappointed", and "strategically changing". The first one, where most of the dropouts belong to, is characterized by a low study motivation and a vocational or subject-specific reorientation. Here, students often become less interested in their study field and their career prospects. The "over-challenged" students mainly do not feel well suited for their study and often do not pass examinations. Additionally, they often have family commitments and financial problems and leave university due to a trainee position offer. The students in the third cluster are mainly disappointed with the study organisation and often change the higher education institution. Students in the last cluster state an incompatibility of the degree course and employment and are supposed to be mainly stu-

dents waiting for their preferred university place. In addition, the authors conclude that different group-specific intervention measures should be implemented to reduce dropout rates. For instance, for both groups "wrong choice" and "over-challenged", which constitute the majority of dropout students (60%), they recommend similarly to Griesbach et al. (1998) specific information programmes concerning study field content, career perspectives or the fit of requirements and own skills.

The study of Heublein et al. (2017) focuses mainly on a factor analysis and a descriptive approach to find and analyse motive groups. Blüthmann et al. (2012) apply a more elaborated method to identify dropout types but focus only on one, although large, university, and use only a very small sample size. Findings may not be transferable to the current inter-university context. Therefore, we use a large nationwide and very recent data set to analyse different motives to leave higher education without degree in detail and, furthermore, aim at identifying distinctive types of dropout students using cluster analysis to provide new insights into the dropout phenomenon. Furthermore, we examine the variation of the motives and types of dropout according to relevant individual characteristics of each student (e.g. gender, study field, social background, parental background, etc.).

The German higher education system: Within the Bologna processes, former German degrees (Diplom, Magister, Staatsexamen) were substituted by the two-tier structure of Bachelor's and Master's degrees. The regular study duration of the Bachelor's degree is 3 to 4 years and aims to provide qualification for labour market entrance (HRK, 2019). The German higher education system is mainly based on two types of institutions: general universities, which are more research-oriented, and universities of applied sciences (or polytechnics), which are more vocationally oriented. The pre-tertiary education system is characterized by early tracking into different high school types, which determine the pathway to further qualifications: the lower pathway is represented by the "Hauptschule", the intermediate pathway by the "Realschule", and the upper pathway mainly by the "Gymnasium". According to the type of high school, students can achieve different tertiary education entrance qualifications. The A-level (typically after 12 to 13 years of schooling at a Gymnasium) enables the student to enter all tertiary education institutions and is the highest and most common entrance qualification. The restricted A-level (typically after leaving school one year earlier) allows access to universities of applied sciences only. Moreover, a lower school leaving certificate and additional apprenticeship training/schooling also entitle students to enter higher education. Most of

the German universities are public institutions, financed by the states. The supremacy of the states in the field of education leads to different regulations. After some experiments with tuition fees, entrance to higher education is presently free. Despite a "numerus clausus" for several fields of study (e.g. medicine, law, business administration), there are usually no further admission rules. The German educational system is regarded as highly socially selective as there is a strong dependency between parental educational background and children's participation at upper secondary schools and tertiary education (Heublein et al., 2017, Watermann et al., 2014). The main students' financial support is a transfer based on the Federal Education and Training Assistance Act (Bundes-Ausbildungsförderungs-Gesetz, BAföG).

6.3 Data set and sample description

6.3.1 Data set

The analysis is based on data of the starting cohort 5 (first-year students) of the National Educational Panel Study (NEPS)[1]. The NEPS is a German panel study, containing 17,910 first-year students of winter term 2010/2011 and more than 3,000 variables covering various fields of student life (Blossfeld et al., 2011).

Students, who participated in the first wave, are regularly interviewed during their study course, in a frequency of about twice a year. Here, we use eleven (11) waves, which were available prior to our analysis. In the last wave, students were interviewed in November/December 2016. The NEPS contains information on the student's study progress over the eleven waves, information according to whether the student is still studying, has successfully graduated or has completely dropped out from studies without degree, as well as dropout motives and a wide range of other determinants found to be important in theoretical and empirical research on dropout (e.g. gender, study field, educational background etc.).

[1]This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort First-Year Students, doi:10.5157/NEPS:SC5:11.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Following Larsen et al. (2013c), we define dropout from a macro perspective as leaving the higher education system without any degree. University or study field changes are not considered as dropouts since they just represent a transfer to another programme or another university. In the data set, we identify dropout students as students who state that they have completely abandoned the studies or students who have not finished the degree programme and are no more studying. One cannot be sure that students later not return to higher education since the data is right censored, but in the interviews, they state no willingness to do this, otherwise this is declared as study break. At the end of the observed time span, we count 9,814 graduates, 840 students who drop out of the higher education system and 2,332 students who are still studying and remain in the panel (right-censored). Unfortunately, 4,924 students (about 27% of the initial sample) definitely left the panel without indicating their study status. Students mainly leave the panel either by retracting their initial willingness to take part or by having not participated in three consecutive computer-assisted interviews.

6.3.2 Sample description

Table 6.1 provides information for graduates and dropouts. The proportion of students of some subject groups is very small, so only the four following major subject groups are considered: (1) Engineering, (2) Mathematics and Natural Sciences, (3) Law, Economics and Social Sciences, and (4) Linguistics and Cultural Sciences. The highest total dropout rate for students is observed in Engineering (10.9%) and the lowest dropout rate in Law, Economics and Social Sciences (6.8%). Moreover, in Linguistics/Cultural Sciences the proportion of men exmatriculated from studies is twice as high as that of women (12.7% vs. 6.2%). In Engineering, for instance, women dropped out more often than men did (12.7% vs. 10.4%).

The present study focuses on dropout students (n = 662), who indicated on a scale of 1 to 6 to what extent each of the 24 different dropout reasons (see Table 6.5) applies to them. These motives include e.g. problems in performance, financial hardships, lack of interest in subject and family issues. Male students did not respond to the pregnancy variable which was imputed by 1 (does not play a role at all). As presented in Table 6.2, the sample consists of 54% female and 46% male students. 26% of the students have exmatriculated from Mathematics and Natural Sciences, 21.8% from Engineering, 21.9% and 23.3% from Law, Economics, Social Sciences and from Linguistics, Cultural Sciences,

Table 6.1: Dropout rates according to gender and study field within the data set

study field	dropouts (840) male (387)	dropouts (840) female (453)	graduates (9,814) male (3,697)	graduates (9,814) female (6,117)	total (10,654) male (4,084)	total (10,654) female (6,570)
Mathematics, Natural Sciences (2,307)	98 (10%)	114 (8.6%)	884 (90%)	1,211 (91.4%)	982	1,325
Engineering (1,594)	129 (10.4%)	45 (12.7%)	1,111 (89.6%)	309 (87.3%)	1,240	354
Law, Economics, Social Sciences (2,777)	82 (7.7%)	108 (6.3%)	986 (92.3%)	1,601 (93.7%)	1,068	1,709
Linguistics, Cultural Sciences (2,852)	59 (12.7%)	147 (6.2%)	(405 (87.3%)	2,241 (93.8%)	464	2,388
other (1,124)	19 (5.8%)	39 (4.9%)	311 (94.2%)	755 (95.1%)	330	794

respectively. In addition, only a small number of students leave higher education without degree after less than a year spent at university (less than 10% in almost all the study fields), while more than one third abandon the degree programme after 3 years of study or more.

6.3.3 Panel attrition

A limitation of the study is caused by panel attrition. As discussed in Behr et al. (2020c), dropouts have a higher probability of leaving the panel than graduates. However, this has no negative effect on the results under the assumptions they made. Under the same assumptions, after separating the 662 dropout students into two groups: panel respondents (n0 = 485) and final panel leavers (n1 = 177), no negative effect on the result has been observed. The analysis of dropout motives also reveals no significant difference between these two groups. Moreover, concerning the one-fifth of the dropout students (178 students more precisely) who missed stating the dropout motives, assumptions can be made. Although we know the distribution of these non-respondents (nearly half are women, 20% dropped out after less than a year, 20% after two years, and 40% after 3 years, equal distribution over the 4 subject fields) which is quite similar to the distribution of the dropout students with available motives, it stands to reason that their dropout motives are missing not at random (MNAR). It is difficult to assess which group of dropout students according to dropout motives might be especially over- or underrepresented since there exist no statistical tests to prove whether the data is MNAR (Kleinke et al., 2020). Some results of the descriptive analysis suggest plausibly

Table 6.2: Composition of the analysed data set containing only dropouts

study field	study years before dropping out	gender male	female	total
Mathematics, Natural Sciences (26.0%)	less than a year	1.7%	1.2%	2.9%
	1 or 2 years	9.4%	9.9%	19.3%
	3 years or more	19.7%	22.1%	41.8%
	NA	14.5%	21.5%	36.0%
	Total:	45.3%	54.7%	100%
Engineering (21.8%)	less than a year	6.9%	3.5%	10.4%
	1 or 2 years	13.2%	4.9%	18.1%
	3 years or more	25.0%	9.1%	34.1%
	NA	27.0%	10.4%	37.4%
	Total:	72.1%	27.9%	100%
Law, Economics, Social Sciences (21.9%)	less than a year	2.1%	7.6%	9.7%
	1 or 2 years	13.1%	15.9%	29.0%
	3 years or more	18.6%	20.6%	39.2%
	NA	9.0%	13.1%	22.1%
	Total:	42.8%	57.2%	100%
Linguistics, Cultural Sciences (23.3%)	less than a year	0.0%	5.2%	5.2%
	1 or 2 years	6.4%	14.9%	21.3%
	3 years or more	10.4%	24.7%	35.1%
	NA	11.7%	26.6%	38.3%
	Total:	28.6%	71.4%	100%
other (7.0%)	less than a year	2.1%	8.5%	10.6%
	1 or 2 years	8.5%	19.1%	27.6%
	3 years or more	17.1%	14.9%	32.0%
	NA	6.4%	23.4%	29.8%
	Total:	34.1%	65.9%	100%

that early and performance-related dropouts might be slightly underrepresented in our data since this is generally regarded as personal failure and some students do not want to blame themselves for dropping out. However, we expect that this problem has no severe consequences on our results.

6.4 Methodological approach

The present article aims at identifying the importance of different motives for dropping out from higher education, bundles of dropout motives and specific student dropout types. Beside some descriptive analysis of the dropout motives, we apply methods of clustering in high-dimensional subspace and clustering in low-dimensional subspace. The following scheme describes the process of analysis:

1. Clustering of the 24 individual dropout motives to motive groups according to their similarity, which is done for a better presentation and interpretation of important motive groups. Moreover, it serves as basis for low-dimensional clustering of students into dropout types.

2. Descriptive presentation and evaluation of the 24 individual dropout motives (to get a detailed impression without a loss of information, but with regard to their related motive group from step 1).

3. Investigation of the association between dropout motives and student characteristics to identify differences between specific students. Statistical tests are conducted to evaluate whether these differences are statistically significant. The choice of characteristics is based on prior theoretical and empirical research on dropout motives.

4. Clustering of students on the basis of their dropout motives to find groups of students with similar single or combined dropout motives (dropout types). We use on the one hand all the 24 individual motives (high dimension) and on the other hand the motive groups found in step 1 (low dimension). For the clustering in low dimension, a principal component analysis is used to condense the information of the individual motives in the different motive groups to a single latent variable for each group without loosing too much information.

6.4.1 Hierarchical clustering

Cluster analysis aims to group the observations $\mathbf{x}_i, i = 1, \ldots, n$ into K distinct and non-overlapping clusters. The dissimilarity between observations in the same cluster should be small, while the dissimilarity between observations in different clusters should be large (Hastie et al., 2009, Bishop, 2006).

We apply a hierarchical clustering approach as it offers an easy interpretable visualization method of the clustering process and yields slightly better cluster performances in terms of the silhouette coefficient. We use agglomerative clustering which starts at the bottom, where every observation $\mathbf{x}_i, i = 1, \ldots, n$ represents its own cluster. In every higher level, the two most similar clusters are merged until there remains only one cluster containing all observations (bottom-up strategy) (Aggarwal, 2015, Hastie et al., 2009). In the first level, agglomerative hierarchical clustering computes $n(n-1)/2$ pairwise

dissimilarities and merges the two clusters (in the first step the two single observations) that are least dissimilar. After merging the two most similar observations, the dissimilarities of the remaining $(n-1)$ clusters are calculated and the two most similar clusters are merged. This process is repeated until all observations are in one single cluster.

For calculating the dissimilarity between two clusters we use Ward's method that aims to minimize the overall cluster homogeneity. As a measure for homogeneity in a cluster, Ward's method applies the squared Euclidean distance of the observations in a cluster to its cluster center (Handl and Kuhlenkasper, 2017). Let $\overline{\mathbf{x}}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i$, $k = 1, \ldots, K$ be the cluster center of the k-th cluster C_k, where $|C_k|$ is the number of observations in this cluster. A homogeneity measure in this cluster is given by

$$H_k = \sum_{i \in C_k} (\mathbf{x}_i - \overline{\mathbf{x}}_k)'(\mathbf{x}_i - \overline{\mathbf{x}}_k). \tag{6.1}$$

The overall homogeneity $H = \sum_{k=1}^{K} H_k$ is minimized by the Ward criterion when merging two clusters.

Since the dissimilarity is monotone increasing the more clusters are merged, the results of hierarchical clustering can be plotted in a binary tree, called dendrogram. The height of the nodes in the dendrogram shows the dissimilarity between the clusters merged at the specific level (Hastie et al., 2009).

6.4.2 Evaluation of cluster models

We use the silhouette coefficient to measure the relationship of a data points' similarity to observations of its own cluster and observations of other clusters. The silhouette values are independent of the number of clusters and therefore an appropriate tool to determine the number of clusters (Dinov, 2018).

The silhouette, introduced by Rousseeuw (1987), indicates whether the object is closer to observations to its own cluster or the observations of the neighbor cluster (Handl and Kuhlenkasper, 2017). Let

$$a(i) = \frac{1}{n_k - 1} \sum_{j \in C_k} d_{ij} \tag{6.2}$$

be the average dissimilarity between observation $\mathbf{x}_i$ to observations of its own cluster, where d_{ij} is the dissimilarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ and n_k is the number of observations in the k-th cluster. Furthermore,

$$b(i) = \min_{j \neq k} \frac{1}{n_j} \sum_{l \in C_j} d_{ij} \tag{6.3}$$

denotes the average distance of observation $\mathbf{x}_i$ to the observations from its closest neighbor cluster. Then the silhouette of observation $\mathbf{x}_i$ can be written as

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}. \tag{6.4}$$

A value near 1 indicates that the observation $\mathbf{x}_i$ is much closer to the observations of its own cluster, while a value near -1 denotes that the observations of the neighbor cluster are closer to $\mathbf{x}_i$. The silhouette coefficient $\overline{s}$ is calculated by averaging all single silhouettes.

6.4.3 Principal component analysis

The problem of clustering in high dimension is also known as "curse of dimensionality" (Bellman, 1961) and has, according to Kriegel et al. (2009), four main impacts:
(1) The clustering results are hard to interpret in high dimensions.
(2) Clustering algorithms usually depend on distance measures and in high dimensions the relative distance of the nearest and the farthest observations converges to zero.
(3) In high dimensions, often irrelevant or noisy features exist.
(4) There usually exist many correlated features, so clusters might also exist in subspaces.
Especially because of point (4), methods of dimension reduction seem to be useful since correlated features are combined to a single variable.

The principal component analysis (PCA) reduces the dimension of the data by using meaningful linear combinations of correlated features (the so called principal components) explaining as much variance as possible. The mathematical details of the PCA are explained, for example, in Aggarwal (2015). The kernel PCA extends the concepts of PCA for nonlinear problems by using a nonlinear kernel function. We choose the radial basis kernel $K(\mathbf{x}, \mathbf{x'}) = exp\left(-\frac{\|\mathbf{x} - \mathbf{x'}\|^2}{2\sigma^2}\right)$ for two samples $\mathbf{x}$ and $\mathbf{x'}$, where $\|\cdot\|$ is the Euclidean norm and $\sigma > 0$ a parameter, that can be selected by the user. To maximize

the amount of variance explained by the first principal component, we plot it against different values of σ. The curve is monotonically increasing but flattens near $\sigma = 2$, so we choose this value for the parameter σ.

6.5 Empirical analysis of dropout motives

6.5.1 Clustering dropout motives

In a first step, we aim at finding groups of similar reasons by conducting a cluster analysis on the 24 dropout motives. They include, among others, problems in performance, high study requirements, financial hardships, study conditions, overcrowded lectures, lack of interest in subject, personal and family issues, job alternative, etc. The complete individual motives are presented in Table 6.5 in the appendix.

Figure 6.1 shows a dendrogram, where the horizontal line denotes the cutpoint for the number of clusters. Similar to Heublein et al. (2017), we find six motive groups. These six clusters are also reasonable concerning our theoretical considerations. The six main motive groups are: **(1)** interest/expectations (interest_field, expect, no_practice, wish_practice), **(2)** performance/requirements (exam, perform, require, material), **(3)** financial aspects (activ, financial, moneymaking), **(4)** study conditions (orga, tuition, anonym, overcrowd), **(5)** job alternative/career (job, suitability, interest_job, opportunities) and **(6)** personal/family aspects (family, ill, child, pregnant, abroad).

6.5.2 Level of importance of the dropout motives

To get a detailed impression without a loss of information, we analyse descriptively the 24 individual dropout motives taking into account their location in the motive groups obtained above. The respondents were asked to indicate on a scale from 1 (plays no role at all) to 6 (plays a major role) the importance of each of these motives. For a better presentation here (and only for the descriptive analysis), we categorize the six possible answers into three importance scales: crucial motives (played a major role, values 5-6), medium motives (played a medium role, values 3-4) and minor motives (played a minor role, values 1-2). The ranking of importance of the different motive is illustrated in Figure 6.2.

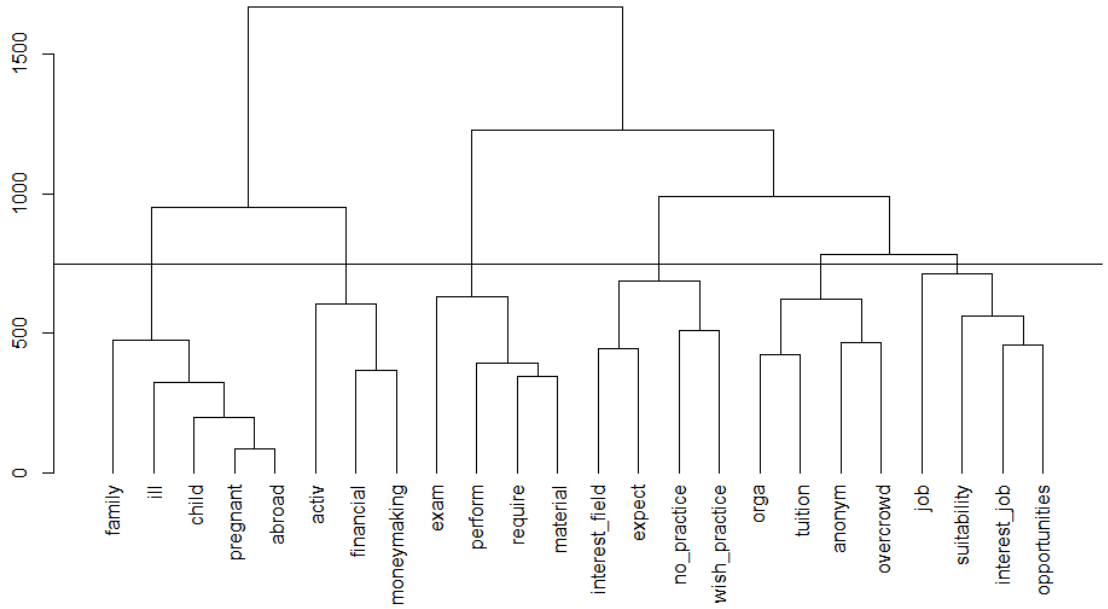

Figure 6.1: Dendrogram of variables

Figure 6.2 shows that the most important motives for dropping out are related to interest and expectation problems (1) as well as to performance and requirement problems (2). Family or personal reasons (6) are rarely decisive for dropping out. These findings are mainly in line with Heublein et al. (2017).

Bundles of students' major dropout motives

As noted, for instance, by Tinto (1975, 1988) and more recently by Heublein et al. (2017), several dropout reasons accumulate and affect each other. For a majority of dropout students, several aspects play a role in withdrawing from university. Despite the multi-causal conditionality of the dropout process, individual crucial motives that drive a student to leave the higher education institution can be identified (Blüthmann et al., 2012).

According to Figure 6.4, for only about 15% of dropout students, only one crucial motive leads them to abandon their studies. Among these students, about one-fourth discontinues studying because of failed examinations. Some of them terminated their studies because they were either offered a lucrative job (13.3%) or because of health problems

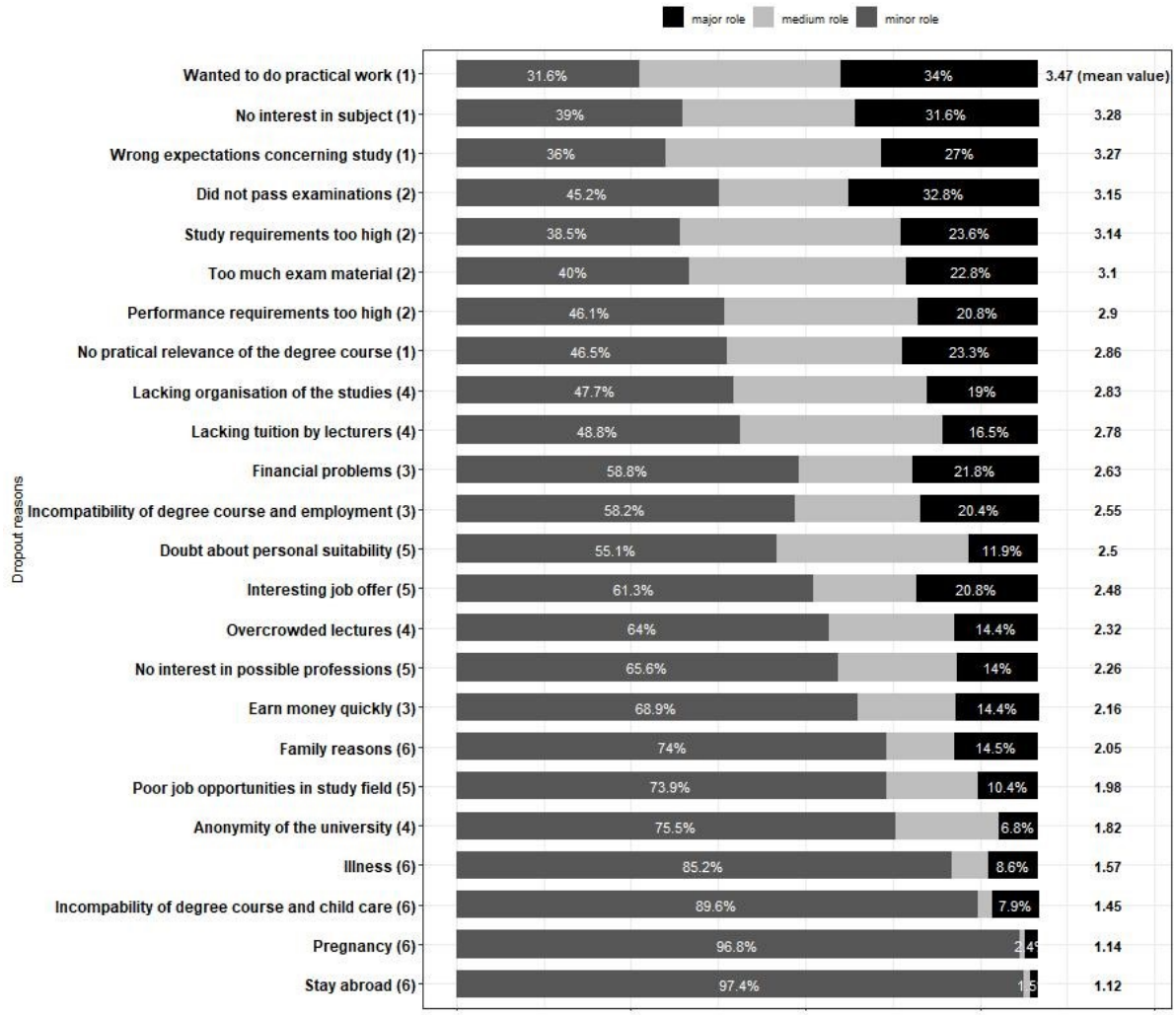

Figure 6.2: Dropout motives according to their importance

(11.2%) and some due to an incompatibility of their degree course and employment or family reasons (8.2%).

Moreover, about 15% of the students indicate two crucial motives for dropping out of university. Interestingly, the most frequently stated two important motives mainly belong to the same area such as performance/ requirements **(2)** (11.5%), interest/expectation related problems **(1)** (8.3%), financial **(3)** (5.2%) or family aspects **(6)** (3.2%).

For the majority of students, three or more main motives caused them to drop out. Some of them are again in the same thematic field, including performance/requirements (10.9%) and financial aspects (5.1%). On the other hand, we also observe many bundles of motives from different areas, such as from the two areas family aspects (family

reasons, child care) and incompatibility of degree course and employment (3.4%) or from the three areas job alternative/career, interest/expectations and financial aspects (1.7%). Students who stated the bundle of family reasons, child care and incompatibility of degree course and employment may be forced to work many hours beside studying because of having to finance family and children. Furthermore, the combination of having an interesting job offer, the wish to do practical work and incompatibility to manage degree courses and employment may point to students already being involved in business and employment. The results indicate, that there are mainly more than one reason and very individual bundles of motives to leave university without degree.

6.5.3 Major dropout motives by student characteristics

Now we investigate the association between dropout motives and student characteristics. The choice of characteristics is based on prior theoretical and empirical research on dropout motives (e.g. Heublein et al., 2017). They include gender, study years, study field, type of university, parental background, immigration as well as educational background. According to the individual characteristics of each student, the dropout motivations may vary to a certain degree. Figures 6.5 and 6.6 provide the proportion of students (in %) in each group for which the specific dropout motives played a major role. All of the 24 motives are examined (in relation to the six motive areas). Moreover, "Pearson's chi-squared tests" (Plackett, 1983) are conducted to infer whether the observed differences are statistically significant, e.g. whether the occurrence of the outcomes of the characteristic "gender" and the indication that a dropout motive has played a major role is significantly dependent.[2]

Dropout motives according to the gender:
Male and female students differ only slightly in their motives for dropping out. Noticeable distinctions are that female students stated more frequently a lack of study organization (21.8% vs. 15.8% for male) and too much exam material (25.1% vs. 20.6% for male) as reasons for their decision. Chi-squared tests show that the gender and the lacking organization of the studies as major dropout motives are at the 10%-level statistically significantly associated (p-value = 0.06). However, the test conduc-

[2]Pearson's chi-squared test is used to evaluate whether there is a statistically significant difference between the expected frequencies and the observed frequencies in a contingency table.

ted for the dropout motive "too much exam material" reveals no statistically significant association between the major importance of this motive and the gender (p-value = 0.145).

Dropout motives according to the study years:
An analysis of the decisive reasons, differentiated according to the duration of the studies until dropping out, provides interesting results. For instance, in the motive group "Performance/Requirements", failure in examinations hardly applies to first-year dropouts (19.1%) but plays an important role for later dropouts (36.1% of 2nd/3dr year dropouts and 34.8% of dropouts after 3 years). The statistical test reveals a significant association between the study years until dropping out and failure in examinations as major dropout motive (p-value= 0.04). Regarding the motive group "Personal/Family aspects", first-year dropouts (21%) seem to indicate family reasons as major dropout motive more often than later dropouts (13% of both groups respectively). However, no statistically significant dependence between the study years and family aspects as major dropout motive is observed (p-value= 0.19). Financial problems are relevant dropout motives at the beginning of the study (about 28% of 1st year and 2nd/3rd year dropout, respectively), and do not represent an issue for later dropouts (19%). At 10%-level there exists a significant association between financial problems as major dropout motive and the study year until dropping out (p-value= 0.08).

Dropout motives according to the study field:
Many students in the Mathematics/Natural Sciences and Engineering rate financial problems as the most crucial dropout motive (about 25%), whereas this seems to be less important for the two other fields (about 17%). More Linguistics and Cultural Sciences students (15%) than students of the other subjects (less than 10% respectively) indicate that the poor opportunities in their study field represent a major dropout motive. However, none of these differences are found to be statistically significant.

Dropout motives according to the type of university:
Not surprisingly, many more students from general universities (36.3%) leave because they wanted to do practical work compared to students from universities of applied sciences (29%). The statistical test reveals at the 10%-level a significant association between the university type and the wish to do practical work as major dropout motive (p-value= 0.08). Another interesting point is that students from general universities more frequently mentioned a lack of study organization as a crucial dropout motive than

their counterparts from universities of applied sciences (20.7% vs. 15.7%). However, no statistically significant association is observed here. Furthermore, the latter group is more confronted with financial problems (26.2% vs. 19.7%) and a significant association at the 10%-level is noted (p-value= 0.07).

Dropout motives according to the parental background:
Dropout students, whose mother and/or father have an above intermediate education, state to abandon their studies due to the wish to do practical work slightly more often compared to their fellow students from less educated households. Students from less educated households more often state financial problems to be responsible for leaving university without degree. Differences in the dropout reasons according to the level of education of parents are, however, not statistically significant.

Dropout motives according to the immigration background:
Here, more students with immigration background indicate that financial problems played a major role in the decision to drop out than students with no immigration background (27% vs. 20%, p-value= 0.08).

Dropout motives according to the educational background:
Regarding the secondary education background, only minor and insignificant differences in the dropout reasons "failure in examinations" and "interesting job offer" are observed.

These results provide important insights into the varying nature of the dropout motives according to students' characteristics. For example, women abandon their studies more often due to study conditions than men and for students from universities of applied sciences financial aspects are more often relevant than for students from general universities. Based on that, customized prevention programs adapted for the specific student groups could be implemented.

6.5.4 Clustering students based on all dropout motives

Considering that the dropout decision is in some cases caused by the accumulation of different motives as observed above, we aim in the next section at finding inter-related dropout motives coming from different areas. This will be examined using a clustering approach based on the 24 different motives. Furthermore, characteristics of the students falling into each group will be investigated.

We use hierarchical clustering, as described above, to find groups of students with similar dropout reasons. To avoid an enhanced impact of features with high variance, the variables are scaled to mean zero and variance one (James et al., 2013). In most clustering algorithms, it is up to the user to select a suitable number of clusters. According to Dinov (2018), we plot the mean silhouette coefficient against the number of clusters (in the range of $k \in \{2, \ldots, 8\}$) to find the optimal number of clusters for our analysis. For a larger number of clusters, the results would be hard to interpret and some resulting clusters may be very small.

Table 6.3: Average silhouette coefficient for hierarchical clustering

k	2	3	4	5	6	7	8
$\overline{s}$	0.123	0.132	0.108	0.107	0.034	0.031	0.036

We obtain the best silhouette coefficient for $k = 3$ clusters. The silhouette coefficient is generally decreasing with the number of clusters, where there is a large gap from $k = 5$ to $k = 6$.

A feature selection, as suggested by Aggarwal (2015), improves the clustering results, but a new silhouette coefficient of $\overline{s} \approx 0.189$ still points to a weak cluster structure in the data. Therefore, we regard the data in its present structure as being not suitable for clustering since the number of dimensions is relatively large. Due to the high correlation between some features applying a reduction technique prior to clustering seems appropriate.

6.5.5 Clustering in reduced dimension

Since clustering is prone to the curse of dimension, we conduct a further cluster analysis based on the six main components of dropout motives as explained above. This makes the results also easier for interpretation. As explained in the methods section, to reduce the dimension of the data, a separate kernel PCA is applied to all six motive groups that were found in Figure 6.1. The three to five variables from each category are reduced to one new latent variable which is the first principal component of each kernel PCA. We finally get six variables of interest explaining as much variability as possible in each category.

In the six categories, the following amount of variance is explained by the first principal component: study conditions (43.05%), performance/ requirements (45.88%), interest/expectations (44.66%), job alternative/career (39.60%), personal/family aspects (41.40%) and financial aspects (50.55%). Note that the amount of variance explained by the first principal component depends strongly on the number of original features in each category. That is why in the category financial aspects there is a high amount of variance explained (there are only three original features).

To make the results in the different clusters comparable, we use the scaled values over all clusters (mean zero and variance one in the complete sample) of the first principal component and not the original principal component values which are not easy to interpret. In (kernel) PCA, high values of the latent variables (the principal components) are not always accompanied by large values of the original variables. This applies to the three variables personal/family, conditions and performance/requirements. The other three latent variables have high values when the original variables have low values. In these situations, the sign of the principal component values is changed to ensure the interpretability of the six latent variables.

Estimate of the silhouette coefficient for the range of values $k \in \{2, \ldots, 8\}$ reaches its minimum value for $k = 5$ clusters with $\overline{s}_{k=5} = 0.479$. This is already a large improvement compared to the model with 24 single variables, and increases rapidly with its maximum value at $k = 8$ with $\overline{s}_{k=8} = 0.634$. As suggested by Dinov (2018), we further use the elbow criterion, where the within-cluster sum of squares is plotted against different numbers of clusters. According to Figure 6.3, $k = 8$ clusters seem to be the appropriate number of clusters within our data.

Table 6.4 displays the means and standard deviations of the scaled principal component values and the number of students n in the eight clusters. High positive values indicate great importance of a variable in a specific cluster, relative to the other clusters (mean zero in the complete sample). Negative values suggest that these dropout reasons play only a minor role.

The last row of the means gives the sums of the mean values of the six dropout reasons in each cluster. As we present standardized values, the weighted mean of the last row with cluster sizes as weights is zero. This row indicates whether students in a specific cluster have multiple reasons for dropping out, like e.g. students in cluster 2. Since also the standard deviation is standardized to 1, values below 1 denote a smaller dispersion

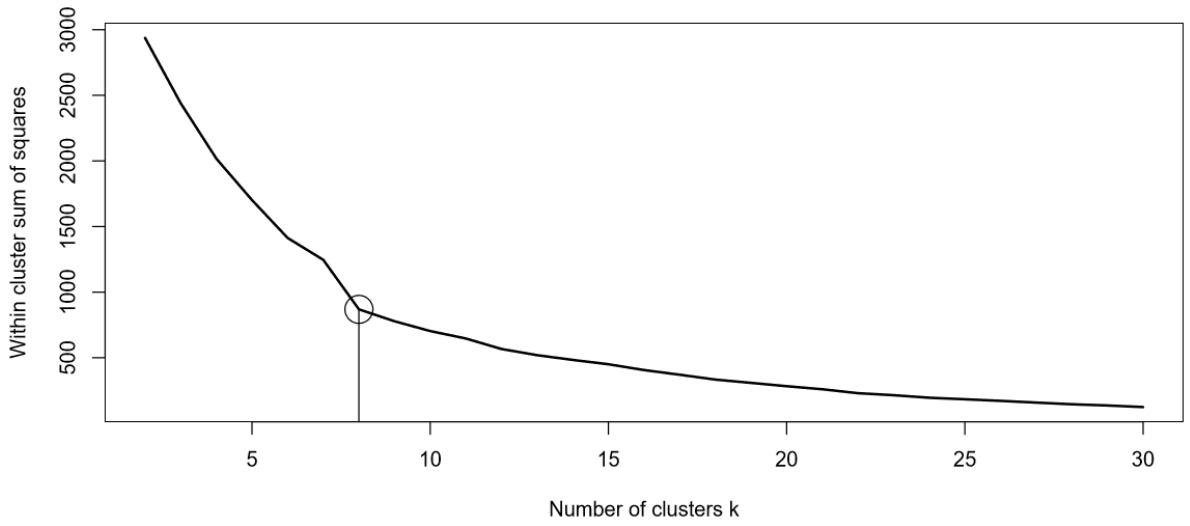

Figure 6.3: Number of clusters k using the elbow criterion.

in the specific cluster than in the complete sample. Especially in the two large clusters, students state very homogeneous values for all variables (low standard deviations), in the six other clusters of smaller size we observe single variables with standard deviations larger than 1.

Cluster 8: "Personal/Family aspects" (5%)
Students in cluster 8 generally state smaller values for dropping out and have mainly personal reasons (e.g. illness, stay abroad) or family reasons (e.g. child care) for leaving the university. Early dropouts are over-represented in this cluster (14.7% VS 7.1% in the whole sample). This is in line with observations made above. In addition, 67.65% of the students are females, while the whole sample contains only 54% female students; 32.3% (VS 21.9%) are from the study field Law, Economics, Social Sciences.

Cluster 7: "Interest/Expectations" (8%)
The only very important dropout motives in this cluster are interest and expectations related problems. Students in this cluster have no interest in their subjects and are more willing to do practical work. For this group, performance hardships seem not to play a relevant role. Over-represented are students from Linguistics, Cultural Sciences (30.2% VS 23.3% in the whole sample) and from general universities (75.5% VS 68.3% in the whole sample).

Cluster 6: "Study conditions & Personal aspects" (3.3%)

Table 6.4: Average values and number of students n in the eight clusters.

Cluster	1	2	3	4	5	6	7	8
n	204	171	41	97	40	22	53	34
Variable	Mean							
Interest/Expecta. (1)	**0.32**	**0.31**	**0.24**	**0.32**	**0.26**	-3.27	**0.26**	-3.27
Perfor./Requi. (2)	**0.37**	**0.37**	**0.36**	**0.37**	**0.36**	0.06	-2.71	-1.81
Financial aspects (3)	**0.58**	**0.59**	-0.03	-1.70	-0.27	-0.04	-0.49	-0.42
Study conditions (4)	**0.41**	**0.41**	**0.42**	**0.38**	-2.44	**0.44**	-0.76	-2.36
Job alternative (5)	**0.42**	**0.39**	-2.44	**0.41**	0.13	-1.07	-0.08	-2.02
Pers./Fam. aspects (6)	-0.86	**1.16**	-0.13	-0.29	-0.05	**0.23**	-0.21	**0.53**
Σ	1.24	3.24	-1.58	-0.52	-2.02	-3.64	-4.02	-9.35
Variable	Standard Deviation							
Interest/Expecta. (1)	0.05	0.11	0.20	0.08	0.18	0.00	0.18	0.00
Perfor./Requi. (2)	0.07	0.06	0.07	0.08	0.10	0.91	0.00	1.41
Financial aspects (3)	0.10	0.07	1.03	0.00	1.13	1.05	1.16	1.16
Study conditions (4)	0.09	0.09	0.11	0.14	0.00	0.10	1.38	0.49
Job alternative (5)	0.09	0.10	0.00	0.12	0.88	1.41	1.08	1.03
Pers./Fam. aspects (6)	0.00	0.11	0.97	0.92	1.00	1.02	0.96	0.97

In this cluster, two main motive groups lead students to abandon their studies, namely study conditions in combination with family/personal reasons. Here, one could assume, that due to some family-related obligations those students can only cope with a well organized study. Over-represented are second and third year dropouts (36.4% VS 22.2% in the whole sample), women (64% VS 54% in the whole sample) and students from Law, Economics and Social Sciences (36.4% VS 21.9% in the whole sample).

Cluster 5: "Interest/Expectations & Performance/ Requirements" (6%)
Dropout reasons for students in cluster 5 are mainly interest and expectation as well as performance related problems. In contrast to cluster 7, where only interest aspects represent the main motives, in this group, wrong expectations concerning the study (field) seem to be associated with poor performance. Students in this cluster are mainly early dropouts (15% VS 7.1%), Mathematics, Natural Sciences (35% VS 26%) and Linguistics, Cultural Studies (32.5% VS 23.3%) students, as well as students from general universities (77.5% VS 68.3% in the whole sample).

Cluster 4 & 3: "Interest/Expectations & Performance/ Requirements & Study conditions" (14.7%, 6.2%, respectively)
In these clusters, a combination of three/four motive areas seems to be responsible for the dropout decision. In addition to both motives areas identified in cluster 5 (i.e. in-

terest/expectation and performance/requirements), study conditions and job alternative also play a major role in the student dropout decision. However, job alternative only plays a minor role in cluster 3, indicating that students from cluster 3 are mainly concerned with study related problems (interest, performance and conditions) and less with external factors (financial aspect, job alternative, family aspect). Students in this group are mainly second and third year dropouts (31.7% VS 22.2%), male students (58.5% VS 46%) and specifically from the study field Engineering (31.7% VS 21.8%). No relevant characteristics are observed among students from cluster 4.

Cluster 2 & 1: "A bundle of all dropout motive areas" (25.8%, 30.8%, respectively)
In the two biggest clusters 1 and 2, students generally state higher values for a large bundle of dropout motives. In cluster 2, mainly a combination of each dropout motive, with family and financial aspects having the greatest importance, leads these students to leave higher education without a degree. It could be assumed here, that in addition to the other motives, family obligations and financial shortcomings force students to quit studying and to earn money quickly. In contrast, in cluster 1 family aspects do not play a role at all, but especially financial aspects in combination with study-related reasons seem to be very important. In both groups not only internal aspects related to study but also external aspects play a major role in the student dropout decision. Regarding the characteristics, no special feature of the students contained in both groups could be identified.

6.6 Discussion and conclusion

This analysis aims at providing a detailed analysis of students' motives do drop out of higher education. Leaving the higher education system without a degree is a long and complex decision-making process and mainly depends on a combination of several reasons. In line with previous findings, our descriptive, as well as the cluster analysis, reveal that rarely there is only one single reason or only reasons from a single area that lead students to leave university.

To effectively reduce dropout rates, a strategy might focus on programs dealing with the most relevant dropout motives. According to our analyses, the most important ones are associated with a lack of interest in the study field and wrong expectations. Here,

the main problem might be the gap between students' expectations concerning study content and organization and the real study situation which is often a result of lacking information (Suhre et al., 2007, Weerasinghe et al., 2017). Therefore, starting points for universities and also secondary schools may be to provide appropriate support for the transition from school to university and to enhance the initial study phase. One suggestion would be to offer and extend general as well as subject-specific information for students already at their qualification phase at school. Especially important seem to be information regarding different study programs, study requirements and organization, as well as job opportunities (in specific fields) and probably study alternatives (e.g. vocational training). Here, an expansion of the cooperation with secondary schools is of considerable relevance (Hetze, 2011). Student information days or more personalized workshops could help students to get an overview of the different study fields and to find study fields matching their interests concerning content and structure (see for instance Griesbach et al., 1998, Blüthmann et al., 2012, Heublein, 2014a). Universities may invite pupils to take part in some well-chosen lectures or at more intensive "try-out courses" and (mentoring)- programs to come in contact with more advanced students may help to obtain clarity on realistic study content, requirements and challenges.

The second main motive area is related to performance problems and excessive demand. Here, a problem might be the gap between students' skills and requirements for the study. Again, better support for the transition from school to university and a more enhanced initial study phase is of considerable importance. Field-specific information programs or online skills self-assessment programs may help to obtain clarity on formal and content-related requirements of the preferred subject and may encourage students to obtain these qualifications and skills already at school (e.g. to choose specific advanced subjects) and to study their subject of choice. If there are only manageable gaps between own skills and requirements of the study of choice, information on interesting alternatives or bridging courses and other preparing seminars could be provided. Another suggestion would be to reduce the number of exams in the first semester(s) to allow freshmen to become adjusted to study life and workload. Furthermore, during the first semesters of study, there often arises problems concerning study organization and workload which may lead to insufficient performance. Possible starting points to support students would be to provide seminars/work-shops on self-organization, time-management or learning techniques. For instance, Härterich et al. (2014) suggest a program "MathePlus" which should help students to improve their learning strategies and their study organization.

This strategy is suitable especially for students who have general skills and interests but need help with post-processing of the lectures and exam preparation since tertiary education requires a higher level of initiative of the students compared to pre-study education. This applies especially for students in cluster 6 (low interest in study plays no role here). The importance of the transition from school to university and the initial study phase as well as some concepts of measures are discussed within the project nexus of the German Rectors' Conference, for instance in Knoke (2018).

Although less important than the previously mentioned motive areas, some dropout reasons related to study conditions, namely lack of study organization and support from lecturers, play a relevant role for students. Here, the starting point for intervention may be to improve teaching and pedagogical skills of lectures. Furthermore, study structure and the "study ability" of degree programs should be intensively reconsidered and discussed in the future.

However, the concentration on individual important reasons for dropping out does not explain the dropout process in detail. Our results indicate, that there are many reasons for dropping out that influence each other, for instance, personal/family and financial reasons (highly correlated). Here, more information programs concerning financial aid services already at school and the very beginning of study may be a promising starting point to help, especially socially underprivileged, students to find opportunities to cover their costs. Moreover, individual counseling services helping students to structure their study and possibly reconcile their different obligations are conceivable.

Moreover, the empirical analysis reveals, that there are differences between student groups or study fields according to dropout motives. For instance, financial problems and the incompatibility of study and job are more important reasons for dropping out in Mathematics/Natural Sciences than in other study fields. This implies that field-specific prevention measures should be implemented. Furthermore, we observe motive differences between the early and late dropouts. For instance, students who drop out in a later stage of study due to an interesting job offer may have the chance, even after a few years of working, to finish their degree course later. Here, a cooperation with the employer would help to combine job and degree courses.

As already mentioned in the data description, the main limitation of the study is due to panel attrition and some missing dropout reasons. While the latter aspect might

have only minor influence on our results (as discussed above), panel attrition affects especially early dropout students. This also leads to lower dropout rates compared to Schnepf (2014). Since early dropouts are predominately performance and interest related dropouts (see cluster 5), we would have had a larger proportion of students in this cluster if the data would not have been affected by panel attrition.

In sum, as students with different (bundles of) problems are rather heterogeneous in their responsiveness to specific programs, higher education institutions wishing to implement more effective prevention measures should focus on more individual or group-specific strategies, especially related to the most important dropout motives and motive clusters.

6.7 Appendix

Table 6.5: Motives for dropping out

Attribute	Description: Reason for dropping out
Interest/Expectations (1)	
drop_interest_field	No interest in subject
drop_expect	Wrong expectations concerning study
drop_wish_practice	Wanted to do practical work
drop_no_practice	No practical relevance of the degree course
Performance/Requirements (2)	
drop_exam	Did not pass examinations
drop_require	Study requirements too high
drop_material	Too much exam material
drop_perform	Performance requirements too high
Financial aspects (3)	
drop_activ	Incompatibility of degree course and employment
drop_financial	Financial problems
drop_moneymaking	Earn money quickly
Study conditions (4)	
drop_anonym	Anonymity of the university
drop_overcrowd	Overcrowded lectures
drop_orga	Lacking organization of the studies
drop_tuition	Lacking tuition by lecturers
Job alternative/career (5)	
drop_job	Interesting job offer
drop_suitability	Doubt about personal suitability
drop_interest_job	No interest in possible professions
drop_opportunities	Poor job opportunities in study field
Personal/Family aspects (6)	
drop_child	Incompatibility of degree course and child care
drop_family	Family reasons
drop_ill	Illness
drop_pregnant	Pregnancy
drop_abroad	Stay abroad

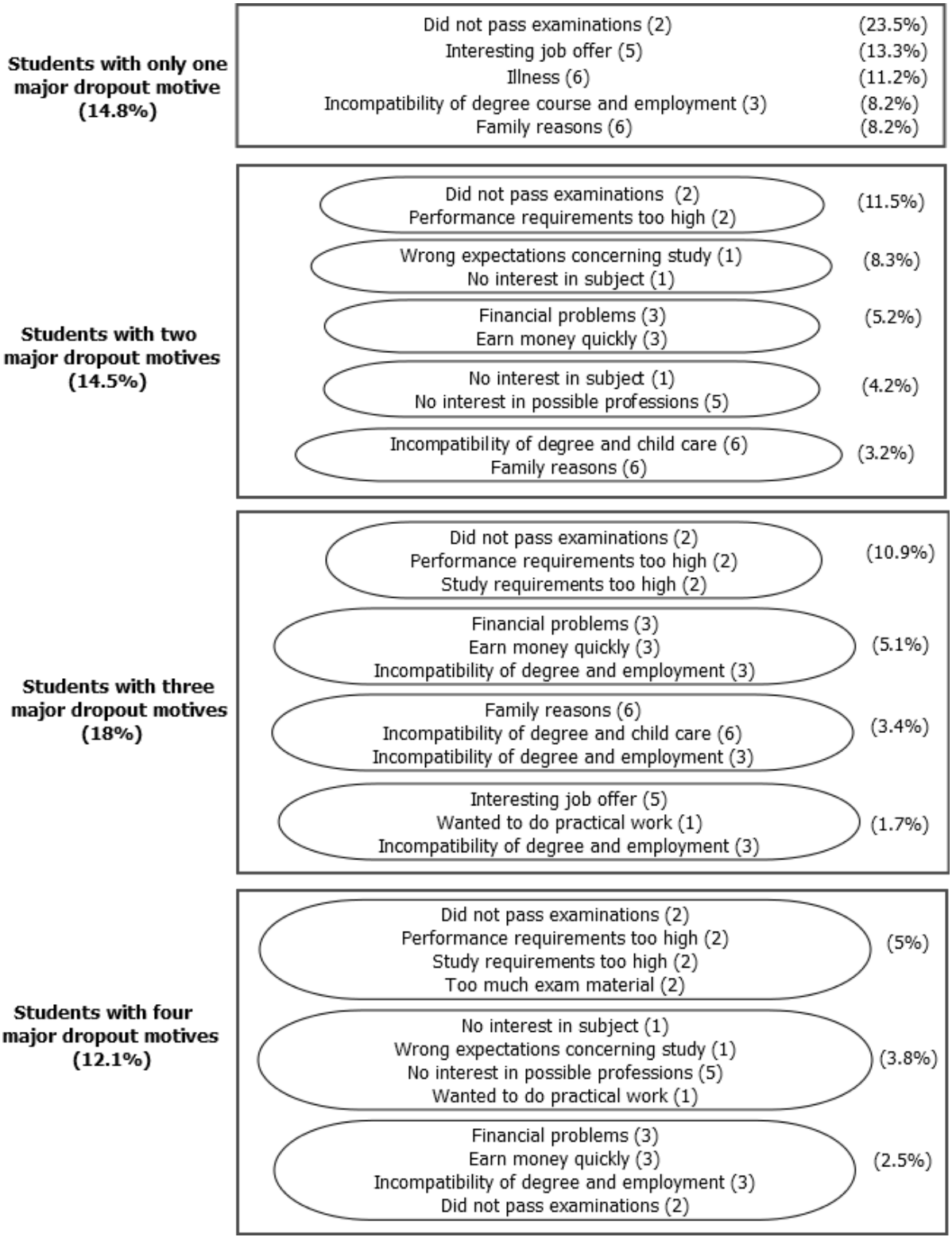

Figure 6.4: Number of major dropout motives.

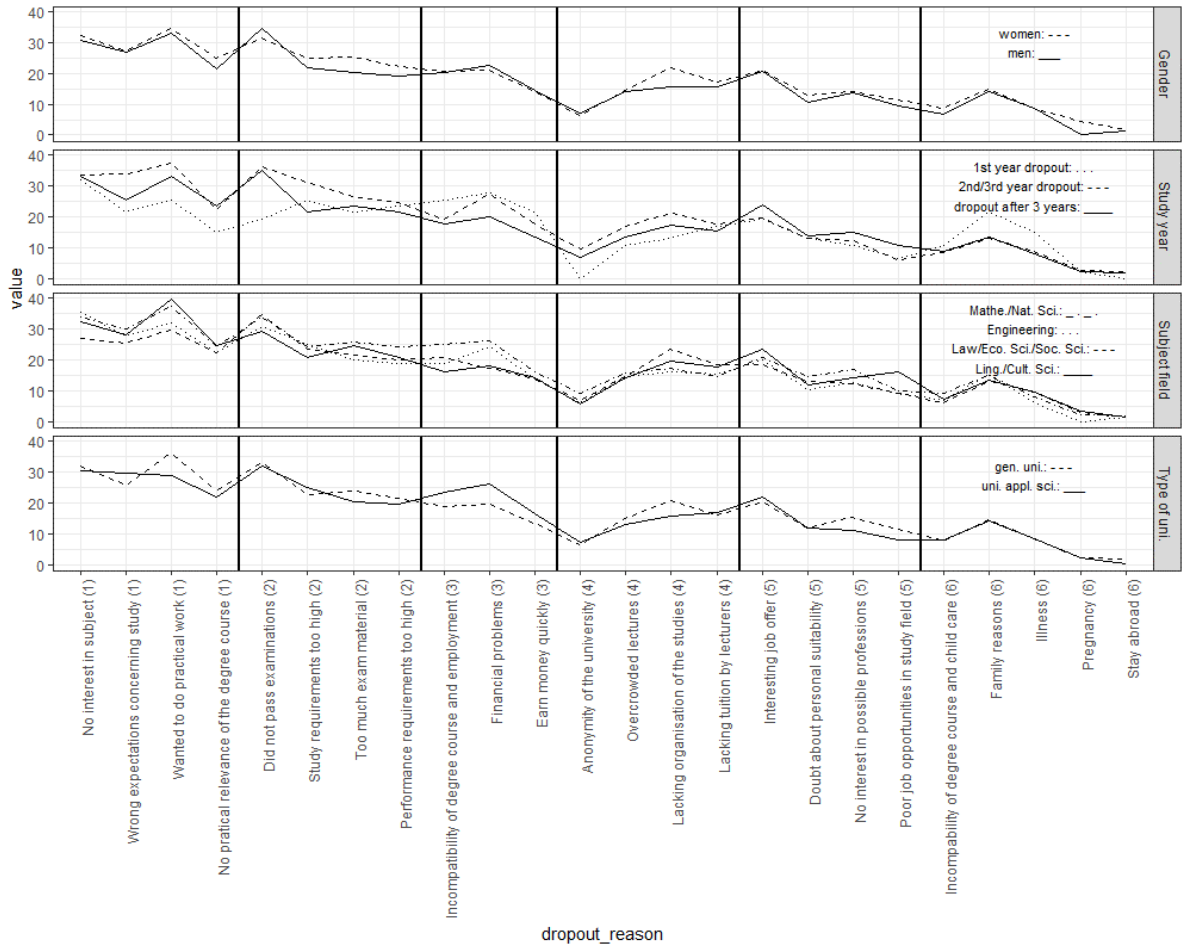

Figure 6.5: Proportion of students (in %) for which the dropout motives played an important role.

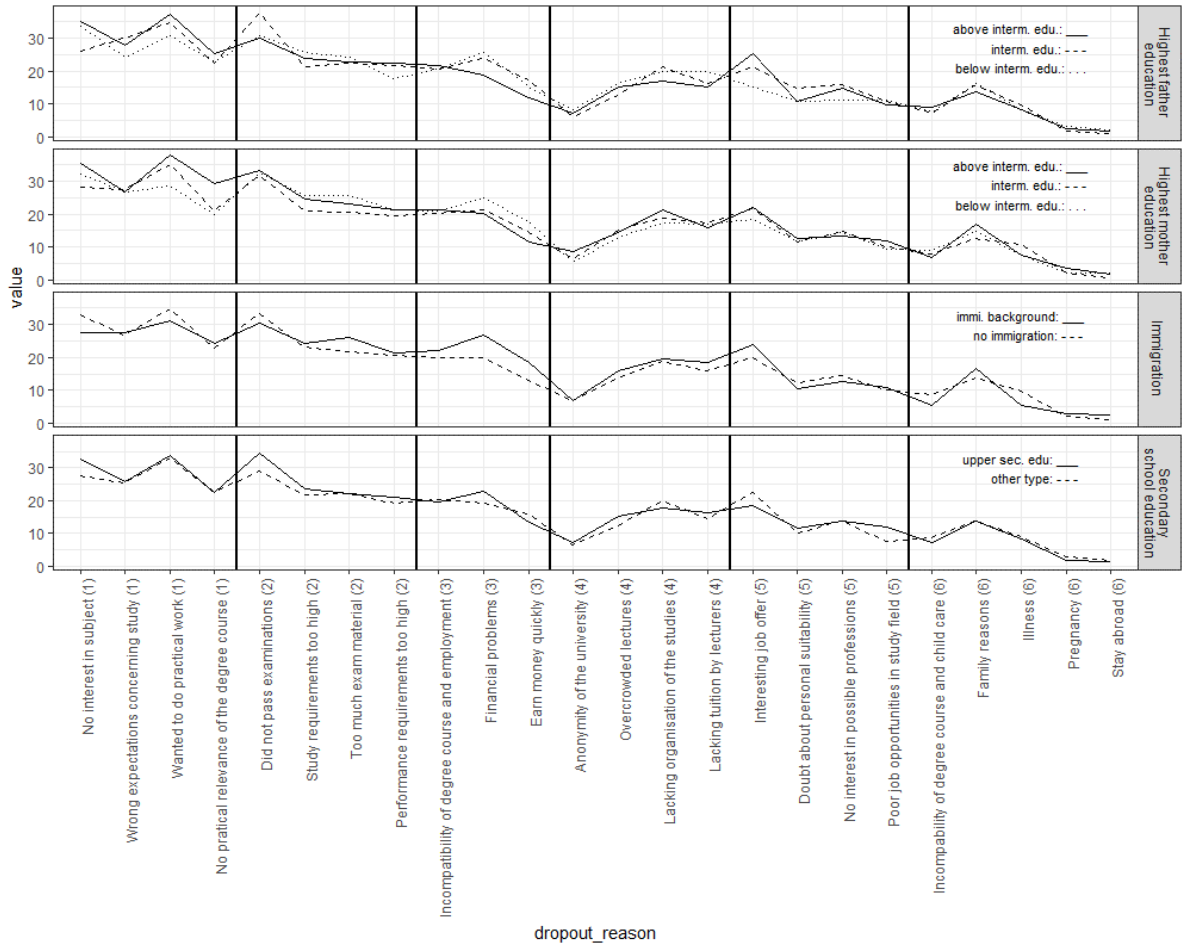

Figure 6.6: Continuation of Figure 6.5.

7 Predicting higher education grades using strategies correcting for panel attrition

Predicting higher education grades using strategies correcting for panel attrition

Marco Giese

Chair of Statistics
University of Duisburg-Essen, 45117 Essen, Germany

Abstract

This study aims to forecast the final grade of the first higher education degree which can be of considerable interest for higher education institutions to implement early warning systems, students themselves, or potential employers. The analysis is based on the National Education Panel Study (NEPS), a large German dataset covering many aspects of students' (educational) life. Since panel attrition concerns 35% of participants the Heckman correction and the inverse probability weight (IPW) estimator are used to reduce the estimation bias. A distinction is made between two scenarios, excluding dropout students and including them with a grade of 5.0. Some predictors reveal significant parameter estimates in the first but not in the second scenario, or vice versa, which means that dropout and study performance is not driven by the same variables. To get an early prediction of grades only variables of a pre-university episode were included in the first step. Afterward, variables of the early study phase are included. For the IPW estimator, the R^2 improves from 0.202 to 0.593 (dropouts included) when adding the additional variables. The best predictors are the grades at secondary school, grades in the first exams, and the type of institution.

Keywords: grade prediction, higher education, students' performance, dropout, Tweedie glm

7.1 Introduction

In the last two decades, numerous studies investigated students' dropout from higher education. But there are only a few studies that forecast the final grade of students. The prediction of grades at an early time of study, long before graduation, can be helpful to implement an early warning system for students at risk (Beck and Davidson, 2001) which assists universities to help students in a more targeted way, for example through special tutorials. Especially the public sector is interested in students with good grades (Velasco et al., 2012) so this can also help universities in recruiting the best student assistants.

The database used in this study is the fifth starting cohort of the National Education Panel Study (NEPS), which is a broad German panel dataset containing almost 18,000 freshmen students of winter term 2010/11 and covering various aspects of students' academic and personal life (Blossfeld et al., 2011)[1]. The grades of the German higher education system are in the range from 1.0 (the best possible grade) to 5.0 (failure), where 4.0 is the worst grade which is just enough to pass.

The study distinguishes between two different scenarios. In the first scenario dropout students are included in the predictions, which are students who finally leave the higher education system without a degree. In the second scenario, only students who earned a first higher education degree are of interest. Comparing the estimated regression coefficients in both scenarios the most interesting aspect is the question which coefficient estimates change dramatically. This would mean that the dropout students have a huge influence on the parameter estimate and the particular variable influence dropout decision and higher education performance in different ways.

Variables of two different points of students' academic careers were used for grade prediction. Firstly, only variables of the pre-university phase were used, which are, for example, demographic variables, migration, and information about secondary schooling and possible vocational training. The advantage of this approach is that we have

[1]This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort First-Year Students, doi:10.5157/NEPS:SC5:12.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

prediction results at a very early stage just before the start of the first semester. The disadvantage is that predictions are less accurate. This problem is mitigated in the second regression step where variables from the early study phase are also included, e.g grades in the first exams, study satisfaction, working status etc. This improves the prediction accuracy at this stage.

From a statistical perspective, this approach leads to two major challenges. The first is caused by panel attrition which is a very common problem when analyzing survey data (Behr et al., 2005). To avoid misunderstandings, students who finally leave tertiary education without a degree, are designated as (study) dropouts, and students who finally leave the panel (panel attrition) are labeled panel leavers or attriters. It stands to reason that the probability to leave the panel depends on academic performance and satisfaction and further variables covering information about the interview process. Since this would lead to biased results in the ordinary least squares (OLS) estimation the Heckman estimation and the inverse probability weight estimator (IPW) shall reduce the bias (Little and Rubin, 2019).

The second major problem occurs in the scenario where dropout students are included in the study. This leads to a mixture of a distribution which is continuous in the range of 1.0 to 4.0 (the graduates) and discrete with a value of 5.0 for the dropouts. The problem can be solved by a general linear model (glm) with the Tweedie distribution as an exponential family which is a novel approach in the education context. The Tweedie glm was, for example, utilized in modeling the zero-catch problem in the fishing industry (Shono, 2008) where the distribution is also a mixture of a continuous and a discrete (in case no fish are caught) distribution.

This study is structured as follows. The second section gives a short overview of previous literature in the field of educational data mining with a focus on the prediction of study performance of higher education students. Furthermore, some aspects of panel attrition are discussed in this section, whereby the methodological aspects are discussed in section 7.4. Some more information about the dataset and a short discussion about the missing values is given in section 7.3. The results, including a comparison of different approaches on how to deal with the panel attrition problem, are presented in section 7.5. Section 7.6 discusses the results in the higher education context and concludes.

7.2 Related work

Dropout prediction: A large number of studies in the research field of educational data mining investigate students' dropout of the tertiary education system as performance indicator using various data mining techniques. Behr et al. (2020a) give a comprehensive literature review regarding dropout of higher education. Widely used methods for this binary classification problem are, among others, artificial neuronal networks (Rios et al., 2013, Jadrić et al., 2010), decision trees and/or random forests (Superby et al., 2006, Aulck et al., 2016, Baradwaj and Pal, 2011), logistic regression (Knowles, 2015) and support vector machines (Mayra and Mauricio, 2018). Although students' dropout and students' grades are strongly correlated, poor study performance is not the only reason to leave university without a degree. Blüthmann et al. (2012) find four different clusters for the dropout students. Only in the cluster "overwhelmed" is the main reason for dropping out the poor study performance. The students in this cluster mainly suffering from a lack of interest in the study field also reveal a poor study performance. There are mainly poor study performance reasons for leaving university. Nevertheless, the grade point average (GPA) is the strongest predictor for study dropout (Stinebrickner and Stinebrickner, 2014).

Grade prediction: The number of studies predicting the final grade at tertiary education is much smaller. The problem of just analyzing students' dropout is, that no distinction is made between excellent students graduating with honors and graduates earning the degree with poor grades that are just enough to pass the specific exam. A regression analysis with the final grade as dependent variable can be seen as a generalization of the binary dropout prediction since all graduates have grades from 1 to 4 and all dropouts get the final grade of 5. Strecht et al. (2015) compares different algorithms for the two problems 1) dropout/graduate (using classification methods) and 2) final grade (using regression methods) with a focus on model performance. The performance for the regression algorithms in terms of the root mean squared error (RMSE) was approximately equal to the classification analysis. Beck and Davidson (2001) find that the academic efficacy and apathy are the most relevant determinants to predict students' final GPA. Sherman (1979) predicts the mathematics performance at high school using linear regression, where the mathematics grade in the previous years is the best predictor.

Differences to other studies: This study stands out from the few studies trying to

predict higher education grades because it follows two new approaches in this research field. On the one hand, it compares methods (Heckman correction and IPW estimation) to correct the distortion of panel attrition. Some other studies in social sciences make use of these methods, e.g. Behr (2006), but, to the best of my knowledge, not in the field of higher education grade prediction. Many studies further ignore missing values in the data that are often a result of the attrition problem. Asendorpf et al. (2014) investigate 35 articles in the International Journal of Behavioral Development in the years 2012 and 2013 and find that 20% of the studies completely ignore the problem of missing data and further 26% use inadequate methods. On the other hand, the Tweedie distribution is used to achieve better modeling of the scores. Other studies use this approach to model the monthly rainfall (Hasan and Dunn, 2011), or the zero-catch problem in the fishing industry (Shono, 2008). As far as I know, this approach is innovative in the field of higher education research. This allows for improved comparison of the two regression models that include and exclude dropouts. Determinants that are mainly significant due to the inclusion of dropouts can be detected by this approach. The central aspect of this article is still on the new finding of relevant results in the research field of higher education. But since it needs advanced statistical methods for the reasons described above (which is probably the reason why there are so few studies trying to predict university grades) the methodology cannot be neglected and should also motivate other researchers to go beyond the statistical standard methods.

7.3 Survey dataset

7.3.1 The National Education Panel Study

The National Education Panel Study (NEPS) is a comprehensive German survey data set. This study uses the starting cohort 5 covering 17,910 freshman students of the winter term 2010/2011 enrolled at German higher education institutions and more than 3,000 variables of various aspects in students' life (Blossfeld et al., 2011). In December 2019 the fifth cohort comprised twelve waves. An overview of the waves, the term of the survey, the number of participants, temporal- and final dropouts is given in Table 7.1.

The dependent variable is the final grade of the first higher education degree, which is, in general, the Bachelor's degree, where 1.0 is the best possible grade and 4.0 the

Table 7.1: Participants and panel dropouts in the current scientific use file (SUF) (LIfBi, 2017, Zinn, 2019, and own calculations). CATI: Computer assisted telephone interview, CAWI: Computer assisted web interview.

wave	instrument	partic. survey	temp. attrition	final attrition	survey term
1st	CATI	17,910	0	0	winter 10/11
2nd	CAWI	12,273	5,591	46	autumn 11
3rd	CATI	13,113	4,560	237	spring 12
4th	CAWI	11,202	6,424	284	autumn 12
5th	CATI	12,694	3,444	600	spring/summer 13
6th	CAWI	10,183	7,039	688	autumn 13
7th	CATI	9,547	7,161	1,139	summer 14
8th	CAWI	8,629	6,024	3,257	autumn 14
9th	CATI	10,096	4,321	3,493	spring/summer 15
10th	CATI	9,090	4,192	4,628	spring/summer 16
11th	CAWI	7,020	5,042	5,848	autumn 16
12th	CAWI	8,551	3,041	6,318	spring-autumn 17

worst possible grade for graduates. Since one scenario also includes dropout students with a final grade of 5.0 it is essential to define dropout. As Tinto (1975) states, the definition can have a huge impact on the study results. Spady (1970) declares there are two general different dropout definitions. The first definition regards dropout from a micro perspective, meaning from a universities or faculties viewpoint. The second definition is from a macro perspective, i.e. dropouts are defined as students who never receive a degree from any higher education institution. Here, the second definition is used since the focus is not on a single faculty or university, but on the entire German higher education system. Furthermore, the data is well suited to use this definition which is generally not possible with administrative data. This dropout definition considers students who changed the institution or the study program as graduates. In this case, the final grade of the subject where the student obtained the first degree was used. All relevant variables up to wave 12 were used to construct the status of a student (dropout, graduate, still studying, or the status is not available). The status variable is truncated on the right side after wave 12 which means that it is missing for students who are still studying after wave 12 (6 years/12 semesters after they start studying) and do not have a higher education degree (Fox, 2015). According to Heublein et al. (2008) 75% of the study programs in Germany have a standard period of six semesters and 25% seven or eight semesters. Only 3% of Bachelor dropouts leave their study program after the 10th semester (Heublein et al., 2017). The median study duration of Bachelor students in Germany was 7.6 semesters in 2018, including study interruptions (DESTATIS, 2019),

so it can be expected that the number of students who are still studying after wave 12 without obtaining their first degree is small.

The explanatory variables used in this study were selected based on a prior descriptive analysis of the NEPS data (Behr et al., 2020b). The most relevant variables with sufficient data quality (not more than 50% missing values) were used in this study. These are variables that become relevant already before study, e.g. demographic variables (e.g. migration, age, gender), secondary education (e.g. final school grade, type of school), parental background, variables describing the phase immediately before study, e.g. is the student studying his subject of choice, what do parents and friends think about the study choice or was there an alternative to study. Finally, variables that are of importance during the study program, e.g. study satisfaction, study commitment, academic integration, off-study work, or financial situation were used to predict the final degree grade. The Tables 7.8, 7.9 and 7.10 in the appendix reveal a more detailed list of variables in the two episodes pre-university and early study phase.

Another educational panel study in Germany is the "Studienberechtigtenpanel" published by the German Centre for Higher Education and Science Research (DZHW) approximately every three years. This panel has, compared to the NEPS, a much smaller number of variables, contains only two waves and reveals a much larger number of panel leavers in wave 2 (Birkelbach et al., 2019). One of the largest survey datasets for educational research covering 15-year old students in OECD countries in 2018 is the Program for International Student Assessment (PISA) (Sellar and Lingard, 2014). In contrast to the NEPS PISA is a cross-sectional dataset whereas the NEPS has a panel structure.

7.3.2 Problems due to non-response and initial selection bias

From a target population of 31,082 freshmen students 13,172 students did not respond, which leads to 17,910 participants in the first wave (Zinn et al., 2017). These students, who did not respond, were not asked for participation in further waves and no information about them is available in the data. Design weights were introduced to overcome the initial bias due to nonresponse and different selection probabilities, e.g. women, students without migration background and students born in 1990 or later are overrepresented in the initial sample (LIfBi, 2017).

The more severe problem of the data regards students who finally leave the panel before graduation or dropout. The extent of final and temporary panel leavers in each wave is displayed in Table 7.1. The contingency Table 7.2 reveals the students' status (graduation, dropout, continue studying, or the status is not available) and the panel attrition (whether a student finally left the panel up to wave 12). Even if the true frequencies for the dropouts are near the expected frequencies under the assumption of independence, this does not hold for the other three groups. The χ^2-test of independence (Hartung et al., 2009) rejects the null with a p-value near zero. The group of students who are still studying and finally left the panel is comparably large in the data. This is mainly caused by students who finally left the panel before graduation or dropout but after wave two (since these students have no available status). Furthermore, one can see that the graduation rate in the sample ($9815/10,657 = 0.921$, if just dropouts and graduates are used) is above the graduation rate of 85.3% that Schnepf (2014) found for Germany using a similar dropout definition.

Table 7.2: Contingency table with students' status and panel attrition

attrition \ status	dropout	graduate	still studying	status not available	Σ
no final attrition	545	8,832	2,215	0	11,592
final attrition	297	983	3,739	1299	6,318
Σ	842	9,815	5,954	1,299	17,910

Since only observations, where the final grade is available, are useful for later regression models the final sample contains 8,727 observations in the situation where university dropouts are included in the study. This disregards all students with unavailable status, who are still studying and all graduates who did not state their final grade. It stands to reason that the probability to leave the panel also depends on the final grade. The point biserial correlation (Bortz and Schuster, 2010) between the grade and the binary attrition variable in the sample containing dropouts and graduates is 0.280 which indicates that the probability to leave the panel rises as grades worsen. But this is mainly caused by dropout students. Excluding the dropouts leads to a point biserial correlation of -0.004 which suggests that grades and panel attrition are uncorrelated. Therefore, in the later analyses both samples are regarded, graduates and dropouts ($n = 8,727$) and only graduates ($n = 7,884$). In the latter sample, the sample size is smaller and dropouts are simply ignored but the attrition bias might be smaller.

7.4 Methodological approach

This section describes the statistical methods needed for empirical analysis. As already described in the introduction there are two major statistical challenges in this study. The first problem covers panel attrition and the resulting missing values in the data. Most, but not all missing values are caused by panel attrition. Section 7.4.1 describes the three major types of missing values and the types of missing values in the NEPS. The following subsection 7.4.2 explains imputation strategies that fill in the missing values. The subsections 7.4.3 and 7.4.4 explain the two methods that reduce (or - if all assumptions met - eliminate) the bias in the parameter estimates which is caused by attrition. The following subsection 7.4.5 introduces the Tweedie distribution which is needed to handle the second major problem of a zero-inflated continuous distribution in the scenario where dropouts are included. Lastly, measures to evaluate the performances of the different models are introduced in 7.4.6. In order not to interrupt the flow of reading by too many formulas, methodological details are included in the appendix.

To compare the different strategies the ordinary least squares estimation (OLS) is used as a benchmark (see Appendix).

7.4.1 Types of missing data

In general, a distinction is made between three types of missing data (Little and Rubin, 2019, Fox, 2015).

(1) Data is missing completely at random (MCAR) if missing data appears randomly independent of the missing variables or any other study variables. MCAR rarely occurs in real data.

(2) Data is missing at random (MAR) if the missing mechanism is not completely random and depends on the observed data. But, conditioned on the observed data, the missing mechanism is independent of the missing data. Statistical methods to prove for MAR do not exist. For example, if students were asked for their actual grades at university, there will be more missing values for freshmen students because they did not write any exams. If the missing mechanism does not depend on the grade itself, the data is MAR.

(3) If the missing mechanism depends on the missing variable itself, the data is not missing at random (NMAR). To continue the example above, if the willingness to disclose the actual university grade depends on the grade itself, e.g. students with bad grades may be less willing to disclose their grades, then the data is NMAR. In this situation, the missing mechanism is nonignorable, since ignoring it would lead to biased results.

NEPS: The NEPS distinguishes between three broad classes of missing data (LIfBi, 2017): a) Item nonresponse, e.g. refused answers or the participant does not know the answer. b) Not applicable, e.g. the variable was not included in a specific survey wave or the variable was filtered (e.g. men were not asked for pregnancy). c) Edition missings, i.e. for some (very special and for this analysis not relevant) variables a remote access is needed, otherwise, the variable is not available. Furthermore, category d) of missing values can be introduced which includes temporal and final panel leavers who are not contained in the NEPS data.

Type c) of missing values is not relevant for this analysis. The 46 cati and 35 cawi variables are deleted. The missing type b) is also of minor interest in this study, since just the survey waves where the specific variable was included were used. Type a) and especially type d) are more problematic because it stands to reason that these kind of missings are nonignorable, even if there is no statistical method to test that hypothesis without making special assumptions (Little and Rubin, 2019). Whereas missing type a) occurs rarely, i.e. in only 0.71% of all non-missing cati variables, type d) emerges frequently from wave 2 as displayed in Table 7.1. It will not be possible to completely eliminate the bias from the estimation since the assumptions of the following sections are very strict and might not be entirely fulfilled. Nevertheless, the bias should be reduced as far as it is possible.

Köhler et al. (2015) investigate the response behavior in competence tests in the NEPS starting cohorts 3, 4 and 6 and come to the result that the response probability is strongly related to the competence of a person but also other person-specific attributes are relevant. This indicates the importance of an adequate estimation of the response probabilities that are needed for the inverse probability weighting in section 7.4.4.

7.4.2 Imputation methods

Imputation methods complete the missing entries in a dataset and make the application of statistical standard methods possible (Fox, 2015). There are two general imputation strategies: 1) Single imputation where all missing values are completed once. The most simple imputation methods are mean or median imputation which usually reduce the variance of the imputed variable dramatically (Little and Rubin, 2019). 2) To mitigate the problem of single imputation, multiple imputation completes all missing entries D times, where each imputed value is drawn from the predictive distribution of the missing value given the observed values. This technique can reflect the uncertainty of the missing data on the costs of additional complexity and computation time (Fox, 2015).

This study uses predictive mean matching (PMM) as imputation technique, introduced by Rubin (1986), which has less stringent assumptions compared to some parametric imputation methods. It has the advantage that real values are sampled being from the same sample space as the original variable which makes it applicable to metric as well as ordinal or categorical variables. The basic idea of PMM is to find possible matching candidates for the missing values in the set of observed values by minimizing the distance of predictive regression values on the variable that is imputed. From these candidates, one value is randomly drawn. In contrast to other imputation techniques, the PMM uses linear regression not for directly imputing missing values, but for matching missing cases with the most similar observed cases (Van Buuren, 2018).

The step of PMM where random values are drawn makes it possible to repeat this step D times to generate different datasets which is known as multiple imputation (Van Buuren, 2018). Averaging the results leads to the combined estimate. Note that the variance of multiple imputation has a within and between component (Little and Rubin, 2019).

Whereas in the previous decade $D = 5$ imputations were the usual, Asendorpf et al. (2014) suggest using at least $D = 20$ imputations since the computation power increased rapidly.

Only the explanatory variables were imputed in this study. It follows that just the complete cases of the dependent variable are included in the OLS-model.

7.4.3 Heckman correction

To overcome the problem of self-selection, i.e. that the students with available final study grades are not representative for the whole population, Heckman (1976) suggested a two-step approach. In the first step the dichotomous response variable

$$R_i = \begin{cases} 1, \text{ if } Y_i \text{ is observed} \\ 0, \text{ else} \end{cases} \tag{7.1}$$

is defined. Via probit regression (Bishop, 2006) the probability of an observed grade given a set of variables $\mathbf{Z}$ is calculated: $P(R = 1|\mathbf{Z}) = \Phi(\mathbf{Z}\gamma)$. Here, γ is a regression parameter estimated by the probit model that is used to estimate the inverse mills ratio $\hat{\lambda}$ (see the Appendix for details) and Φ is the probability function of the Gaussian distribution. In the second step, the estimates $\hat{\lambda}$ are used as additional regression parameter to estimate the final grade (Fox, 2015). The matrix $\mathbf{Z}$ in the probit regression contains variables used to estimate the response (1 for respondents and 0 for non-respondents), γ is the parameter vector optimized by the model. The matrix $\mathbf{Z}$ can contain variables that are also in the design matrix $\mathbf{X}$ but it also contains additional variables describing the response behavior but have no influence on the target variable y, e.g. information about the interviewer and the number of contact attempts. Table 7.10 gives an overview of these interview specific variables used in the study.

7.4.4 Weighting methods

Weighting is one of the most widely used methods when panel attrition induces a bias in the estimates in the common OLS model (Vandecasteele and Debels, 2007). The two steps of the inverse probability weighted estimator, described by Robins et al. (1995), are very simple and intuitive.

In the first step one estimates the response probabilities R_i using the variables in $\mathbf{Z}$, defined above for the Heckman correction, for example via logistic regression.

These are denoted with $\hat{\pi}_i = \hat{P}(R_i = 1|\mathbf{Z}_i)$, $i = 1, \ldots, n$ and the $n \times n$ matrix $\hat{\Pi} = diag(\hat{\pi})$ contains the vector $\hat{\pi}$ on the diagonal and all other entries are zero.

In the second step a weighted OLS estimation with $\mathbf{X}$ as exploratory and Y as dependent variable is conducted, where the inverse weights from first step are used to put more

weight on participants with a large non-response probability:

$$\hat{\beta}_{IPW} = \left[\mathbf{X}'\hat{\Pi}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\hat{\Pi}^{-1}Y. \tag{7.2}$$

The inverse probability weighted (IPW) estimator results in a consistent estimation of β if the response probabilities are known (Robins et al., 1995). Therefore, it is essential to get unbiased estimates of the response probabilities in the first step.

7.4.5 Tweedie distribution

The final grade from 1 to 4 in the German higher education system is rounded down on one decimal. Indeed, the grade distribution would be continuous in the interval [1,4] if it were not rounded to one decimal place. A problem occurs if dropout students are included in the model with a 5.0 which leads to a semicontinuous distribution. Mixture models (Van Buuren, 2018) can handle such distributions where a discrete and a continuous part occurs. These are often used in zero-inflated models. Standard applications are the modeling of daily rainfall (many days without rain) (Hasan and Dunn, 2011) or for insurance companies the loss amount of individual policyholders in a certain period (Jørgensen and Paes De Souza, 1994). To transform the grade variable to a zero-inflated model the transformation

$$\tilde{y}_i = \sqrt{5} - \sqrt{y_i} \tag{7.3}$$

is used, where y_i is the original grade variable of the i-th student and $\tilde{y}_i$ the transformed grade. The square root is used because it best eliminates the skewness of the data.

The Tweedie distribution, introduced by Tweedie (1984), can overcome the problem of zero-inflation. It is a generalization of some other distributions, including the Gaussian distribution, but here the Poisson-Gamma distribution is of interest to model the zero-inflated data (Shono, 2008). If the random variable K is discrete Poisson-distributed and $Z_1, \ldots, Z_K$ are independent, identical random variables following a Gamma distribution, a Poisson-Gamma distributed variable Y can be written as

$$\mathbf{Y} = \sum_{k=1}^{K} Z_k. \tag{7.4}$$

This leads to the zero-inflation in cases where $K = 0$. This distribution is used as exponential family in a generalized linear models (glm) (Fahrmeir and Tutz, 2013) to model the final grades including dropouts.

7.4.6 Model comparison

To evaluate the model performance on new, unseen observations the dataset is divided into training data to fit the model and test data (50% of the complete dataset in each group) (Hastie et al., 2009). The training and test sets are different samples for all of the $D = 20$ imputed datasets and the results were aggregated as explained in section 7.4.2. As evaluation measures, I used the $R^2 \in [0, 1]$, which quantifies the variance explained by the model, and the mean squared error (MSE) (Aggarwal, 2015), which measures the squared error between the predicted values $\hat{y}_i$ and the observed values

$$\text{MSE} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / n. \tag{7.5}$$

A good model should have a large R^2 and a small MSE, whereby the latter strongly depends on the variance of the dependent variable.

Furthermore, the parameter estimates are compared especially to the OLS model where the parameters are expected to be biased. Other studies, such as Behr (2006), conduct a bias analysis but this is only possible under strong assumptions which do not apply to the NEPS data.

Note that (1.) model performance and (2.) parameter estimation are two completely different topics. The IPW estimator and the Heckman correction mainly correct for the bias in the parameter estimates but they may also improve the model performance. In the two-step approach of the Heckman model, we have an additional explanatory variable (the inverse Mills-ratio) that might also improve the model performance. Since underrepresented students in the training data are also underrepresented in the test data it can be expected that the model performance gap between the OLS model and the two model correcting for attrition increases in favor of the latter two models if they were applied in real situations where no group of students is over- or underrepresented.

To eliminate implausible continual values larger than 4.0, all predicted grades larger than 4.0 were set to 5.0. This also brings us closer to the true mixture distribution which is discrete for dropouts (grade 5.0) and continuous otherwise.

7.5 Empirical results

This section presents the empirical results to answer the two major research questions 1) How do parameter estimates differ when dropouts are included? 2) How far does the model improve when additional variables from a later point in time are added to the model?[2]. To answer the second major research question this section is divided into two main parts. In section 7.5.1 just pre-university variables are used where the number of missing values in the explanatory variables is small and therefore the data is less sensitive to imputation. Adding additional variables of the early study phase, what is done in section 7.5.2, improves the model performance in terms of MSE and R^2 due to the additional information in the data. However, the added variables are not only from wave 1 but mainly from waves 2 and 3 where the data contains more missing values caused by panel attrition. This makes the model more sensitive for a potential bias caused by missing data and the prediction is only possible at a later time in study.

To find answers on the first major research question, four models were compared in the scenario including dropouts and three models when dropouts are excluded. In the latter case, the Tweedie glm model is missing since the problem a zero-inflated mixture distribution only applies for the scenario with dropouts included. One further important aspect of this article is an adequate handling of panel attrition and missing values in survey data and therefore two models (IPW and Heckman) are compared. While the Tweedie glm should mainly improve the model performance, the IPW and Heckman model should reduce the bias in the parameter estimates. The IPW estimator and the Heckman correction were also embedded in the Tweedie model. The OLS model serves as a benchmark model.

[2]For all calculations the statistical software R version 3.6.1 was used (R Core Team, 2019)

7.5.1 Pre-university variables

Here only pre-university variables are used as explanatory variables which means variables up to the end of secondary education or vocational training that have nothing to do with higher education or the study decision process. An overview of the variables is given in the Tables 7.8, 7.9 and 7.10 in the appendix. The number of missing values in each variable (% NA) is calculated on basis of students ($n = 8,727$) where the degree grade is available including study dropouts, except for the degree grade itself where the percentage of missing values is calculated based on all 17,910 participants of wave 1.

Table 7.3 reveals the out-of-sample performance results of the four models in both scenarios. Note that the grades were retransformed to the original form for better interpretation, which is equally applicable to Table 7.5. The transformation in equation 7.3 was only used for better modeling properties. Since a general linear model with a Gaussian exponential family is nothing else as the usual OLS regression, the Tweedie model reveals the same results as the OLS model when dropouts are excluded.

Table 7.3: R^2 and MSE of the three different methods using only pre-university variables (standard errors over the 20 imputations in parenthesis)

Measure	Include dropouts	OLS	Tweedie	IPW	Heckman
MSE	yes	0.857	0.855	0.852	0.851
		(0.022)	(0.021)	(0.022)	(0.021)
	no	0.197	0.197	0.199	0.194
		(0.003)	(0.003)	(0.003)	(0.003)
R^2	yes	0.192	0.204	0.202	0.205
		(0.012)	(0.016)	(0.021)	(0.023)
	no	0.169	0.169	0.198	0.183
		(0.009)	(0.009)	(0.011)	(0.011)

The glm with the Tweedie distribution slightly outperforms the OLS benchmark model in terms of R^2 and MSE in the situation with dropouts. The best models regarding the model performance are the Heckman model and the IPW estimator which slightly outperform the two other models. The Heckman correction includes estimates of the inverse Mills ratio $\hat{\lambda}$ as an additional explanatory variable in the second step which is has a significant influence as demonstrated in Table 7.4. The relatively small amount of

variance explained by the models is caused by the fact that just pre-university variables were used but the dropout process and study performance are also affected by many study related variables as stated in section 7.5.2.

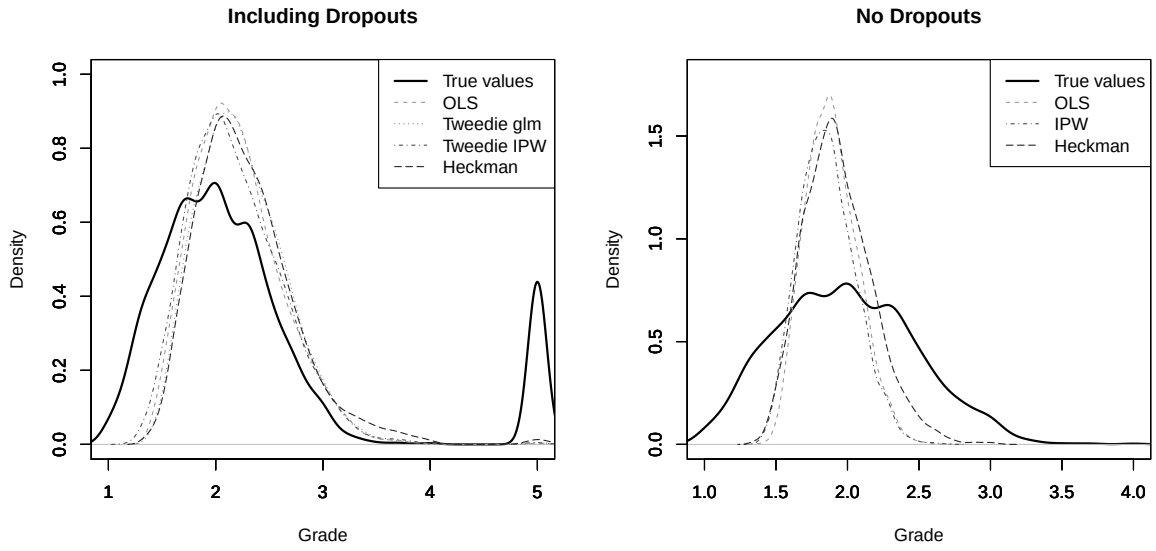

Figure 7.1: Kernel density estimation of true grades and the out-of-sample predictions of the four models including dropout students (left panel) and excluding them (right panel) using only data of the pre-university phase.

Figure 7.1 reveals the kernel density estimation of the true grades and the kernel density of the out-of-sample predictions of the four models. The distribution of the models is concentrated near the median of the true distribution. In the left panel, one can see that only the Heckman estimator accomplishes to predict a notable amount of university dropouts. It is just too early to get reasonable predictions of study performance. The variance of the predicted grades is much smaller when dropouts are excluded.

Table 7.4 shows the parameter estimates of the four models. The parameter estimates were averaged over the 20 imputations as in equation 7.8 the standard deviation was calculated following equation 7.9. All observations have been used for the regression since it is not necessary to hold out test data as in the performance analysis.

Note that in Table 7.4 as well as in Table 7.6 and 7.7 in the next section the transformed degree grade of equation 7.3 is used since back-transformation is not possible here. This

Table 7.4: Parameter estimates and standard errors for the different models using only pre-university variables. The reference category for the region of origin is Africa/Asia. Significance codes: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$.

	including dropouts								no dropouts					
	OLS		Tweedie glm		Tweedie IPW		Heckman		OLS		IPW		Heckman	
Variable	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.
(Intercept)	-11.010***	1.397	-11.706***	1.849	-9.008***	1.925	-11.895***	1.865	7.116***	1.397	5.254***	1.406	8.850***	1.393
genstat	0.009	0.005	0.002	0.008	0.015	0.008	0.003	0.008	0.000	0.005	-0.010	0.005	-0.002	0.005
rep_class	-0.059***	0.005	-0.062***	0.006	-0.050***	0.007	-0.064***	0.006	-0.021***	0.005	-0.029***	0.005	-0.014**	0.005
familylife	0.031	0.006	0.027**	0.009	0.032***	0.009	0.028**	0.009	0.004	0.006	-0.000	0.006	0.002	0.005
school_type	0.051***	0.006	0.066***	0.009	0.068***	0.009	0.067***	0.009	0.036***	0.006	0.036***	0.006	0.034***	0.005
qualif_max	0.008	0.005	-0.010	0.008	-0.011	0.008	-0.010	0.008	-0.010	0.005	-0.016**	0.005	-0.009	0.005
grade_school	-0.154***	0.005	-0.147***	0.008	-0.153***	0.008	-0.152***	0.008	-0.095***	0.005	-0.101***	0.005	-0.078***	0.005
math_points	0.001	0.001	0.002	0.001	0.002	0.001	0.002	0.001	-0.001	0.001	-0.002*	0.001	-0.001	0.001
ger_points	0.005***	0.001	0.005**	0.002	0.007***	0.002	0.005**	0.002	0.005***	0.001	0.006***	0.001	0.005***	0.001
exam_german	0.011	0.005	-0.000	0.009	0.004	0.009	0.001	0.009	0.009	0.005	0.013**	0.005	0.006	0.005
exam_adv_german	0.001	0.004	0.011	0.007	0.001	0.007	0.011	0.007	0.008	0.004	0.007	0.004	0.009*	0.004
exam_maths	0.008	0.005	0.009	0.009	0.029**	0.009	0.008	0.009	0.001	0.005	-0.004	0.005	0.003	0.005
exam_adv_maths	-0.017*	0.004	-0.013	0.007	-0.015*	0.007	-0.014	0.007	-0.012**	0.004	-0.015***	0.004	-0.010*	0.004
birthyear	0.006***	0.001	0.006***	0.001	0.005***	0.001	0.006***	0.001	-0.003***	0.001	-0.002**	0.001	-0.004***	0.001
gender	-0.031***	0.004	-0.034***	0.007	-0.032***	0.007	-0.035***	0.007	-0.024***	0.004	-0.020***	0.004	-0.019***	0.004
mother_qualif	0.002	0.001	0.003	0.002	0.003	0.002	0.003	0.002	0.001	0.001	-0.001	0.001	0.001	0.001
father_qualif	-0.001	0.001	-0.001	0.002	-0.002	0.002	-0.001	0.002	-0.001	0.001	-0.000	0.001	-0.001	0.001
mother_job	0.001*	0.000	0.001*	0.000	0.001**	0.000	0.001*	0.000	0.000	0.000	0.0003*	0.000	0.000	0.000
father_job	0.000	0.000	0.000	0.000	0.001*	0.000	0.000	0.000	0.000	0.000	0.0003*	0.000	0.000	0.000
voctrain	0.042***	0.006	0.050***	0.009	0.048***	0.009	0.051***	0.009	0.008	0.006	0.005	0.006	0.003	0.006
fail_prestudy	-0.010	0.011	-0.010	0.016	0.003	0.016	-0.011	0.016	0.012	0.011	0.007	0.010	0.015	0.011
mother_alive	0.009	0.015	0.032	0.023	0.035	0.024	0.033	0.023	-0.002	0.015	0.005	0.014	-0.008	0.015
fath_alive	-0.013	0.010	-0.016	0.017	-0.031	0.017	-0.016	0.017	-0.016	0.010	-0.011	0.010	-0.015	0.010
bilingual	-0.021	0.009	-0.036*	0.015	-0.004	0.016	-0.038*	0.015	-0.016	0.009	-0.012	0.010	-0.012	0.009
lang_ger	0.006	0.014	0.014	0.023	-0.024	0.022	0.013	0.023	0.036*	0.014	0.040**	0.015	0.041**	0.014
household	-0.002*	0.001	-0.002	0.002	0.000	0.002	-0.002	0.002	-0.002*	0.001	-0.001	0.001	-0.002*	0.001
America/Australia	-0.021	0.043	-0.025	0.067	-0.167**	0.060	-0.031	0.067	0.032	0.043	0.042	0.043	0.045	0.043
Germany	-0.031	0.029	-0.038	0.047	-0.081	0.050	-0.046	0.048	0.014	0.029	0.040	0.027	0.035	0.029
eastern Europe	0.003	0.028	-0.000	0.046	-0.024	0.049	-0.005	0.046	0.015	0.028	0.019	0.026	0.027	0.028
western Europe	-0.006	0.034	-0.014	0.057	-0.007	0.060	-0.021	0.057	0.016	0.034	0.031	0.033	0.035	0.034
inv. Mills ratio	-	-	-	-	-	-	-0.319***	0.042	-	-	-	-	-0.547***	0.047

means that a positive sign of the coefficient estimates mean better study grades if the value of the regressor is increasing.

The most important variables (measured by the lowest p-value) of the pre-university episode to predict the final degree grade are the overall grade at secondary school (better school grades generally lead to a better university grade), the school type (students who attended a general Gymnasium perform better at higher education), the gender (females perform better), the year of birth, the number of repeated classes in their school career (more repeated classes lead to worse university grades) and the final points in the school subject German (the better the results in German the better the university grades). The large importance of German grades at school is mainly caused by the fact that students who have good school grades in German tend to choose study fields like linguistics and cultural sciences more frequently. Students in these "soft" study fields have on average better grades than students of science, technology, engineering, mathematics (STEM) fields (Heublein et al., 2017), which can also be found in this data. For the same reasons students who had Mathematics as advanced course at school in some models perform significantly worse. These students are more frequently enrolled in "hard" study fields like Engineering or Mathematics.

Comparing the various modeling strategies minor differences in the parameter estimates can be found. The IPW estimates for some variables slightly differ from the other estimation strategies. The Heckman correction does generally not change the estimation results dramatically. Even though, the coefficient of the inverse Mills ratio is significant. Its negative coefficient indicates that students with worse grades are more prone for panel attrition as it was expected.

7.5.2 Early study phase

In this section, 54 additional variables describing the early study phase were added to the pre-university variables to a total number of 80 explanatory variables modeling the first higher education degree grade. This includes the selected study field and type of the higher education institution, the average grade of the first higher education exams, early study satisfaction, academic and social integration, financial aspects, off-study work, study commitment, and the big five personality traits.

Table 7.5: R^2 and MSE of the three different methods using using pre-university and early study phase variables (standard errors over the 20 imputations in parenthesis)

Measure	Include dropouts	OLS	Tweedie	IPW	Heckman
MSE	yes	0.705	0.703	0.731	0.693
		(0.013)	(0.021)	(0.028)	(0.013)
	no	0.120	0.120	0.124	0.120
		(0.003)	(0.003)	(0.003)	(0.003)
R^2	yes	0.415	0.550	0.593	0.550
		(0.016)	(0.029)	(0.054)	(0.031)
	no	0.522	0.522	0.545	0.523
		(0.022)	(0.022)	(0.025)	(0.022)

Table 7.5 illustrates the performance results of the four models regarding MSE and R^2. Additional information on the first semesters at university improves the regression performances of all models by far especially in terms of the R^2. Regarding the MSE the IPW estimator performs slightly worse than the other models but this model explains most of the variance. In the situation with zero-inflated distribution, where dropouts are included, the benchmark OLS model underperforms dramatically. In the situation without dropouts, all models reveal similar results since the usual OLS estimation is less problematic.

Figure 7.2 visualizes the out-of-sample predictions of all models by its kernel density estimation. In the left panel, where dropouts are included, one can see that the OLS estimator has massive problems to model the tails of the true distribution. Nevertheless, the other models also have problems in the prediction of dropouts with a grade of 5.0. If the interest is mainly on classification (dropout or graduate) and the specific grade estimation is not relevant, classification models are preferable to regression models. But even classification models tend to underestimate the proportion of the minority class (here the dropout students) by far. Behr et al. (2020c) used random forests to classify dropouts and graduates, whereby they adjusted the probability threshold to generate a larger number of classified dropouts.

The right panel of Figure 7.2 highlights that the IPW estimator is slightly left-shifted. The Heckman- and the OLS estimator are very similar in this situation.

Table 7.6 and 7.7 present the parameter estimates. The regression contains 90 ex-

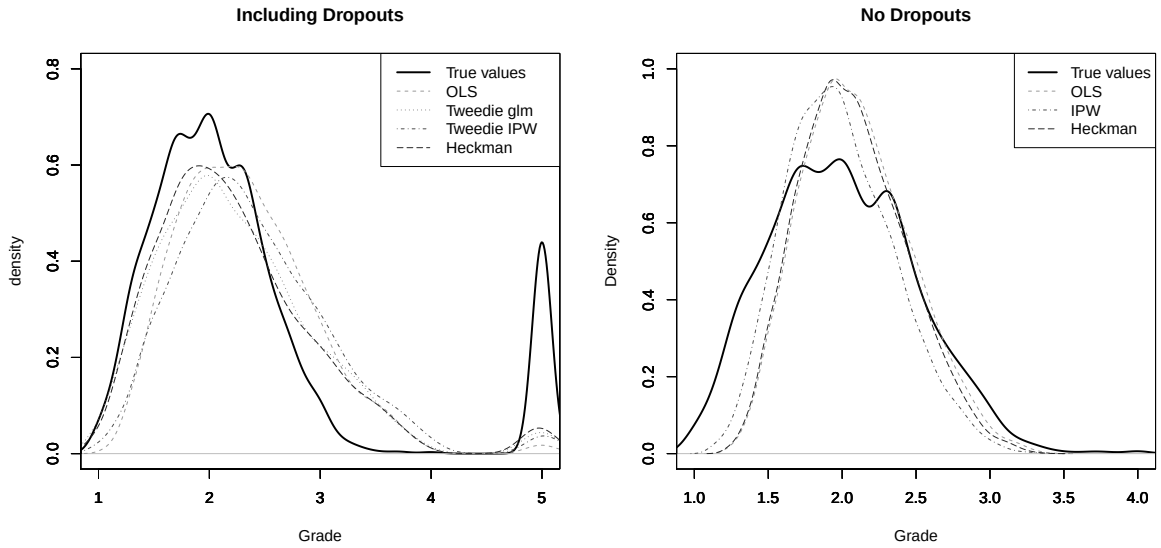

Figure 7.2: Kernel density estimation of true grades and the three models including dropout students (left panel) and excluding them (right panel) using variables of pre-university and early study phase.

planatory variables since two character variables (region of origin and study field) were converted to dummies. To reduce the number of coefficients displayed in the two tables, only coefficients were the estimate is significant to the 5% level for at least two of the seven models or where it is significant for at least one model to the 0.1% level. The other variables listed in the appendix but not shown in one of the Tables 7.6 or 7.7 were used for the regression but do not have a significant parameter estimate for more than one model.

The coefficient stronger varying between the models including dropouts. Furthermore, coefficient estimates from variables of the pre-university episode slightly differ from the estimates in Table 7.4 caused by the inclusion of other correlated variables which are also significant. For example, the points in the school subject German were highly significant in Table 7.4 but after adding the subject groups this effect decreases in the situation without dropouts and disappears in the situation including dropouts.

The most important new variables from the early study phase are the grade point average after the first exams at university, the own performance evaluation of the students, the type of institution and neuroticism (students with more confidence generally perform better). For most of these variables it is obvious why they are important and these findings are already widely discussed in the literature. The grade point average is even

Table 7.6: Parameter estimates and standard errors for the different models in the early study phase.
Significance codes: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$.

	including dropouts								no dropouts			
	OLS		Tweedie glm		Tweedie IPW		Heckman		OLS		IPW		Heckman	
Variable	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.
(Intercept)	-8.693***	1.973	-13.916***	1.426	-16.63***	1.473	-15.193***	1.307	3.02*	1.172	1.541	1.139	2.941*	1.171
rep_class	-0.035***	0.007	-0.027***	0.005	-0.015**	0.005	-0.02***	0.005	-0.009*	0.004	-0.015***	0.004	-0.007	0.004
familylife	0.015	0.008	0.02**	0.008	0.024**	0.008	0.023**	0.007	0.000	0.004	-0.004	0.004	-0.001	0.004
school_type	0.034***	0.008	0.039***	0.007	0.038***	0.007	0.035***	0.007	0.017***	0.004	0.017***	0.004	0.016***	0.004
grade_school	-0.097***	0.007	-0.106***	0.007	-0.118***	0.007	-0.093***	0.007	-0.045***	0.004	-0.051***	0.004	-0.044***	0.004
ger_points	0.000	0.001	0.001	0.002	-0.004**	0.001	0.001	0.002	0.002**	0.001	0.003***	0.001	0.002**	0.001
exam_adv_maths	-0.007	0.006	-0.015*	0.007	-0.011	0.007	-0.015*	0.006	-0.003	0.003	-0.007*	0.003	-0.003	0.003
birthyear	0.005***	0.001	0.008***	0.001	0.009***	0.001	0.008***	0.001	-0.001	0.001	0.000	0.001	-0.001	0.001
gender	0.017**	0.006	0.035***	0.007	0.031***	0.007	0.032***	0.007	0.009**	0.003	0.012***	0.003	0.009*	0.003
mother_job	0.001**	0.000	0.001***	0.000	0.001***	0.000	0.001***	0.000	0.000*	0.000	0.000**	0.000	0.000*	0.000
voctrain	0.015*	0.008	0.005	0.008	0.016*	0.008	0.007	0.008	-0.001	0.005	-0.003	0.004	-0.001	0.005
gpa_cur	-0.187***	0.006	-0.173***	0.006	-0.18***	0.006	-0.154***	0.006	-0.174***	0.004	-0.180***	0.004	-0.172***	0.004
study_alternative	-0.074***	0.011	-0.095***	0.009	-0.085***	0.009	-0.102***	0.008	-0.013*	0.007	-0.001	0.007	-0.013*	0.007
study_judge_parent	-0.003	0.004	0.005	0.004	0.015***	0.004	0.006	0.003	-0.004*	0.002	-0.006**	0.002	-0.004*	0.002
study_restrict	0.040***	0.006	0.057***	0.006	0.056***	0.006	0.054***	0.006	0.011**	0.003	0.009**	0.003	0.011**	0.003
partic_people	0.013*	0.006	0.011	0.006	0.016*	0.007	0.019**	0.006	0.004	0.003	0.001	0.003	0.005	0.003
partic_orga	0.001	0.005	-0.009	0.006	-0.022***	0.006	-0.013*	0.006	0.002	0.003	0.003	0.003	0.002	0.003
partic_facil	-0.002	0.005	-0.012*	0.006	-0.003	0.006	-0.012*	0.006	0.000	0.003	0.002	0.003	0.000	0.003
partic_acadskills	0.016**	0.006	0.022**	0.007	0.028***	0.007	0.017**	0.006	0.008*	0.003	0.007*	0.003	0.008*	0.003
preparation	0.003	0.005	0.010	0.006	0.012*	0.006	0.006	0.006	-0.005	0.003	-0.008**	0.003	-0.006*	0.003
workload_match	0.021***	0.003	0.040***	0.003	0.042***	0.003	0.031***	0.003	0.001	0.002	-0.003	0.002	0	0.002
performance_eval	0.019***	0.005	0.019***	0.006	0.037***	0.006	0.02***	0.006	0.010***	0.003	0.010***	0.003	0.010***	0.003
selfconcept	0.006	0.004	0.010*	0.004	0.005	0.004	0.012**	0.004	0.001	0.002	0.003	0.002	0.002	0.002
study_informed	-0.007*	0.003	-0.015***	0.003	-0.008*	0.003	-0.015***	0.003	-0.004	0.002	-0.003	0.002	-0.003	0.002

Table 7.7: Continuation of Table 7.6.
Reference category for the field are other minor subject groups. Significance codes: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$.

	including dropouts								no dropouts					
	OLS		Tweedie glm		Tweedie IPW		Heckman		OLS		IPW		Heckman	
Variable	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.	$\hat{\beta}$	std.
socint_instructors	-0.010	0.007	-0.017**	0.006	-0.032***	0.006	-0.026***	0.006	-0.002	0.004	0.001	0.004	-0.003	0.004
socint_students	0.005	0.004	0.014**	0.004	0.014**	0.004	0.011*	0.004	-0.007**	0.002	-0.007**	0.002	-0.007**	0.002
commit_necessary	-0.005	0.003	-0.010***	0.003	-0.011***	0.003	-0.011***	0.003	-0.001	0.002	-0.003*	0.002	-0.001	0.002
commit_enjoy	0.006	0.004	0.013***	0.004	0.008	0.004	0.015***	0.004	0.004	0.002	0.002	0.002	0.004	0.002
commit_identification	0.003	0.004	0.003	0.004	0.006	0.004	0.007	0.004	-0.005*	0.002	-0.004	0.002	-0.005*	0.002
job_semester	-0.001***	0.000	-0.001***	0.000	-0.001***	0.000	-0.002***	0.000	0.000	0.000	0.000	0.000	0.000	0.000
costs_direct	0.008*	0.003	0.011***	0.003	0.019***	0.003	0.010***	0.003	0.001	0.002	0.000	0.002	0.001	0.002
financialaid_bafoeg	-0.013*	0.005	-0.020***	0.006	-0.018**	0.006	-0.020***	0.006	-0.003	0.003	0.002	0.003	-0.002	0.003
change_field	0.015	0.011	0.027*	0.013	0.068***	0.014	0.021	0.013	0.009	0.006	0.005	0.006	0.009	0.006
satisf_enjoy	0.006*	0.003	0.006*	0.003	0.001	0.003	0.007**	0.003	0.004*	0.001	0.002	0.001	0.003*	0.001
satisf_whole	0.012***	0.002	0.019***	0.002	0.009***	0.002	0.016***	0.002	0.000	0.001	0.001	0.001	0.000	0.001
satisf_interesting	-0.005*	0.002	-0.01***	0.002	-0.005*	0.002	-0.011***	0.002	0.000	0.001	0.000	0.001	0.000	0.001
satisf_tired	0.001	0.001	0.001	0.001	0.005***	0.001	0.000	0.001	-0.001	0.001	-0.002**	0.001	-0.001	0.001
insttype	-0.054***	0.007	-0.065***	0.008	-0.058***	0.008	-0.032***	0.008	-0.035***	0.004	-0.031***	0.004	-0.030***	0.004
big5_extraversion	0.002	0.003	0.003	0.004	0.011**	0.004	0.004	0.003	-0.004*	0.002	-0.003	0.002	-0.004*	0.002
big5_agreeable	-0.002	0.005	-0.005	0.005	0.001	0.005	-0.006*	0.005	-0.004	0.003	-0.007**	0.003	-0.004	0.003
big5_conscientious	0.003	0.004	0.002	0.004	0.006	0.004	0.003	0.004	0.006**	0.002	0.005*	0.002	0.006**	0.002
big5_neuroticism	0.020***	0.004	0.031***	0.004	0.028***	0.004	0.029***	0.004	0.006**	0.002	0.007***	0.002	0.006**	0.002
arts	0.063**	0.022	0.071*	0.028	0.089**	0.028	0.075**	0.027	0.056***	0.012	0.076***	0.012	0.055***	0.012
mathematics and natural sciences	0.001	0.017	0.016	0.018	0.024	0.018	0.015	0.018	0.022*	0.009	0.029**	0.009	0.021*	0.009
linguistics and cultural sciences	0.013	0.017	0.015	0.018	0.012	0.018	0.021	0.018	0.021*	0.009	0.028**	0.009	0.021*	0.009
inv. Mills ratio	-	-	-	-	-	-	-0.175***	0.010	-	-	-	-	-0.337***	0.008

included in the output variable if only by a small percentage of the final grade. Students at universities of applied sciences have better grades, which might be the result of a lower requirement level. The variable indicating study restrictions is significant in all models but this is mainly caused by the large correlation with the secondary school grade.

There are much less significant coefficients in the situation where dropouts are excluded. Students who state higher values of conscientiousness have significantly better grades only in the models where dropouts are excluded.

Students in arts, linguistics and cultural sciences get significantly better grades. This also applies to mathematics and natural sciences but only if dropouts are excluded because there are larger dropout rates that were also found by Heublein et al. (2012) caused by many exams with large failure rates at the beginning of the study program.

7.6 Discussion and conclusion

This analysis aims to estimate the final degree grade of the first higher education degree of German freshman students who first enrolled in the winter term 2010/11. The data used for the study comes from the National Education Panel Study and contains in total 17,910 students and more than 3,000 variables.

Two different scenarios were analyzed in the study: 1) including study dropouts with a grade of 5.0 and 2) excluding study dropout in the regression models.

Furthermore, two sets of variables were used. The first set of variables only contains pre-university variables, which became relevant after the secondary education degree even before the final study decision process. In the second step variables of the early study phase were added to the models to investigate the model improvement in the second step when additional information is available.

A glm with the Tweedie distribution as exponential family is used to model the zero-inflation of the data if dropouts are included in the model. The predictive performance improves markedly when adding the additional variables of the early study phase from 0.204 to 0.550 in terms of R^2 when dropout students are included. The benefit of

the pre-university model is that predictions at a very early point directly after secondary school graduation are possible at the expense of model performance. Behr et al. (2020c) found similar results in a binary dropout-graduate classification analysis.

When dropouts are included in the regression model the results regarding influencing variables are predominantly in line with the previous dropout literature presented at the beginning of section 7.2. Interestingly, significant coefficients in the scenario including dropouts are not significant if dropouts are excluded from the model. These are mainly parameters that were found to be significant in the dropout literature if the reason for dropping out is not performance-related. Consequential, these variables mainly influence the dropout process but have only a minor influence on the final grade of graduates.

In some situations even the sign of the estimated coefficient changes. This applies, for example, for the age of a student, where a rising age has a negative influence in the dropout literature (Sarcletti and Müller, 2011), but has a positive influence on the grade if the student does not drop out. Müller and Schneider (2013), Lassibille and Gómez (2009) and Montmarquette et al. (2001) also found a larger dropout probability for older students. The possible reasons are higher opportunity costs for older students who already have experience in the labor market and the increasing financial (and social, if they have children) pressure if they already have a family. However, if older students graduate they can profit from their higher life experience.

Other variables that are good predictors for dropout but not for performance are study alternative, enjoying the degree program, study satisfaction, direct study costs, or the working hours during the semester. Many students are forced to work during the semester to raise the costs of their studies and already have a study alternative (their job). If they do not enjoy their degree program and are not satisfied they may leave the higher education institution despite good performance. Stinebrickner and Stinebrickner (2014) have an economic explanation for the dropout phenomenon. Students want to maximize their lifetime utility and if opportunity costs become too high or they expect only a minor increase of their salary with a higher education degree they frequently tend to leave the system without a degree.

The most important variables from the pre-university episode are the final grade at secondary school, the number of repeated classes in students' school career, the school

type, and the age. In the first semesters at the higher education institution, especially the average grades of the first exams become relevant. Whether the student is studying at the institution of choice, has an alternative (e.g. vocational training) to the degree program, whether there are study restrictions, and the type of institution, university or university of applied sciences, are important determinants already before the start of the first semester. During the early study phase, also the study satisfaction, the match of study workload and curriculum plan, and a weakly developed neuroticism have a positive influence on the final degree grade.

The limitations of the study are mainly data-driven. As in most survey datasets in panel design, the NEPS data also suffers under panel attrition, which leads to an overrepresentation of well-performing graduates. The Heckman correction and the inverse probability weight estimation should correct for this problem to get (ideally) unbiased parameter estimates. Strongly related to (temporary) panel attrition is the problem of missing values in the explanatory variables. The MAR assumption that is made for imputation is presumably not fulfilled for all missing values. This can also influences parameter estimates especially if the relative number of missing values is large which is especially the case for some early study variables (see Tables 7.8, 7.9 and 7.10).

The models presented in this study can help higher education institutions to implement early warning systems for students at risk. In contrast to other early warning systems that are only based on dropout prediction, e.g. (Knowles, 2015), this system can also send a warning to the students if they fail to meet their performance targets (e.g. a specific grade they want to reach). Students themselves can get extra motivation if they get early feedback from their institution.

A more detailed dataset, for example combining survey data with administrative data, covering more detailed information about the credit points earned and grades in single exams, would lead to further improvement of the models.

While this study predicts the final grade of the first higher education study program, which is generally a Bachelor program, a further research question would be to predict the grades of a Master program using information from the Bachelor courses. Since dropout rates in Master programs are lower (Heublein et al., 2017) different results can be expected and the estimated parameters of models including and excluding dropout students will be less markedly different. When doing this with the NEPS data the right censoring problem arises since there is a considerable number of students actually still

studying in the Master's program. Therefore also administrative data can be used where the problem of panel attrition does not exist but many "soft" variables as satisfaction are not available.

7.7 Appendix

OLS regression

In the OLS regression one is interested to get an estimate of the parameter vector β in the following equation

$$Y = \mathbf{X}\beta + \varepsilon \tag{7.6}$$

where $\mathbf{X} \sim (n \times k)$ denotes the design matrix containing the explanatory variables with n observations and k variables and $Y \sim (n \times 1)$ is the dependent variable containing the final grades for the first higher education degree and $\varepsilon \sim (n \times 1)$ is an error term (Hastie et al., 2009). The best estimate $\hat{\beta}$ is calculated as follows

$$\hat{\beta} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'Y. \tag{7.7}$$

Details of the imputation methods

Predictive mean matching (PMM): Let $x_i, i = 1, \ldots, k$ be one specific variable of the design matrix $\mathbf{X}$ containing missing values that should be imputed and x_{-i} be the other variables of $\mathbf{X}$ without x_i. Vink et al. (2014) suggest to split the PMM algorithm into the following steps:

1. Produce a linear regression, where the observed values x_i^{obs} with the regressor variables x_{-i}^{obs} are used to estimate a coefficient vector $\hat{\beta}$.

2. Randomly draw a coefficient vector β^* from the distribution of $\hat{\beta}$. The vector β is multinomial normal distributed with mean $\hat{\beta}$ and the empirical covariance estimation $Cov(\hat{\beta}) = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}[(x_{-i}^{obs})'x_{-i}^{obs}]$ with regression residuals $\hat{\varepsilon}$.

3. Generate predictions with the coefficient vector β^* for observed ($\hat{x_i^{obs}}$) as well as for missing values ($\hat{x_i^{mis}}$).

4. For each estimated missing value $x_{j,i}^{\hat{mis}}$, $j = 1, \ldots, m_i$ (where m_i indicates the number of missing values in variable i) calculate the distances to the observed values $\Delta_j = |x_{j,i}^{\hat{mis}} - x_i^{\hat{obs}}|$.

5. Take one random value of the d closest distances of Δ_j and insert the corresponding x_i^{obs} as imputed value for $x_{j,i}^{mis}$. Van Buuren (2018) suggests a value of $d \in \{5, 6, \ldots, 10\}$ if the dataset is not too small (less than 100 observations). Since the number of observations is relative large $d = 10$ is a good choice, even though the difference is negligible.

6. Follow the previous steps for all variables with missing values and repeat this to generate D complete datasets.

Multiple imputation: In multiple imputation D different datasets without any missing values are generated. After imputation, statistical standard methods can be applied to the complete data sets. Let $\hat{\theta}_d$ denote a point estimate and V_d the variance of a single imputed dataset $d, d = 1, \ldots, D$. The calculation of the combined estimate is done by simply averaging

$$\overline{\theta}_D = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d. \tag{7.8}$$

The variance of 7.8 has a within-component $\overline{W}_D = \frac{1}{D} \sum_{d=1}^{D} V_d$ averaging the D single imputation variances, and a between-imputation component $B_D = \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \overline{\theta}_D)^2$. The total variance of the combined estimate $\overline{\theta}_D$ is

$$\mathrm{Var}_D^{total} = \overline{W}_D + \frac{D+1}{D} B_D. \tag{7.9}$$

MICE-algorithm: As one can see in the first step of the PMM-algorithm, the method is constructed for univariate imputation, which means that only the imputed variable has missing entries. Otherwise, the linear regression would not be possible since linear regression cannot handle missing data. Multivariate imputation by chained equation (MICE), described by Van Buuren (2018), uses PMM as base model for imputation.

Any other univariate imputation model can also be used. MICE is an iterative algorithm where random samples from the observed values are used in the first iteration to make use of PMM which is applied to every variable with missing values. The missing values are updated in every iteration of the MICE-algorithm. Usually, five iterations are sufficient, which is also the default in the R-package MICE (van Buuren and Groothuis-Oudshoorn, 2011).

The best imputation method is a widely discussed topic in the literature. Saar-Tsechansky and Provost (2007) and Garciarena and Santana (2017) find different methods and strategies of dealing with missing data leading to the best out-of-sample performance results depending on the analyzed dataset. Several imputation strategies, like random forest imputation, hot deck imputation and pmm have been compared. The best out-of-sample R^2 was found for the pmm-method. Trivial methods like mean or median imputation can just be used for single imputation and have the disadvantage that the variance in the imputed variables is severely underestimated with consequences on confidence intervals and statistical tests (Kleinke et al., 2020).

Heckman correction

The mathematical details of the Heckman correction model are illustrated in Fox (2015). The first step is a regression on the latent response variable ξ which illustrates the observed values of Y:

$$\xi_i = \mathbf{X}_i\beta + \varepsilon_i. \tag{7.10}$$

In a second step the variable ϑ describes whether ξ is observed or not:

$$\vartheta_i = \mathbf{Z}_i\gamma + \delta_i. \tag{7.11}$$

We just observe the variable

$$Y_i = \begin{cases} \text{not available, if } \vartheta_i \leq 0 \\ \xi_i, \text{ if } \vartheta_i > 0. \end{cases} \tag{7.12}$$

The errors are assumed to be bivariate normal with mean zero, variances $Var(\varepsilon) = \sigma_\varepsilon^2$, $Var(\delta) = 1$ and correlation $Cor(\varepsilon, \delta) = \rho_{\varepsilon\delta}$. The expected value of Y_i given that Y_i is observed is

$$\begin{aligned} E(Y_i|\vartheta_i > 0) &= \mathbf{X}_i\beta + E(\varepsilon_i|\vartheta_i > 0) \\ &= \mathbf{X}_i\beta + \sigma_\varepsilon\rho_{\varepsilon\delta}m(-\mathbf{Z}_i\gamma), \end{aligned} \tag{7.13}$$

where $\lambda_i \equiv m(-\mathbf{Z}_i\gamma) = \phi(-\mathbf{Z}_i\gamma)/[1 - \Phi(-\mathbf{Z}_i\gamma)] = \phi(\mathbf{Z}_i\gamma)/[\Phi(\mathbf{Z}_i\gamma)]$ is the inverse Mills ratio. In a regression model where Y is only regressed on the design matrix $\mathbf{X}$ the additional effect λ is omitted which leads to biased estimates if the coefficient $\beta_\lambda \neq 0$ in the Heckman correction model in the following equation

$$(Y_i|\vartheta_i > 0) = \mathbf{X}_i\beta + \beta_\lambda\lambda_i + \nu_i. \tag{7.14}$$

Via probit regression we get estimates $\hat{\gamma}$ as defined in section 7.4.3. These are used to get estimates $\hat{\lambda}_i = \phi(\mathbf{Z}_i\hat{\gamma})/[\Phi(\mathbf{Z}_i\hat{\gamma})]$. In the second step model 7.14 can be estimated via OLS where the estimates $\hat{\lambda}$ are used instead of the true λ.

The main criticism of the model is that the estimation is inconsistent if the assumption of jointly normality of the error terms (ε, δ) is not fulfilled.

Tweedie distribution

Since the assumption of Gaussian distributed error terms in the OLS model is strongly violated for the data in the scenario where dropouts are included and the distribution of the modeled variable is not completely continuous generalized linear models (glm) are briefly introduced. More detailed information is in Fahrmeir and Tutz (2013). These models have three main components:

1. The exponential family, which is the Gaussian distribution in the standard OLS model.
2. The linear prediction $\eta = \mathbf{X}\beta$ which is also well known from the OLS model.
3. The link function g so that $\eta = g(\mu)$, where $\mu = E(\mathbf{Y}|\mathbf{X})$. The link function can also be nonlinear in glm's. In the OLS model g is just the identity function.

The family of Tweedie distributions are special cases of exponential dispersion models, which are a generalization of the exponential family. Therefore, the Tweedie distribution can be used in glm's (Shono, 2008). The density function of a Tweedie distributed random variable $\mathbf{Y}$ can be written as

$$f_{\mu,\sigma^2,p}(y) = a_{\sigma^2,p}(y)exp[-\frac{1}{2\sigma^2}d_{\mu,p}(y)], \tag{7.15}$$

where μ is the location parameter with $E(\mathbf{Y}) = \mu$, σ^2 is the dispersion parameter and p is the power parameter with $Var(\mathbf{Y}) = \sigma^2\mu^p$. For a power parameter p of 1, 2 and 3 one gets the Poisson, Gamma and the inverse Normal distribution, respectively. In this situation, with a zero-inflated continuous distribution one should select $p \in (1, 2)$ which leads to a compound Poisson-Gamma distribution. The parameter p is tuned via a 5-fold cross-validation in the set of values $p \in \{0\ , 1,\ 1.1,\ 1.2, \ldots, 1.9,\ 2, 3\}$. The Poisson-Gamma distribution combines a discrete Poisson distributed random variable $K \sim pois(\frac{\mu^{2-p}}{(2-p)\sigma^2})$ and iid random variables $Z_1, \ldots, Z_K \sim \Gamma(\frac{2-p}{p-1},\ \frac{\mu^{1-p}}{(p-1)\sigma^2})$ to the mixture distribution of equation 7.4.

The model comparison between the Tweedie distribution model and other models can be difficult and is usually limited to the comparison of the predictive results, as explained in section 7.4.6. Another disadvantage of the Tweedie model is that quasi-likelihood estimation is used to transform the Tweedie distribution to an exponential family. This impedes the calculation of widely used information criterions such as the Akaike information criterion or the Bayesian information criterion.

The tuned power parameter of the Tweedie distribution is $p = 1.6$ in the scenario with dropouts, which is a Poisson-Gamma distribution, and $p = 0$ when dropouts are excluded from the prediction, which is the Gaussian distribution.

Table 7.8: Description of pre-university variables.

Name	Description	Values and Notes	Wave	% NA
genstat	Generation status in Germany	1.0 to 4.0 (no immigration background)	1	0.000
rep_class	Number of repeated school classes	0 to 4	1	0.000
familylife	Living together with parents up to the age of 14	1 = with both parents, 0 else	1	0.023
school_type	Type of secondary school attended	1 = upper secondary education, 0 = other	spell	2.189
qualif_max	School-leaving qualification obtained	2 = general university entrence qualification, 1 = university of applied science qualification, 0 = other	spell	0.046
grade_school	Overall grade on final school certificate	1.0 (excellent) to 5.0 (poor)	1	1.765
math_points	Points in Mathematics in the last school year	0 (poor) to 15 (ecellent)	spell	30.675
ger_points	Points in German in the last school year	0 (poor) to 15 (ecellent)	spell	30.159
exam_german	German as examination subject for graduation	0 = no, 1 = yes	spell	6.337
exam_adv_german	German as first examination subject	0 = no, 1 = yes	spell	5.901
exam_maths	Mathematics as examination subject for graduation	0 = no, 1 = yes	spell	6.234
exam_adv_maths	Mathematics as first examination subject	0 = no, 1 = yes	spell	5.855
birthyear	Birth year of the target person	1950 to 1994	1	0.000
gender	Gender of the target person	0 = female, 1 = male	1	0.000
mother_qualif	CASMIN: mother's highest general school leaving qualification	0 to 8 (highest)	1	0.630
father_qualif	CASMIN: father's highest general school leaving qualification	0 to 8 (highest)	1	2.292
mother_job	ISEI-08: mother's occupation	11.74 to 88.96	1	26.366
father_job	ISEI-08: father's occupation	11.56 to 88.96	1	21.061
voctrain	Successfully completed vocational training before study	0 = no, 1 = yes	spell	0.000
fail_prestudy	Number of dropouts from other trainings before study	0 = no, 1 = yes	spell	0.000
mother_alive	Mother alive	0 = no, 1 = yes	1	0.046
fath_alive	Father alive	0 = no, 1 = yes	1	1.111
bilingual	Grew up bilingual	0 = no, 1 = yes	1	0.000
lang_ger	Speak German fluently	0 = no, 1 = yes	1	0.000
household	Number of household members	0 to 20	1	0.057
country	Region of origin	Germany, Western Europe, Eastern Europe, Afrika, Asia, America, Australia	1	0.011
Dependent variable				
degree_grade	Final degree grade in the first higher education degree	1 (excellent) to 4/5 (poor/dropout)	all waves	51.273

Table 7.9: Description of study related variables (first part).

Name	Description	Values and Notes	Wave	% NA
gpa_cur	Current grade point average	1.0 (excellent) to 4.0 (poor)	2	24.155
fieldofchoice	Studying the field of first choice	0 = no, 1 = yes	1	18.345
institofchoice	Studying at the institute of higher education of choice	0 = no, 1 = yes	1	14.151
study_alternative	Would you have started something else instead of a degree	0 = no, 1 = yes	1	14.335
study_judge_parent	Parents' opinion concerning study/field	1 to 5 (strong support)	1	14.060
study_judge_friend	Friends' opinion concerning study/field	1 to 5 (strong support)	1	14.048
study_restrict	Study with admission restrictions or selection procedure?	0 = no, 1 = yes	1	12.513
partic_people	Participation in university events aimed getting to know people	0 = no, 1 = yes	1	12.937
partic_orga	Participation in university events on study organization	0 = no, 1 = yes	1	13.682
partic_facil	Participation in university events with respect to central facilities	0 = no, 1 = yes	1	14.461
partic_acadskills	Participation in university events on academic skills	0 = no, 1 = yes	1	14.610
partic_course	Participation in university events on bridging courses	0 = no, 1 = yes	1	13.601
preparation	General preparation for study	0 (poor) to 4 (perfect)	2	22.917
skills_prep	Necessary knowledge acquired for subject field	1 (poor) to 4 (perfect)	2	23.261
workload_match	Matching study workload to curriculum plan?	1 (much less) to 5 (much more)	2	26.492
performance_eval	Satisfaction with academic performances till yet	1 (low) to 4 (high)	2	25.438
selfconcept	Academic self-concept (talent, learning new material etc.)	1 (low) to 7 (high)	2	26.263
study_informed	Informed about study	1 (poor) to 5 (good)	1	0.286
socint_instructors	Social integration: acceptance by instructors	1 (poor) to 4 (good)	2	25.450
socint_students	Social integration: contacts with students etc.	1 (poor) to 4 (good)	2	25.576
commit_necessary	Commitment to degree course: do no more than necessary	1 (does not apply) to 5 (applies completely)	2	25.656
commit_enjoy	Enjoyment of degree program	1 (does not apply) to 5 (applies completely)	2	25.839
commit_demands	High demands on self	1 (does not apply) to 5 (applies completely)	2	25.931
commit_energy	Invest a lot of energy for academic success	1 (does not apply) to 5 (applies completely)	2	25.828
commit_identification	Identification with degree program	1 (does not apply) to 5 (applies completely)	2	26.080
helplessness	Study-specific helplessness	1 (does not apply) to 5 (applies completely)	2	26.309
job_semester	Time spent in a week for employment during semester time	0 to 60 hours	2	24.682
study_semester	Time spend in study-oriented activities during semester	0 to 142 houres	2	24.648
job_break	Time spent in a week for employment during semester break	0 to 99 houres	2	24.728
study_break	Time spend in study-oriented activities during semester break	0 to 99 houres	2	24.739
costs_direct	Difficulty to pay direct costs of higher education	1 (difficult) to 5 (easy)	1	0.321
costs_opportunity	limitation of the possibilities to earn own money	1 (not at all) to 5 (a lot)	1	0.401
financialaid_bafoeg	Currently receive student financial aid (BAfoeG)	0 = no, 1 = yes	2	46.884
funding	Total amount of financial resources per month	0 to 10899 Euro	2	26.871

Table 7.10: Description of study related variables (second part) and interviewer variables.

Name	Description	Values and Notes	Wave	% NA
change_field	Have changed at least once the subject field	0 = no, 1 = yes	spell	0.000
change_uni	Have changed at least once the institution	0 = no, 1 = yes	spell	0.000
change_degree	Have changed at least once the type of degree (e.g. Bachelor instead of state examination)	0 = no, 1 = yes	spell	0.000
satisf_enjoy	Study satisfaction: enjoy degree course	0 (no) to 10 (absolutely)	3	16.501
satisf_conditions	Wishing better study conditions	0 (no) to 10 (absolutely)	3	16.523
satisf_match	Degree course and other obligations are hard to match	0 (no) to 10 (absolutely)	3	16.512
satisf_whole	On the whole, satisfied with actual studies	0 (no) to 10 (absolutely)	3	16.501
satisf_frustrating	External circumstances are frustrating	0 (no) to 10 (absolutely)	3	16.741
satisf_kill	Degree course kills me	0 (no) to 10 (absolutely)	3	16.523
satisf_interesting	Degree course is interesting	0 (no) to 10 (absolutely)	3	16.501
satisf_concerns	Concerns of students are not taken into account	0 (no) to 10 (absolutely)	3	16.695
satisf_tired	Often tired due to degree course	0 (no) to 10 (absolutely)	3	16.501
insttype	Type of institution	1 = university of applied sciences, 2 = university	spell	0.023
nontrad	Non traditional student	0 = no, 1 = yes	2	1.134
big5_extraversion	Big five personality traits: Extraversion	1 (outgoing) to 5 (reserved)	3	13.074
big5_agreeable	Agreeableness	1 (challenging) to 5 (friendly)	3	13.029
big5_conscientious	Conscientiousness	1 (efficient) to 5 (easy going)	3	13.017
big5_neuroticism	Neuroticism	1 (nervous) to 5 (confident)	3	13.017
big5_openness	Openness to experience	1 (curious) to 5 (cautious)	3	13.017
field	First field of study	engineering, mathematics and natural sciences law, economics and social sciences, linguistics and cultural sciences, arts, medicine, other minor subject groups	1	0.321
Interview specific variables to calculate the attrition probability				
nr_contact	Number of contact tries	0 to 129	all waves	0.000
minutes	Minutes of interview	13 to 136	all waves	0.000
int_gender	Gender of the interviewer	0 = female, 1 = male	all waves	0.000
int_age	Age of the interviewer	4 levels from 1 (young) to 4 (old)	all waves	0.000
int_school	Highest school degree of the interviewer	1 (no degree) to 4 (Gymnasium)	all waves	0.000
int_problems	Problems during the interview	1 = no problems, 0 = problems	all waves	0.000

8 Prediction of time-dependent dropout and graduation rates in higher education under the presence of panel attrition

Prediction of time-dependent dropout and graduation rates in higher education under the presence of panel attrition

Marco Giese

Chair of Statistics
University of Duisburg-Essen, 45117 Essen, Germany

Abstract

The prediction of students who drop out of higher education or graduate becomes increasingly essential to integrate early warning systems at universities. This study uses an extensive survey data set, namely the National Education Panel Study (NEPS), covering 14 waves and almost 18,000 students from German higher education institutions to model dropout, graduation, and study continuation probabilities of students from wave to wave. Synthetic oversampling of the dropout students is conducted to reduce a potential bias in the estimated probabilities since the dropout group is severely affected by panel attrition, which is common in survey data. The random forest was used as classification method since it reveals the best performance results for this data. The model reveals the best classification performance at the end of the standard study period, where the difference between graduates and dropouts increases. Variables covering prior education, as the final school grade, are most important at the study beginning, while study-related variables, e.g. the current grade point average, are important during the entire study period. Individual trajectories with dropout and graduation probabilities show that some students with large dropout chances at the study beginning later get the turnaround and graduate.

Keywords: panel attrition, higher education, educational data mining, oversampling, inverse probability weighting

8.1 Introduction

Study success and dropout are becoming more important since the dropout rate in Germany in 2016 in Bachelor programs is at a constantly high level of 29% (Heublein et al., 2017). On the one hand, dropout is often associated with personal failure and raises personal costs, especially opportunity costs (Behr et al., 2020a). Public expenditures in higher education institutions accumulate to almost 1% of the German gross domestic product in 2018 (in Zahlen, 2020). Study dropout is often associated with a waste of public financial resources, which leads to growing interest of policy-makers.

Students' decision to leave the higher education system without obtaining a first degree, which is, in general, a Bachelor's degree, is often a long process and rarely depends on a single determinant (Behr et al., 2020a). Modern data mining methods can help to combine many different features in a single model, while many data mining models do not need strong assumptions on the data as it is often the case for parametric models (Pochiraju and Seshadri, 2018).

In the research field of educational data mining (Baker et al., 2010) one aims, for example, to predict the higher education dropouts and graduates using classification models. In these situations, one is often just interested in predicting this binary target variable and it does not matter when the dropout or graduation took place (Kemper et al., 2019). Also, the longitudinal character of the dropout/graduation process is most often not incorporated in the model (Asif et al., 2014). Other studies focus only on graduation timing (Theune, 2015), where methods of survival analysis have been used.

Early identification of students at risk for dropping out can help universities to implement early warning systems (Knowles, 2015). Therefore, it is helpful when students' risk status, i.e. the probability to leave the higher education system in the next semester, is regularly updated when new information is available.

This study aims to identify higher education dropouts and graduates with classification methods taking the study duration into account. The data basis is the National Education Panel Study (NEPS), a large panel dataset covering many aspects of the German higher education system. This data is well suited to answer the research questions since the waves were surveyed approximately every six months, which is almost the same

rhythm as the semester in Germany (see Table 8.5 for details). Variables in a survey wave were used to predict a student's status (dropout, graduation, or study continuation) in the next wave. This approach incorporates the time-dependent prediction of students' status in the next wave.

The study's central research question is how well dropouts and graduates can be predicted by the classification model in different discrete time points. One further important question is which variables drive the process in the various waves. Does a study dropout already indicate itself in the previous waves by an increased dropout risk? The last major research questions deal with the development of the probability of dropping out, graduation, or study continuation over discrete time.

Since dropout students are more affected by panel attrition, a common problem in survey data, this group is underrepresented in the data. To generate a dropout proportion that is in line with dropout rates reported by Heublein et al. (2017), these students are synthetically oversampled. Additionally, inverse probability weights (IPW) should correct for over- and underrepresentation of students inside the three classes. Using only IPW would not generate the aimed dropout and graduation proportions in this situation.

The study is structured as follows. Section 8.2 gives an overview of related studies in the field of educational data mining. The dataset and its main limitations are discussed in section 8.3. Section 8.4 gives an overview of the general methodical approach and a brief discussion of the statistical methods used in this study. Empirical results are presented in section 8.5. A critical discussion of the results and a short conclusion are given in section 8.6.

8.2 Literature review

Binary dropout models: The field of educational data mining is a fast-growing branch of higher education research (Baker and Yacef, 2009). Many studies, e.g., Bayer et al. (2012), Abu-Oda and El-Halees (2015) or Sales et al. (2017), use models for binary classification to predict students' chances to graduate or to drop out. Baars et al. (2017) forecast medical students' failure after the first year of study, which circumvents the right censoring problem of up-to-date survey data sets. Therefore, they use the widely used logistic regression, which is also used by da Silva et al. (2017). Other popular

methods for the binary classification problem dropout vs. graduate are, for example, random forests, e.g. (Aulck et al., 2016), (Rovira et al., 2017), neural networks, e.g. (Saarela and Kärkkäinen, 2015), (Jadrić et al., 2010), support vector machines, e.g. (Manhães et al., 2014), (Mayra and Mauricio, 2018), and "weak learners" as Naïve Bayes (Mortagy et al., 2018), (Ramaswami and Bhaskaran, 2009), or decision trees (Shannaq et al., 2010), (Quadri and Kalyankar, 2010). Weak learners (Hastie et al., 2009) have the advantage of being fast in computation but usually show worse model performance than "strong learners". Also, the assumption of independent explanatory variables made by the Naïve Bayes classifier is strongly violated since the features in the educational context are usually correlated.

Model performance: The model performance is hard to compare between different studies since this depends mainly on the study's dataset. Behr et al. (2020c) revealed that one is usually interested in early prediction of study dropout to help students at risk before it is too late. The disadvantage of early prediction, using only variables of the pre-university episode, is that the model performance is worse compared to models, which also include variables of the second and third study semesters. Furthermore, university grades and credit points earned in the first study exams generally explain most of the dropout behavior, and studies using these variables mostly have an excellent model performance with accuracy values of up to 95% (Kemper et al., 2019). The model accuracy usually lies between 57.35% in the study of Superby et al. (2006) and 98% achieved by Mayra and Mauricio (2018). Since the accuracy is no good measure if the two classes are unbalanced (generally, there are more graduates than dropouts in the data), James et al. (2013) suggest to choose the area under the ROC-curve (AUC), which is not influenced by different proportions of the tow classes. This measure was used, for example, by Behr et al. (2020c).

Important variables: Behr et al. (2020a) give a detailed literature review where important variables influencing the dropout decision were described in more detail. They distinguish between three categories of variables depending on 1) the national education system, 2) the higher education institutions and 3) the individuals.

1) Müller and Schneider (2013) find that students from the upper secondary education pathway have better chances to complete tertiary education in Germany. Johnes and Taylor (1989) found a similar result for the UK. Students with a lower socio-economic background are generally disadvantaged regarding their educational performance (Schnepf, 2003). Dustmann (2004) also revealed the importance of parental

background for educational success. A country's financing policy can reduce the dropout rate by 2.6% for every 1,000 Euro per semester that students receive from the government for financial aid (Glocker, 2011). Major changes in the higher education system as the Bologna process in 1999 can also influence students' dropout decision even when Horstschräer and Sprietsma (2015) did not find significant differences in dropout rates before and after the Bologna process.

2) On the institutional level, Sarcletti and Müller (2011) exhibited larger dropout rates for private institutions compared to public institutions. Heublein et al. (2014) revealed that students at universities of applied sciences have a larger probability of graduating than students of general universities, where students are also prepared for a scientific career. The highest dropout rates are observed in "hard study fields" like engineering, natural sciences (Lassibille and Navarro Gómez, 2008), and mathematics, while low dropout rates were observed in "soft fields" like arts (Heublein et al., 2017). Furthermore, study conditions such as the teaching quality influence the dropout process (Georg, 2009).

3) Widely used variables are demographic factors such as gender and migration background. Most studies, e.g. (Ghignoni, 2017) and (Van Bragt et al., 2011b), find that female students tend to drop out of higher education programs with lower probability. A migration background has negative effects in most countries (Reisel and Brekke, 2009) but this also depends on the immigration policy and the country's financial aid system. Younger students were found to have better chances for graduation (Lassibille and Navarro Gómez, 2008). One of the most essential variables in many studies is the grade-point average (GPA) at secondary school. The GPA at secondary school is often highly correlated with the grades at higher education, and therefore the dropout probability at tertiary education (Stinebrickner and Stinebrickner, 2014), (Voelkle and Sander, 2008). Furthermore, some "soft" variables that are only available in survey data, as self-confidence (Brandstätter et al., 2006), study motivation (Schiefele et al., 2007), study organization (Schiefele et al., 2007) and degree program satisfaction (Suhre et al., 2007) affect students' dropout decision.

Longitudinal studies: While most studies in higher education research use cross-sectional data for empirical analysis, some studies use longitudinal or panel data. Haas and Hadjar (2019) find 27 studies since the turn of the millennium investigating students' trajectories in higher education using longitudinal data. US studies are strongly overrepresented in this research field. Chen (2012) investigates US students' dropout

behavior in a four-year period using methods of event history analysis. Therefore he used time constant features, e.g., gender, and other time-varying variables, and finds the largest dropout rates in the first study year with 17.7%. Aarkrog et al. (2018) investigate for a small sample of 31 Danish students their motivation over a period of eight weeks by weekly interviews. They find positive, negative, stable, and unstable development in students' attitudes regarding dropout. Meggiolaro et al. (2015) and Clerici et al. (2014) both use the same Italian administrative dataset covering more than 32,000 students from 2002 to 2005. Both studies consider the temporal dimension of the data. Pre-university variables, as well as socio-demographic variables, were important for a successful university career of students. Müller and Schneider (2013) analyze Germany's educational pathways, finding that the traditional way of students attending the "Gymnasium", which is the highest secondary education institution in Germany, decreases the dropout probability in tertiary education.

The novel approach of this study in the field of educational research is that it combines aspects of longitudinal studies with modern methods of data mining that are usually used in cross-sectional settings. The risk status is updated after every wave, which can help universities to implement an early warning system that is updated when new information about a student is available. Therefore, the model also covers information of previous waves to improve the predictive performance.

8.3 Survey dataset

The data covering the freshmen students (starting cohort 5) of the National Education Panel Study (NEPS) actually (August 2020) contains 14 waves. It covers the cohort of first-year students first enrolled at German higher education institutions in winter term 2010/2011. In the first wave, 17,910 students participated in the study. However, caused by panel attrition, the number of participants reduced to only 5,161 in wave 14. A detailed overview of the waves and the amount of temporary and final panel attrition is given in Table 8.5. More than 3,000 variables were surveyed, but not all of them can be used for the analysis. Some variables suffer from a low data quality (large proportion of missing values). This includes the actual number of students' credit points, which is a strong predictor in this classification problem (Baars et al., 2017), but only 2.9% state a value for this variable. Some other variables occur several times with different encoding, where only one of them was chosen. Other variables do not have

substantial relevance for our research question. The final dataset contains 98 explanatory variables, which are described in Tables 8.6, 8.7, 8.8 and 8.9 ordered by five thematic fields:

1. Prior education (15 variables)
2. Demographic and family variables (12 variables)
3. Higher education-related variables (30 variables)
4. Variables covering students' personal life (20 variables)
5. Satisfaction and personality of students (21 variables).

Furthermore, the out-of-sample estimates for graduation and dropout of the previous waves were included as explanatory variables in the next waves to improve classification performance, which can be seen as a sixth category of variables.

The variable of interest is the status of student i in wave t which is defined as

$$Y_i(t) = \begin{cases} 2, & \text{if the student has no degree and is still studying} \\ 1, & \text{finally leaving the higher education system without a degree} \\ 0, & \text{if a student gets a first higher education degree.} \end{cases}$$

This definition is based on Larsen et al. (2013c) and means that students changing their program or institution are not considered as "dropout" but as "still studying" which is also the initial status of all freshman students in the sample ($Y_i(1) = 2 \; \forall i = 1, \ldots, 17,910$). The time $t = 1$ can be seen as starting point of the higher education program since wave 1 was surveyed at the beginning/middle of the first semester before the first exams. If student i earns a Bachelor's degree in wave t^*, his status will be "graduate" for all following waves ($Y_i(t) = 0 \; \forall t \geq t^*$) even if he later drops out of a Master's program. The same is assumed for dropouts, since they state that they finally (and not temporary) left the higher education system, although it is possible that some dropout students return to higher education after wave 14. Variables regarding the status of a student were not surveyed in wave 1, which is about two months after the start of their studies.

Similar to the status $Y_i(t)$ also the explanatory variables were defined as time-dependent variables $X_i(t) = [X_i^{(1)}(t), \ldots, X_i^{(p_t)}(t)]$ where p_t is the time-dependent number of explanatory variables in each wave. Note that the number of available variables is not constant over time and most variables were not surveyed in each wave. Determinants that remain constant over time are mainly demographic variables as gender, year of birth, migration background and others. Further features covering information of secondary education or former vocational training also do not change after entering tertiary education. Time-varying variables as motivation, effort, social and academic integration, academic performance, information about the financial situation, off-campus work etc. were surveyed in irregular periods and not in each wave, sometimes just once. There are nine time-independent variables in the study which are widely discussed in the literature and also found to be important in a previous study of Behr et al. (2020c). These nine variables are used in all waves as explanatory variables to improve the classification performance, namely generation status, number of repeated school classes, family life up to the age of 14, school type, school-leaving qualification, grade at secondary school, year of birth, gender and whether the student completed a vocational training before he started studying.

Similar to the number of variables p_t used for prediction of the status in the next wave, also the number of participants is varying, denoted by n_t, which is the sample size used to predict the status in wave $t+1$ based on variables from wave t.

8.3.1 The attrition problem

The most severe problem of the data is caused by panel attrition, which occurs in most survey datasets (Little and Rubin, 2019). The percentage of final attrition accumulates to 44.76% in wave 14, but most attriters have either a degree or dropped out, while I am mainly interested in the question of what happens to students who are still studying. The attrition in the data is non-monotone (Seaman et al., 2016), which means that we have temporary survey dropouts who later return to the study. The non-monotone data pattern makes the application of many missing-data methods infeasible without discarding many non-monotone observations (Little and Rubin, 2019).

Especially, the class of dropout students is affected by panel attrition. Consequently, the dropout students are underrepresented in the study.

Heublein et al. (2017) report 29% dropout students in Bachelor programs in the same starting cohort of winter-term 2010/11 and 13.8% dropouts in Staatsexamen programs (calculated as the weighted mean of Medicine (11% dropouts), Law (24%) and Teaching (13%) in the graduation year 2014). This leads to an overall weighted mean of 24.348% dropout students based on the proportions of the NEPS data. This study finds only 1,013 dropout students (5.656%), 10,648 graduates (59.453%) and 6,249 students (34.891%) who are still studying or left the panel before graduation or withdrawal from the higher education system.

Table 8.1 reveals the underestimation of study dropouts and covers the theoretical proportion of dropouts (in percent of all dropouts) in each wave. Heublein et al. (2017) find considerably larger dropout rates in the first seven waves, so the observed number of dropouts is also adjusted by a theoretical number of dropouts that would have been observed for the NEPS data assuming the dropout rates found by Heublein et al. (2017). The column "dropout in %" states the percentage of dropouts in a specific wave based on the total number of dropouts, e.g., 41.132 % of all dropouts leave the higher education system in the first two semesters (Heublein et al., 2017). To overcome the strong underestimation of dropouts in the early waves, the theoretical number of dropouts in the last column of Table 8.1 is used to reduce the bias in the empirical results. Therefore, the dropouts are oversampled, as described in section 8.4.2.

Table 8.1: Number of observations n_t and predictors p_t used for prediction in each wave, the number of dropouts $Y(t+1) = 1$, graduates $Y(t+1) = 0$ and students still studying $Y(t+1) = 2$ in the following wave, the dropout rates reported by Heublein et al. (2017) (dropouts in %) and theoretical number of dropout in each wave.

t	n_t	p_t	observed dropouts	graduates	still studying	dropouts in %	theoretical dropouts
1	12.273	39	154	0	12119	41.132	1349
2	9.676	49	0	63	9613	13.451	328
3	9.245	32	161	87	8997	13.451	308
4	8.929	47	80	88	8761	5.266	115
5	8.150	36	0	1587	6563	5.266	106
6	5.089	50	0	1397	3692	3.630	45
7	3.431	25	0	638	2793	3.630	31
8	3.455	50	94	1384	1977	4.413	37
9	2.427	28	55	972	1400	2.021	12
10	983	39	14	267	702	2.021	5
11	700	40	7	275	418	2.021	3
12	574	34	23	264	287	2.021	3
13	165	39	8	37	120	2.021	1

8.4 Methodological approach

In a usual classification setting one is, for example, interested in estimating the dropout probability using early study phase variables up to wave 3, i.e. in estimating $P(Y_i = 1|X_i(t \leq 3) = x)$ Behr et al. (2020c). That means the dropout timing is not relevant in these studies as long as $t \leq 14$ (right censoring problem) and the individual participates in the study until the event of interest (attrition problem).

In this approach, the timing of dropout and graduation is of particular relevance and, in contrast to the approach of Behr et al. (2020c), the output has three possible outcomes, namely graduation, dropout, and study continuation. To answer the research questions, some statistical methods are briefly introduced. After a general methodical approach, the synthetic minority oversampling technique (SMOTE) is introduced, which is necessary to overcome the underrepresentation of dropouts. The basic concept of the classification algorithm, which is the random forest based on conditional inference trees (Hothorn et al., 2006) is provided in section 8.4.3. The following section presents measures to evaluate the classification performance of the model in a setting with three classes. Finally, I explain the principles of inverse probability weighting (IPW) and its application in the present situation.

8.4.1 General methodological approach

This study models the probability of graduating, dropping out, or continuing the study in the next wave. The students' actual status in wave t must be "still studying" to be of interest for the prediction of wave $t+1$, meaning that the focus lies on the estimation of the following probability:

$$\mathbf{P}\Big(Y_i(t+1) = y \mid Y_i(t) = 2,\ X_i(t) = [x_i^{(1)}(t), \ldots, x_i^{(p_t)}(t)]\Big). \tag{8.1}$$

This approach differs from many other studies in this research field, which are merely interested in graduation or dropout at any time of the study (see section 8.2).

Since student's status in wave $t+1$ is predicted with variables from wave t, the student has to participate in both waves t and $t+1$. Thus, forecasts are only possible up to wave 13, where the status of wave 14 is predicted.

The classification method can be applied in each wave similar to a cross-sectional classification setting (Aggarwal, 2015). Even when panel data is used in this study, the methods in this study come rather from the field of data mining than from event history analysis.

There are different statistical techniques to reduce, or ideally eliminate, the bias evoked by panel attrition. Some methods like inverse probability weighting (IPW) (Robins et al., 1995) or the Heckman correction (Heckman, 1976) take the response behavior of survey participants into account. Here, oversampling (see sections 8.3.1 and 8.4.2) is used to get realistic dropout proportions. Inverse probability weights are embedded in the classification algorithm to correct for the different response behavior of students inside the same class. Graduates and students who continue their studies are less affected by panel attrition than dropouts. The proportion of graduates after oversampling the dropout students already coincides with the graduation proportions reported by Heublein et al. (2017).

The statistical software R (Version 4.0.2) is used for all calculations in this article (R Core Team, 2019).

8.4.2 Dealing with panel attrition - synthetic minority oversampling technique (SMOTE)

SMOTE is a method to oversample minority class observations in a classification setting (Jeatrakul et al., 2010). In unbalanced classification settings, the classifier usually tends to prefer the majority class for new instances (Hastie et al., 2009) or even classify every new observation as the majority class. In this situation, the group of study dropouts is in most waves the minority class and strongly affected by panel attrition. This class is also underrepresented compared to the dropout rates found by Heublein et al. (2017). The dropout students up to wave 7 are oversampled with the SMOTE algorithm to generate this more realistic fraction of dropouts. In waves 2, 5, 6 and 7 no dropouts in the data emerge. This problem is solved using data of dropout students who participate in the previous wave $t-1$ but not in wave t ($t \in \{2,5,6,7\}$) that should be estimated so that they do not occur as dropouts in Table 8.1. From 1,013 dropout students found in the data, only 596 dropouts also appear in Table 8.1. The other 417 dropout students did not participate in the wave before they state that they dropped out and are potential candidates for waves 2, 5, 6 and 7 as long as they participate in wave $t-1$ that contains

the explanatory variables. Furthermore, next wave's status after wave $t-1$, where the student participates, should be "dropout".

The SMOTE algorithm, described by Jeatrakul et al. (2010), uses the observations in the minority class to generate new, synthetic instances in the minority class. Contrary to simple oversampling, where exactly the same observations are used multiple times, SMOTE reduces the problem of overfitting. Let $x := [x_i^{(1)}(t), \ldots, x_i^{(p_t)}(t)]$ be an observation of the minority class that should be oversampled. SMOTE searches for the k nearest neighbors of x (here I use $k = 5$) in the class of minority instances. One random observation $\tilde{x}$ from the k nearest neighbors is sampled and the difference $d = x - \tilde{x}$ is calculated. The new, synthetic observation of the minority class is $x_{new} = x + r \cdot d$, where r is a random number drawn from a rectangular distribution on the interval $[0, 1]$.

One crucial aspect of oversampling is that in performance prediction, the oversampled instances should only be used to train the model (Hastie et al., 2009). Otherwise, the oversampled instances in the train and test data can be (nearly) identical, leading to over-optimistic prediction results. To evaluate the predictive performance of the model ten-fold cross-validation (CV) repeated 20 times to reduce the variance of the prediction (Krstajic et al., 2014) is applied. The random splitting in the ten groups takes place before the oversampling.

The R package "DMwR" (Torgo, 2010) contains the SMOTE algorithm used in this study.

8.4.3 Random forest based on conditional inference trees (cforest)

There is no general classification algorithm that always shows the best performance on all datasets (Pochiraju and Seshadri, 2018). A prior study of Behr et al. (2020c) reveals that cforests are well suited for this dataset. Further model comparisons illustrate that tree-based models outperform other classification models (e.g. Naive Bayes, logistic regression, support vector machine) by far. Moreover, the currently hyped neuronal networks (Pochiraju and Seshadri, 2018) have been tested on this data but they also perform slightly worse compared to cforests. Decision trees have the advantages that they can handle missing values (with surrogate splits), are robust against extreme

outliers, have a kind of inner feature selection, can handle all types of variables (nominal, ordinal, and metric), and do not make particular assumptions (Hastie et al., 2009, Pochiraju and Seshadri, 2018).

Conditional inference trees (Hothorn et al., 2006) split the data, beginning in a root-node, by testing the independence hypothesis between Y and X with a multiple permutation test. Therefore, the single p_t hypothesis of independence between $Y(t)$ and $X^{(1)}(t)$, $\ldots$, $Y(t)$ and $X^{(p_t)}(t)$ were tested, whereby the global significance level is adjusted with the Bonferroni correction. The variable $\tilde{X}^{(q)}, q \in 1, \ldots, p_t$ with the smallest p-value is selected for the splitting rule. The splitting stops if the test cannot reject the null hypothesis.

Cforests aggregate B ctrees to a single classifier to overcome the problem of the large variance of a single tree (Breiman et al., 1984). The observations for every single tree are drawn via bootstrapping (with replacement). An increasing number of trees generally improves the classification performance on costs of a larger computation time. In this situation, $B = 100$ should be sufficient for performance prediction. The variable importance measure is affected by much larger variance. Therefore $B = 1,000$ trees were used for the importance ranking where also no test-holdout is necessary. To calculate the importance of a single variable, the classification problem is conducted in a first step with all variables and in a second step without the variable of interest. The difference between both predicted performances is the importance. If the variable is important, the classification performance should be significantly worse without this variable. Since absolute importance values calculated by the cforest are hard to interpret, the relative importance of variable i is computed as function of the discrete-time t (here the wave) as
$imp_i^{rel}(t) = imp_i^{abs}(t) / \sum_{j=1}^{p_t} imp_j^{abs}(t)$. The cforest is implemented in the "party" package (Hothorn et al., 2018) in R.

8.4.4 Performance measures

The classification problem in this situation is a multi-class problem with $c = 3$ different classes (dropout, graduation, study continuation). Most performance measures for binary classification, such as accuracy, recall, or precision, can be expanded to multi-class problems (Sokolova and Lapalme, 2009). The confusion matrix contains the true class in the rows and the predicted class in the columns and states, for example, how often

dropouts are also classified as dropouts or misclassified as graduates or students who continue their studies. Misclassification often leads to different costs, e.g., classifying a dropout as a graduate can be more "expensive" than classifying a graduate as a student who continues the study (which is the majority class in all waves). Therefore, the costs matrix, which is usually determined by the organization that uses the model (here, e.g., a university), states the costs of misclassification and has the same dimension as the confusion matrix (Krawczyk et al., 2014).

Since the measures above, calculated with the entries of the confusion matrix, are influenced by the fact that the classes are imbalanced and strongly depend on the entries in the cost matrix, the area under the curve (AUC) is used as the primary performance measure which is independent of the cost matrix and not affected by different class sizes (Han et al., 2011, James et al., 2013). The AUC takes values between 0 and 1, when plotting the true positive rate against the false positive rate in the binary classification setting. A value of 0.5 is a random guess and 1 is a perfect classification. Hand and Till (2001) generalize the concept of the AUC for more than two classes by merely averaging over the $\binom{c}{2}$ pairwise AUC values.

8.4.5 Inverse probability weighting (IPW)

The IPW estimation should correct for different response probabilities (Robins et al., 1995). Participants with higher response probabilities are generally overrepresented in the data, which is compensated with smaller weights.

Note that a student is only relevant for estimating the status in wave t, $t = 2, \ldots, 14$ if he or she also participated in wave $t-1$ containing the predictor variables. The IWP estimation is a very intuitive method to correct for different response behaviors. It is a two-step procedure wherein the first step response probabilities for participation in both waves of interest (t and $t-1$) are calculated. Here, the cforest is used for this binary classification problem to estimate the participation probability $\hat{\pi}_i^t$, in the next wave t for student i with variables from the last wave $t-1$. Additionally, three variables regarding the interview (contact tries, problems during the interview and length of the interview) and three variables describing the interviewer (age, gender and school degree) in wave $t-1$ were used for classification since these variables have been found to influence the response probability (Zinn, 2019). Weights are estimated using the inverse response probabilities $\hat{w}_i^t = 1/\hat{\pi}_i^t$. To obtain the estimated participation probability in

two following waves $\hat{\pi}_i^{t,t+1}$ of interest the participation probabilities of the two relevant waves are multiplied under the assumption of stochastic independence of the single participation probabilities: $\hat{\pi}_i^{t,t+1} = \hat{\pi}_i^t \hat{\pi}_i^{t+1}$. The assumption of independence might slightly be violated since Zinn (2019) reports a significant coefficient for participation in prior waves.

In the second step, the weights are integrated into the cforest, which is now used for status prediction. In the cforest weighting means that in the bootstrap sampling each observation is drawn with a probability proportional to their weight (Hothorn et al., 2018).

Dropouts have on average slightly higher weights than students of the two other groups. At this step, the dropout group is already oversampled to generate the same dropout proportions as Heublein et al. (2017). Further overweighting of the dropout group would lead to an overestimation of the dropout probabilities. To avoid this problem, the average weight in each of the three classes (graduate, dropout, study continuation) should be equal by dividing each weight by the class-specific average weight.

Contrary to many other methods dealing with panel attrition (Van Buuren, 2018), IPW estimation makes comparably soft assumptions. It leads to unbiased estimates if the response probabilities are known. Therefore, precise estimation of students' response behavior is essential (Robins et al., 1995).

Since the model is tested with out-of-sample observations from the same population (which might not represent the total population of German higher education students), it is not expected that IPW estimates are better than estimations with equal weights regarding their model performance. However, it should give better results if the model is applied in practice and, additionally, the variable importance ranking should be less biased.

Since a major aim is to get preferably unbiased estimates of graduation-, dropout-, and study continuation probabilities for the next wave, both steps, oversampling and IPW are necessary. The IPW alone would not put enough weight on the dropout group in the first waves, leading to an underestimation of dropout probabilities. For example, this can be caused by a biased estimation of response probabilities or biased estimates of dropout and graduation proportions reported by Heublein et al. (2017). Oversampling alone only generates the correct proportions in each of the three groups. However, it would only lead to unbiased estimates if the non-respondents in each of the three classes are missing

at random (MAR) (Kleinke et al., 2020). This means that the participants in each class and wave must be representative for the respective population. This assumption is not realistic for this situation since the response rates also depend on other variables such as gender, age and study field (Zinn, 2019).

8.5 Empirical results

8.5.1 Performance results

Table 8.2 shows the AUC values and standard errors of the classification model in each wave, where wave t means that explanatory variables of wave t were used to predict the status in wave $t + 1$. As expected, the model performance for all waves, tested with out-of-sample observations inside a ten-fold cross-validation, is about 3% worse than a model ignoring the attrition problem (without oversampling and IPW estimates). Best predictions were made in wave 8 where we have many predictors for estimation, a relatively large number of graduates, and enough observations to train the model (see 8.5). Furthermore, many graduation and dropout probabilities of previous waves can be used to fit the model, which are relatively strong predictors, as we will see in 8.5.2.

After wave 8 the classification performance worsens, which has two main reasons. On the one hand, the number of training- (and test-) observations is monotonically decreasing because only students who are still studying and still participating in the survey are relevant for the model in the next wave. This has considerable consequences in wave 13 where classification is not much better than a random guess with huge variance. On the other hand, after wave 8 not many variables with large predictive power for this research question were surveyed. This is discussed in more detail in the next section.

Additionally, Table 8.2 illustrates the pairwise AUC values, calculated by only taking test data of two classes and their specific class probabilities into account. There are only two available classes in the waves with missing entries in the test data, meaning that the multi-class AUC is also the pairwise AUC of these two available classes. Due to the small number of graduates and dropouts at the beginning of the study in waves 3 and 4 the model has problems to distinguish between these two classes. In

Table 8.2: Classification performance measured by AUC in the various waves (standard deviations in parenthesis) for the three-class problem and pairwise AUC values.

wave	1	2	3	4	5	6	7
AUC	0.763	0.686	0.791	0.738	0.676	0.782	0.718
sd(AUC)	(0.068)	(0.089)	(0.087)	(0.069)	(0.082)	(0.072)	(0.086)
dropout-graduate			0.560 (0.233)	0.731 (0.128)			
dropout-continue	0.763 (0.068)		0.897 (0.105)	0.840 (0.097)			
graduate-continue		0.686 (0.089)	0.917 (0.135)	0.643 (0.141)	0.676 (0.082)	0.782 (0.072)	0.718 (0.086)

wave	8	9	10	11	12	13
AUC	0.809	0.687	0.711	0.687	0.722	0.528
sd(AUC)	(0.117)	(0.089)	(0.081)	(0.104)	(0.099)	(0.223)
dropout-graduate	0.858 (0.262)	0.901 (0.229)	0.930 (0.132)	0.795 (0.292)	0.791 (0.189)	0.661 (0.418)
dropout-continue	0.849 (0.188)	0.534 (0.117)	0.628 (0.165)	0.493 (0.132)	0.573 (0.189)	0.460 (0.338)
graduate-continue	0.719 (0.138)	0.630 (0.103)	0.662 (0.117)	0.706 (0.093)	0.756 (0.102)	0.578 (0.354)

this situation, the estimated variance is also comparably large. In later waves, where the number of graduates is large enough to have sufficient training data, the model best distinguishes between graduates and dropouts since these two groups differ the most.

In the early semesters, the model precisely separates between higher education dropouts and students who continue their studies, which is relevant, e.g., to implement early warning systems where an early prediction of dropouts is important (Knowles, 2015). From wave 9 the model has problems separating these two classes, which is again caused by the very small number of dropouts in later waves.

8.5.2 Variable importance

In the description of the dataset in section 8.3 was already explained that the variables were divided into five thematic fields plus a sixth category of variables covering

the dropout and graduation probabilities of the previous waves. The out-of-sample estimates of dropout and graduation probabilities were obtained by a 20-fold repeated cross-validation in the specific wave. Figure 8.1 reveals the relative importance of the variables in the six fields for each wave. Since the relative importance of groups of variables is calculated as the sum of the relative importance of the single variables in the specific group, it strongly depends on the (relative) number of variables surveyed in each category and wave, which is also plotted. As before, results in wave 13 are unreliable due to the minimal number of instances for training the model.

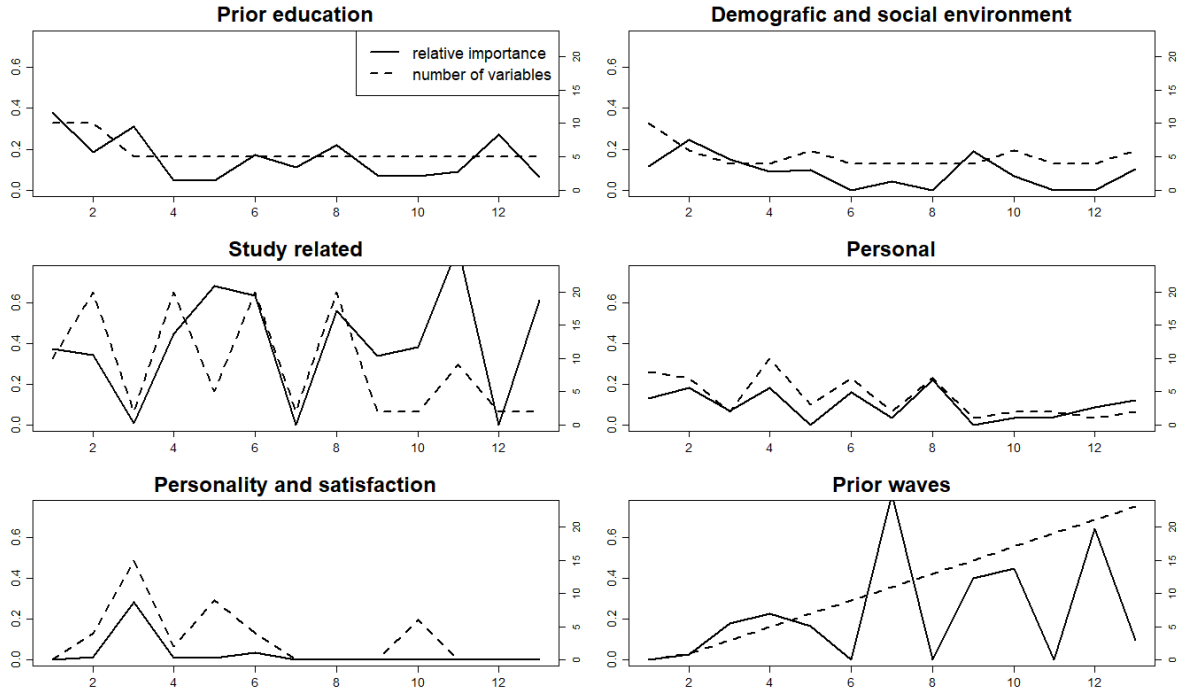

Figure 8.1: Relative importance (left axis, solid line) and the number of variables (right axis, dashed line) of the six groups of variables dependent on the wave (x-axis).

Dropout and graduation probabilities of prior waves become more important in later waves since the number of variables is rising linearly (two more variables each wave). We have more information on students' prior probabilities of leaving the higher education system (with or without a degree). While study related variables are essential during the entire educational career, prior education is especially relevant in the first semesters of study. Former studies with this data (Behr et al., 2020c) reveal that the grade point average at secondary school is the most relevant variable describing the dropout process in a time-invariant binary classification model. While demographic determinants, which

become relevant before entering the higher education system, play a role at the beginning of the studies caused by different starting conditions, social disadvantages decrease after the early study phase. The problem of personality and satisfaction variables in this dataset is that most of these variables were surveyed only in wave 3 and 5, making time-dependent statistical modeling less meaningful, also including econometric models like fixed- or random effects models (Baltagi, 2008). Variables covering students' personal (off-campus) life have roughly a time constant relevance for students' status at higher education but are less relevant than other features.

Table 8.3 gives an overview of the 5 most important single variables in each wave. We have already seen in Figure 8.1 that there are not enough observations in wave 13 to get a reliable ranking, so it is excluded.

Table 8.3: Top 5 variable ranking from wave 1 to 12 and relative importance.

wave	Top 5 variable ranking				
	1.	2.	3.	4.	5.
1	probsuccess 0.151	exam_maths 0.130	study_alternative 0.071	qualif_max 0.067	school_type 0.047
2	birthyear 0.197	job_semester 0.110	grade_school 0.096	courses_semester 0.075	probsuccess 0.043
3	rep_class 0.138	birthyear 0.107	school_type 0.100	dropout_w2 0.087	satisf_whole 0.066
4	probsuccess 0.167	dropout_w3 0.125	graduate_w3 0.095	study_importance 0.090	job_semester 0.077
5	field 0.436	dropout_w3 0.076	gender 0.063	dropout_w2 0.055	qualif_max 0.051
6	probsuccess 0.212	courses_semester 0.145	job_break 0.105	learn_semester 0.104	perform_better 0.073
7	graduate_w6 0.232	graduate_w5 0.189	field 0.178	birthyear 0.153	gender 0.109
8	courses_semester 0.378	graduate_w7 0.209	grade_school 0.189	workload_match 0.126	perform_better 0.110
9	field 0.346	gender 0.102	graduate_w5 0.093	dropout_w5 0.073	dropout_w8 0.071
10	field 0.383	dropout_w9 0.089	graduate_w5 0.075	dropout_w2 0.067	dropout_w8 0.064
11	workload_match 0.345	grade_school 0.248	courses_semester 0.185	perform_expect 0.144	graduate_w10 0.064
12	graduate_w11 0.312	grade_school 0.250	dropout_w2 0.147	health 0.101	dropout_w10 0.054

Note that high importance does not directly mean that this variable is important for splitting graduates from dropouts since there are three classes. It also does not say anything about the values of a variable in each of the three groups.

Students' subjective probability of success, the current grade point average and the final grade at secondary school are three of the most important variables. The dropout or graduation probabilities of the (two) previous waves also have a considerable influence. For example, students who have large graduation probabilities in wave 5 but continued their studies also have larger graduation probabilities in wave 7. This is shown in more detail in section 8.5.3.

8.5.3 Dropout and graduation trajectories

This article's central research question is how the probabilities of graduation, dropout, and study continuation develop over time for different students. Students were allocated into three groups according to their final status in the last wave of participation. Since early panel attrition is often accompanied by the final status "still studying", students in this group also have to participate in waves 8 or one of the following waves. Wave 8 was surveyed after the 8th semester of study, which is more than two semesters longer than the standard period in most study fields, so these students can be regarded as "long term students". Table 8.4 shows the number of students in each of these groups in the 13 waves.

Table 8.4: Number of participants in each wave in the groups of graduates, dropouts and students who are still studying.

wave	1	2	3	4	5	6	7	8	9	10	11	12	13
Graduates	8367	7406	7266	7360	6950	4427	2886	2954	1928	738	491	298	37
Dropouts	655	431	448	303	182	170	162	165	118	38	27	30	8
Still studying	923	796	775	735	651	384	383	336	381	207	182	246	120

The rapidly decreasing number of graduates after wave six is caused by the fact that only students who are still studying were used for classification in the next wave. For dropouts the numbers are already decreasing from the study beginning since most dropout students leave the higher education system in the early semesters.

Figure 8.2 reveals the average probabilities for graduation, dropout, and study continuation in each of the three groups from wave to wave. The highest graduation probabilities are, unsurprisingly, in the group of graduates and the lowest in the group of dropouts, while in the latter group dropout probabilities are by far the largest. In wave 8 one can observe an increase of dropout probabilities caused by many dropout-specific features

surveyed in that wave (see also Table 8.1). Note that the probabilities in Figure 8.2 cannot be calculated with the entries of the confusion matrix of each wave because here the three classes were allocated by the final status of a student while the confusion matrix regards the wave specific status.

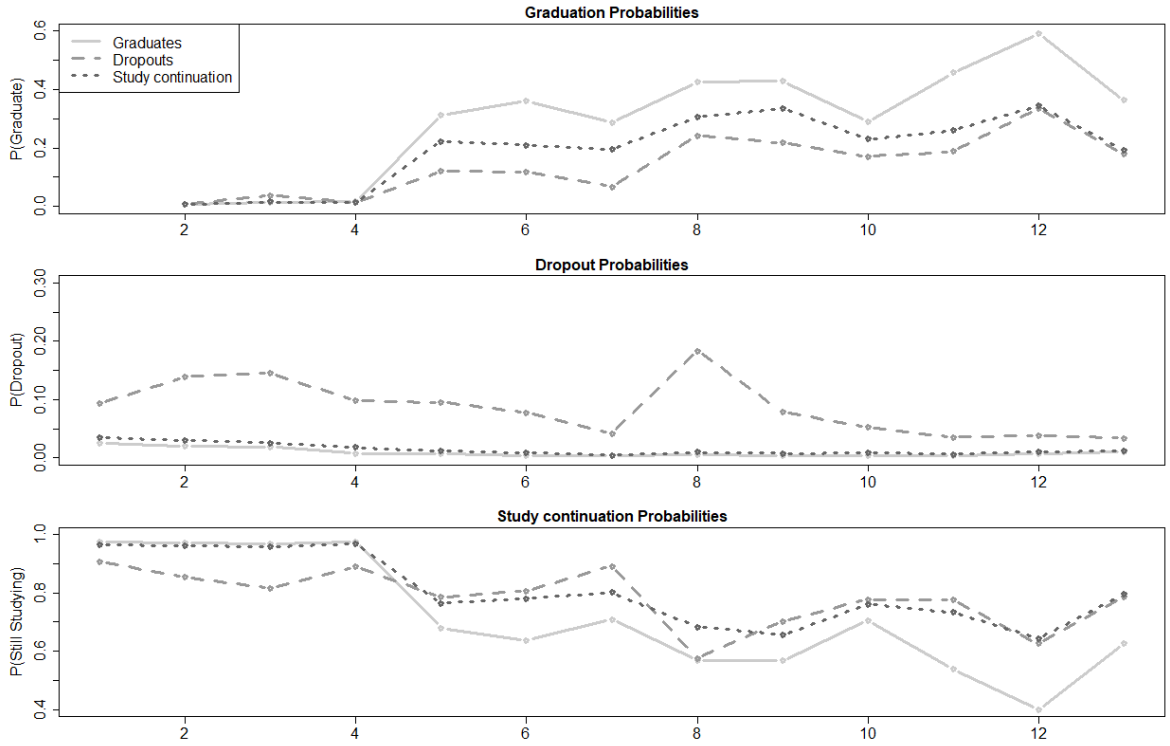

Figure 8.2: Mean trajectories of graduation, dropout and study continuation probabilities.

For the study continuation probabilities, the situation is less clear. Up to wave 4 dropouts have the lowest continuation probability since most of them leave the higher education system in the first two years (Heublein et al., 2017). After wave 5 the group of graduates has the lowest study continuation probability since this is the time where most students finish their studies. It becomes clear that the model can better distinguish between dropouts and graduates than between one of these groups and study continuers.

To take a closer look at the two groups of graduates and dropouts, Figure 8.3 compares these two groups regarding their dropout and graduation probabilities. The four spaghetti plots reveal individual trajectories until wave 12. The plot excludes individuals who did not participate in the last two waves before they state their final status

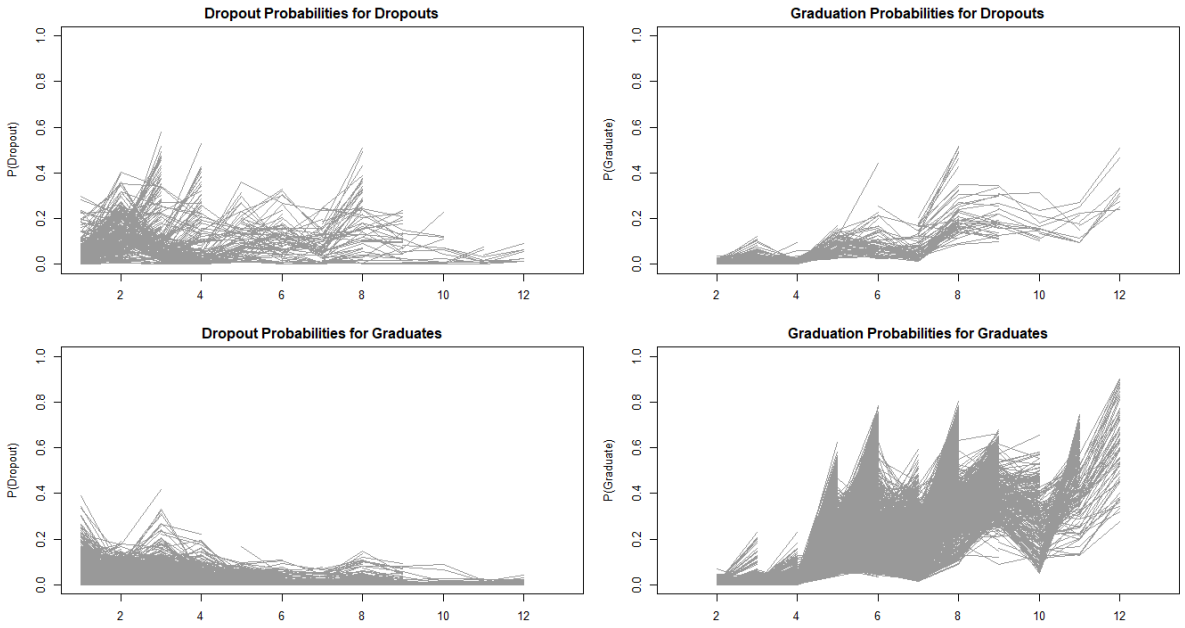

Figure 8.3: Graduation and dropout probabilities in the groups of observed graduates and dropouts

because then no probability estimates would be available for the final status of a student. One can see that the trajectories of the highest dropout probabilities for the group of dropouts and the trajectories of the largest graduation probabilities for graduates often end after the specific wave. This generally means that the students reach their final status in that wave and the model also predicts a high probability for this status.

8.6 Discussion and conclusion

This study aims to estimate the time-dependent status of higher education students. The status of a student is defined as a categorical variable with three possible values: (0) graduation, (1) dropout, (2) study continuation. This leads to a time-dependent classification problem with three possible outcome classes. The random forest based on conditional inference trees is best suited for this problem and outperforms other popular methods like neuronal networks.

The main limitations of this study are data-driven. Panel attrition is a severe problem in many survey datasets, including the NEPS, which has been used for this study. The

attrition problem mainly concerns the class of higher education dropouts, where the rate is clearly under the dropout rates of Heublein et al. (2017). The class of dropouts was oversampled to generate classes in each wave comparable with the current dropout literature. Synthetic oversampling reduces the problem of overfitting, compared to usual oversampling where the model is trained with the same observations of the minority class multiple times. However, oversampled dropout instances are similar to the original dropout instances.

Furthermore, it is likely that missing instances, even inside each of the three classes, are not missing at random (Little and Rubin, 2019). Therefore, inverse probability weights are introduced to put more weight on the students with lower response probabilities. Inverse probability weights without oversampling would lead to an underestimation of dropout probabilities in this situation. Oversampling without inverse probability weighting would ignore the structure behind the missing mechanism inside each class (Kleinke et al., 2020). To avoid that students in the dropout group are overweighted a second time (oversampling can also be seen as a kind of weighting), the weights in each class are divided by the mean weight of the class.

Behr et al. (2020c) reveal that under some assumptions, the problem of panel attrition does not have significant negative consequences for their machine learning models with the NEPS data. Behr et al. (2020c) expect that performance-related dropout students in early waves are more frequently affected by panel attrition. Only by making strong (and not realistic) assumptions for this scenario it would be possible to eliminate the bias caused by panel attrition completely (Van Buuren, 2018). However, the approach in this study should help to reduce the bias.

One further problem regarding the data was the vast number of missing values for some variables, especially for the earned credit points which are presumably a strong predictor for this classification problem. These variables were excluded from the analysis. Many variables were only surveyed in one or two waves but not in each wave (as necessary for many statistical panel data models). These two aspects will have negative consequences for model performance.

The AUC is used as central performance measure for classification since it is not influenced by different class sizes. The model performance in this study depends, on the one hand, on the number of instances in the training data, which is no problem with more than $\backsim$1,000 instances (wave 1 to 9) but dramatically decreases with less than

$\backsim$500 instances (wave 13). On the other hand, it depends on the number, quality, and importance of available variables in the specific wave. The AUC reaches its maximum in wave 8 (after the 8th semester at university) with a value of 0.825. In the later waves, the groups of dropouts and graduates become more dissimilar, making it easier for the model to distinguish between these two groups.

The importance of the variables in the models depends strongly on the survey wave. In the early study episode (the first three waves) prior education variables, e.g., the final grade at secondary school or the number of repeated school classes, are of particular importance. This also applies to demographic variables, such as the year of birth; however, these variables are less important than prior education variables. Study related variables, e.g., the study field, current grade point average, or the weekly hours spent in the study, have approximately the same (high) relevance during the complete study duration. At the end of the study, information on prior dropout and graduation probabilities becomes much more important and such variables are the best predictors for later waves.

Finally, the individual dropout, graduation, and continuation probabilities reveal that after wave 4 the groups of the later dropouts and graduates differ more strongly than in the early study phase, while the group of long-term students, who did not earn a degree at the end of the study, is somewhere between dropouts and graduates. Individual trajectories also reveal outliers in all groups, such as later graduates with huge dropout risk in wave 3.

The latter aspect also shows the potential value of early warning systems at universities. Such models can be implemented in practice, and if a student has a large chance of dropping out, the institution can initiate countermeasures, such as extra tutorials for learning techniques or personal mentoring for students at risk. Furthermore, the students can be informed, e.g., after each semester, and can use this feedback to get extra motivation for the next semester.

In practice, such models would be fitted with administrative data that are available at all institutions. This would not lead to the attrition and missing value problem that exists in survey data. Instead, the number of variables would be limited, and information about satisfaction, motivation, personality, and many study-related variables would not be available. These variables can be implemented by an additional survey, leading to a combination of administrative- and survey data.

8.7 Appendix

Table 8.5: Participants and panel dropouts in the current scientific use file (SUF) (Würbach, 2020, and own calculations). CATI: Computer assisted telephone interview, CAWI: Computer assisted web interview.

wave	instru-ment	partic. survey	temp. attrition	final attrition	survey term	semester
1st	CATI	17,910	0	0	winter 10/11	during 1st
2nd	CAWI	12,273	5,591	46	autumn 11	after 2nd
3rd	CATI	13,113	4,558	239	spring 12	after 3rd
4th	CAWI	11,202	6,424	284	autumn 12	after 4th
5th	CATI	12,694	4,620	596	spring/summer 13	during 6th
6th	CAWI	10,183	7,041	686	autumn 13	after 6th
7th	CATI	9,611	7,158	1,141	summer 14	during 8th
8th	CAWI	8,629	6,025	3,256	autumn 14	after 8th
9th	CATI	10,096	4,323	3,491	spring/summer 15	during 10th
10th	CATI	9,090	4,191	4,629	spring/summer 16	during 12th
11th	CAWI	7,020	5,042	5,848	autumn 16	after 12th
12th	CAWI	8,551	3,042	6,317	sprint-autumn 17	during 14th
13th	CATI	7,293	3,316	7,301	spring/summer 2018	during 16th
14th	CAWI	5,161	4,733	8,016	autumn 2018	after 16th

Table 8.6: Prior education, demographic and family variables and depended variable.

Name	Description	Values and notes	Waves
		Prior education	
rep_class	Number of repeated school classes	0 to 4	all waves
school_type	Type of secondary school attended	1 = upper secondary education, 0 = other	all waves
qualif_max	School-leaving qualification obtained	2 = general university entrence qualification, 1 = university of applied science qualification, 0 = other	all waves
grade_school	Overall grade on final school certificate	1.0 (excellent) to 5.0 (poor)	all waves
voctrain	Successfully completed vocational training before study	0 = no, 1 = yes	all waves
exam_german	German as examination subject for graduation	0 = no, 1 = yes	1
exam_adv_german	German as first examination subject	0 = no, 1 = yes	1
exam_maths	Mathematics as examination subject for graduation	0 = no, 1 = yes	1
exam_adv_maths	Mathematics as first examination subject	0 = no, 1 = yes	1
fail_prestudy	Number of dropouts from other trainings before study	0 = no, 1 = yes	1
math_prep	Math knowledge acquired before study	1 to 4 (very much)	2
ger_prep	German knowledge acquired before study	1 to 4 (very much)	2
skills_engl	English skills acquired before university	1 (few) to 4 (many)	2
skills_comp	Computer skills acquired before university	1 (few) to 4 (many)	2
preparation	General preparation for study	0 (poor) to 4 (perfect)	2
		Demographic variables and family	
genstat	Generation status in Germany	1.0 to 4.0 (no immigration background)	all waves
familylife	Living together with parents up to the age of 14	1 = with both parents, 0 else	all waves
birthyear	Birth year of the target person	1950 to 1994	all waves
gender	Gender of the target person	0 = female, 1 = male	all waves
mother_qualif	CASMIN: mother's highest general school leaving qualification	0 to 8 (highest)	1
father_qualif	CASMIN: father's highest general school leaving qualification	0 to 8 (highest)	1
mother_job	ISEI-08: mother's occupation	11.74 to 88.96	1
father_job	ISEI-08: father's occupation	11.56 to 88.96	1
statusp_father	Importance of father's status preservation	1 to 5 (very important)	1, 5, 10, 13
statusp_mother	Importance of mother's status preservation	1 to 5 (very important)	1, 5, 10, 13
religiosity	Religiosity	1 to 4 (very religious)	2, 14
religion	Member of a religious confession	0 = no, 1 = yes	2, 14
		Dependent variable	
status	Status (Y) of a student	0 = graduate, 1 = dropout, 2 = still studying	2, 3, ..., 14

Table 8.7: Higher education related variables.

Name	Description	Values and notes	Wave
Higher education related variables			
institofchoice	Studying at the institute of higher education of choice	0 = no, 1 = yes	1
insttype	Type of institution	1 = university of applied sciences, 2 = university	1
study_restrict	Study with admission restrictions or selection procedure?	0 = no, 1 = yes	1
fieldofchoice	Studying the field of first choice	0 = no, 1 = yes	1
study_alternative	Would you have started something else instead of a degree	0 = no, 1 = yes	1
study_informed	Informed about study	1 (poor) to 5 (good)	1, 5
probsuccess	Subjective probability of completing degree course	1 (low) to 5 (high)	1, 2, 4, 5, 6, 8
study_useful	Is the study useful for a good job	1 (no) to 5 (absolutely)	1, 5
helplessness	Study-specific helplessness	1 (low) to 5 (high)	2, 6
socint_instructors	Social integration: Acceptance by instructors	1 (low) to 4 (high)	2, 4, 6, 8
fear_failure	Fear of (academic) failure	1 (low) to 4 (high)	2, 4, 6, 8
study_importtance	Study important step for life goals	1 (low) to 4 (high)	2, 4, 6, 8
commit_necessary	Commitment to degree course: do no more than necessary	1 (does not apply) to 5 (applies completely)	2, 4, 6, 8
commit_enjoy	Enjoyment of degree program	1 (does not apply) to 5 (applies completely)	2, 4, 6, 8
commit_demands	High demands on self	1 (does not apply) to 5 (applies completely)	2, 4, 6, 8
commit_energy	Invest a lot of energy for academic success	1 (does not apply) to 5 (applies completely)	2, 4, 6, 8
commit_identification	Identification with degree program	1 (does not apply) to 5 (applies completely)	2, 4, 6, 8
learninggroup	Learning group participation	0 = no, 1 = yes	4, 8
courses_semester	Time spend for lectures during semester time	0 to 90 (hours)	all cawi
learn_semester	Time spend for private study during semester time	0 to 99 (hours)	all cawi
study_semester	Time spend for other study-oriented activities during semester	0 to 142 (hours)	all cawi
study_break	Time spend for study-oriented activities during semester break	0 to 99 (hours)	all cawi
gpa_cur	Current grade point average	1.0 (excellent) to 4.0 (poor)	all cawi
workload_match	Matching study workload to curriculum plan?	1 (much less) to 5 (much more)	all cawi
perform_better	Study performance better than expected	1 (worse) to 4 (better)	all cawi
perform_expect	Performance requirements fulfilled	1 (not at all) to 4 (absolutely)	all cawi
perform_satisfied	Satisfied with study performance	1 (not at all) to 4 (absolutely)	all cawi
field	First field of study	engineering, mathematics and natural sciences law, economics and social sciences, linguistics and cultural sciences, arts, medicine, other minor subject groups	all cati
change_uni	Have changed at least once the institution	0 = no, 1 = yes	all cati

Table 8.8: Variables describing the personal life of students.

Name	Description	Values and notes	Wave
	Personal life		
partic_people	Participation in university events aimed getting to know people	0 = no, 1 = yes	1
partic_orga	Participation in university events on study organization	0 = no, 1 = yes	1
partic_facil	Participation in university events with respect to central facilities	0 = no, 1 = yes	1
partic_acadskills	Participation in university events on academic skills	0 = no, 1 = yes	1
partic_course	Participation in university events on bridging courses	0 = no, 1 = yes	1
direct_costs	Difficulty to pay direct costs of higher education	1 (difficult) to 5 (easy)	1, 5
opportunity_costs	Limitation of the possibilities to earn own money	1 (not at all) to 5 (a lot)	1, 5
health	Health condition	1 (bad) to 5 (very good)	all cati
sports	Frequency of sport activities	1 (never) to 5 (daily)	13
smoking	Experience with smoking	1 (never) to 4 (daily)	3, 7, 10
student_assoc	Activity in student association	0 = no, 1 = yes	4
uni_sport	Activity at campus sport	0 = no, 1 = yes	4
study_peers	Number of friends at university	1 (nobody) to 7 (everyone)	4
socint_instructors	Social integration: acceptance by instructors	1 (poor) to 4 (good)	2, 4, 6, 8
socint_students	Social integration: contacts with students etc.	1 (poor) to 4 (good)	2, 4, 6, 8
bafoeg	Currently receive student financial aid (BAfoeG)	0 = no, 1 = yes	2
finance_family	Financial aid by family	0 to 8,000 (Euro)	2, 4, 6, 8
finance_get_by	Getting by with available money	1 (bad) to 5 (very good)	2, 4, 6, 8
job_semester	Time spent in a week for employment during semester time	0 to 60 hours	all cawi
job_break	Time spent in a week for employment during semester break	0 to 99 hours	all cawi

Table 8.9: Description of study related variables (second part) and interviewer variables.

Name	Description	Values and notes	Wave
	Satisfaction and personality		
satisf_enjoy	Study satisfaction: enjoy degree course	0 (no) to 10 (absolutely)	3, 5
satisf_conditions	Wishing better study conditions	0 (no) to 10 (absolutely)	3 , 5
satisf_match	Degree course and other obligations are hard to match	0 (no) to 10 (absolutely)	3, 5
satisf_whole	On the whole, satisfied with actual studies	0 (no) to 10 (absolutely)	3, 5
satisf_frustrating	External circumstances are frustrating	0 (no) to 10 (absolutely)	3, 5
satisf_kill	Degree course kills me	0 (no) to 10 (absolutely)	3, 5
satisf_interesting	Degree course is interesting	0 (no) to 10 (absolutely)	3, 5
satisf_concerns	Concerns of students are not taken into account	0 (no) to 10 (absolutely)	3, 5
satisf_tired	Often tired due to degree course	0 (no) to 10 (absolutely)	3, 5
big5_extraversion	Big five personality traits: Extraversion	1 (outgoing) to 5 (reserved)	3, 10
big5_agreeable	Agreeableness	1 (challenging) to 5 (friendly)	3, 10
big5_conscientious	Conscientiousness	1 (efficient) to 5 (easy going)	3, 10
big5_neuroticism	Neuroticism	1 (nervous) to 5 (confident)	3, 10
big5_openness	Openness to experience	1 (curious) to 5 (cautious)	3, 10
selfesteem	Global self-esteem	13 (low) to 50 (high)	3, 10
talent	Academic talent	1 (low) to 7 (high)	2, 6
easy_learning	Learning new things is ... for me	1 (hard) to 7 (easy)	2, 6
study_skills	Skills regarding study field	1 (low) to 7 (high)	2, 6
manage_tasks	Task management skills	1 (low) to 7 (high)	2, 6
persistence	Persistence	5 (low) to 25 (high)	4, 14
flexibility	Flexibility	5 (low) to 25 (high)	4, 14

9 General conclusion and summary

This thesis shed more light on students' higher education careers in Germany. The broad survey dataset used for all six empirical studies is the fifth starting cohort of freshmen students from winter-semester 2010/2011 of the National Educational Panel Survey (NEPS). The data gave new insights into students' educational pathways, including study fields, grades, but also "soft" determinants that are generally not available in administrative datasets, such as study satisfaction, academic integration, or their general effort for the study. Additionally, the NEPS covers determinants about early life in childhood, prior education and vocational training, and about students' life beyond the campus, for example, their hobbies, friends, family, financial situation, living conditions and more.

With the help of this large and in this extent unique dataset with data from German universities many variables have been found in a bivariate analysis that affect the dropout decision. These include many features that were already found relevant in the previous literature, as gender, parental background, secondary school grades, or study field. For other determinants, the effect on the dropout decision has rarely been investigated and, if so, predominantly in very small samples and with data from other countries where the education and financial aid system are different from those of Germany. These are mainly "soft" determinants belonging to the thematic field of motivation and satisfaction. Students who are satisfied and enjoy the degree course are less likely to drop out. Another new finding is that performance-related extrinsic motivation is stronger related to the binary dropout variable than intrinsic motivation.

Students' dropout from tertiary education is usually caused by a bundle of different reasons. The binary analysis gives just an overview of variables that might be important for multivariate models. The main aim of the thesis was to implement models that detect students at risk to drop out at an early stage or even before they start their higher education career. Tree-based models as the random forest and the AdaBoost algorithm have proven to be best suited for this classification task. Decision trees have the advantages

that they can handle missing data, are robust against outliers, are easily comprehensible, and can handle all common types of variables. Both tree-based methods also reveal better results in model performance compared to the naive Bayes, logistic regression, and the support vector machine. Unsurprisingly, the model gets more accurate if additional variables from later periods are included. Nevertheless, additional features from the study decision phase, which begins after secondary education or vocational training but before the start of university, only lead to a minor improvement of the model that only contains pre-study variables.

The most important variables in the multivariate model are largely in line with the important variables found in the bivariate analysis and important features that were already found by other studies. The overall grade at secondary education, the number of repeated classes at school, age, study preparation, overall satisfaction, and students' study commitment belong to the most important variables.

In a further article, different reasons for dropping out were analyzed. Students were grouped in eight clusters containing different types of dropout students. For the majority of students, different reasons were responsible for leaving the higher education system. The most frequently mentioned reasons are performance and interest related.

Since the models with a binary target variable (dropout vs. graduation) do not distinguish between good and poor students, a grade prediction was conducted. Two separate regression models include and exclude dropout students from the model, respectively. The two models reveal that many variables influence the dropout decision of students but not the grade if a student finally finished the first higher education degree. A rising age has even a positive influence on the grade, but a negative impact on the graduation probability. A potential reason for this phenomenon is that older students get less financial aid from their parents and are under higher financial pressure. They more often have an off-campus job and sometimes they even have their own family. This increases their opportunity costs of studying which may lead them to drop out.

In the final article of this thesis, I take advantage of the panel structure of the data to create a time-dependent dropout model. This updates the model when new data are available, approximately after each semester. Additionally, the models show which variables are important in each study phase. At the start of studies, demographic variables

and determinants of secondary education are of high relevance but they lose importance with progressing study duration.

My findings can mainly be used in practice by higher education institutions to implement an early warning system (EWS) for students. The EWS is supposed to help students at risk to drop out and suggests specific countermeasures that are tailored to the specific problems of the student. As the cluster analysis shows, many students realize that they are not interested in their subject and find that the study field has no practical relevance in the early stage of their study program. Therefore, three models for different stages of study have been fitted, where the first model uses only pre-study variables. This model can detect students at risk of dropping out even before they enroll for the study field and give a warning to the student and the institution that the selected subject might not be the right choice. Additional information events for students at secondary school, cooperations with secondary and tertiary education institutions (e.g. study test weeks) and test semesters where students can try courses of different study fields can also help to find the right field of study.

A change in the study field is also possible for students with lacking interest in their subject. In many cases, passed examinations can also be credited for other subjects.

Other students have performance problems in their study field. (Compulsory) bridging courses for students with poor secondary school grades, extra tutorials, and organized learning groups can help to close the gap to the high-performing students and lead to better academic integration. The numbers in the introduction exhibit that other highly developed countries as the UK and the US spend much more money on each student. Even if extra courses must be financed, Germany would probably still be below the costs of these two countries.

The financial aid system in Germany already belongs to the best in the world, which is shown by the low relevance of financial variables. As stated in the literature review in chapter 2, international studies find higher relevance of financial aid variables in countries offering less financial help for students. Closely related to the financial situation of students is the working status. Some students are forced to work even outside the semester-break.

The models also unfold which variables drive the dropout process in which stage of the study program. If the EWS detects a student at risk it can also suggest strategies suitable for the specific needs of a student. A poor grade point average at secondary

school might lead to performance problems, while uninformed students possibly lose interest in the field.

The models must be updated regularly with new data. Changes in the education system, as the Bologna process in 1999, can also lead to deviations from models that were trained with old data. Also, the actual COVID-19 pandemic will influence the models. They will perform worse with current data since the consequences of the pandemic are not included in the models. This crisis forced higher education institutions to turn classroom teaching into online events. Furthermore, it aggravates the financial situation of some students who have lost their off-campus jobs due to the crisis. Other economic crises can also influence enrollment and dropout rates. As described in the literature review, an increase in the unemployment rate decreases the opportunity costs for many students who do not find a job.

Institutions that will use an EWS in practice will probably use administrative data instead of survey data due to the easier availability. The problems of administrative data are that many "soft" determinants and background information of the students are not available. Instead, there will be more detailed information about earned credit points and grades for every single examination, which are strong predictors for dropout models. However, these variables are only available after the first examination phase, which is for some students already too late to intervene. Other, in the thesis widely discussed problems of panel surveys, are panel attrition and missing values. Panel attrition will occur in administrative data only if a student changes the institution and there is no data transfer between the two institutions. But higher education institutions (or single faculties) might be more interested in a dropout definition from a micro-perspective, meaning that university (or field) changes are defined as dropout. Contrary to that I constantly define dropout from a macro-perspective, so only leaving the complete higher education system without obtaining the targeted degree is considered as dropout.

In further research, an EWS must be tested before its exhaustive application in the German higher education system. The question of how students react to automated feedback from the model has not yet been sufficiently investigated.

Bibliography

Aarkrog, V., Wahlgren, B., Larsen, C. H., Mariager-Anderson, K., and Gottlieb, S. (2018). Decision-making processes among potential dropouts in vocational education and training and adult learning. *International Journal for Research in Vocational Education and Training (IJRVET)*, 5(2):111–129.

Abu-Oda, G. S. and El-Halees, A. M. (2015). Data mining in higher education: University student dropout case study. *International Journal of Data Mining & Knowledge Management Process*, 5(1):15.

Aggarwal, C. C. (2015). *Data Mining: The Textbook*, volume 1. Springer Science & Business Media.

Aina, C. (2013). Parental background and university dropout in Italy. *Higher Education*, 65(4):437–456.

Asendorpf, J. B., van de Schoot, R., Denissen, J. J. J., and Hutteman, R. (2014). Reducing bias due to systematic attrition in longitudinal studies: The benefits of multiple imputation. *International Journal of Behavioral Development*, 38(5):453–460.

Asif, R., Merceron, A., and Pathan, M. K. (2014). Predicting student academic performance at degree level: a case study. *International Journal of Intelligent Systems and Applications*, 7(1):49.

Assaad, R., Krafft, C., and Yassin, S. (2018). Comparing retrospective and panel data collection methods to assess labor market dynamics. *IZA Journal of Development and Migration*, 8(17).

Aulck, L., Velagapudi, N., Blumenstock, J., and West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*.

Baars, G. J., Stijnen, T., and Splinter, T. A. (2017). A model to predict student failure in the first year of the undergraduate medical curriculum. *Health Professions Education*, 3(1):5–14.

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7(3):112–118.

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1):3–17.

Baltagi, B. (2008). *Econometric analysis of panel data*. John Wiley & Sons.

Baradwaj, B. K. and Pal, S. (2011). Mining educational data to analyze students' performance. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 2(6):63–69.

Baraldi, A. N. and Enders, C. K. (2009). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37.

Batista, G. E. A. P. A. and Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6):519–533.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T., and Popelinsky, L. (2012). Predicting drop-out from social behaviour of students. *International Educational Data Mining Society*.

Bean, J. and Eaton, S. B. (2000). A psychological model of college student retention. In Braxton, J. M., editor, *Reworking the student departure puzzle*, pages 48–61. Nashville: Vanderbilt Univ. Press, 1 edition.

Bean, J. and Eaton, S. B. (2001). The psychology underlying successful retention practices. *Journal of College Student Retention: Research, Theory & Practice*, 3(1):73–89.

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in higher education*, 12(2):155–187.

Bean, J. P. (1985). Interaction effects based on class level in an explanatory model of college student dropout syndrome. *American educational research journal*, 22(1):35–64.

Bean, J. P. and Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of educational Research*, 55(4):485–540.

Beck, H. P. and Davidson, W. D. (2001). Establishing an early warning system: Predicting low grades in college students from survey of academic orientations scores. *Research in Higher education*, 42(6):709–723.

Becker, R. and Hecken, A. E. (2007). Studium oder berufsausbildung? eine empirische überprüfung der modelle zur erklärung von bildungsentscheidungen von esser sowie von breen und goldthorpe/university or vocational training? an empirical test of the rational choice model of educational choices suggested by esser as well as breen and goldthorpe. *Zeitschrift für Soziologie*, pages 100–117.

Beerkens, M., Mägi, E., and Lill, L. (2011). University studies as a side job: causes and consequences of massive student employment in estonia. *Higher Education*, 61(6):679–692.

Behr, A. (2006). Comparing estimation strategies for income equations in the presence of panel attrition. *Jahrbücher für Nationalökonomie und Statistik*, 226(4):361–384.

Behr, A., Bellgardt, E., and Rendtel, U. (2005). Extent and determinants of panel attrition in the european community household panel. *European Sociological Review*, 21(5):489–512.

Behr, A., Giese, M., Kamdjou, H. D. T., and Theune, K. (2020a). Dropping out of university: a literature review. *Review of Education*.

Behr, A., Giese, M., Teguim, H. D., and Theune, K. (2020b). Dropping out from higher education in Germany - an empirical evaluation of determinants for Bachelor students. *Open Education Studies*.

Behr, A., Giese, M., Theune, K., et al. (2020c). Early prediction of university dropouts–a random forest approach. *Jahrbücher für Nationalökonomie und Statistik*, 1(ahead-of-print).

Behr, A. and Theune, K. (2016). The causal effect of off-campus work on time to degree. *Education Economics*, 24(2):189–209.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Belloc, F., Maruotti, A., and Petrella, L. (2010). University drop-out: an Italian experience. *Higher Education*, 60(2):127–138.

Berens, J., Schneider, K., Görtz, S., Oster, S., and Burghoff, J. (2018). Early detection of students at risk – predicting student dropouts using administrative student data and machine learning methods. *CESifo Working Papers No. 7259*.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365.

Birkelbach, R., Vietgen, S., and Wallis, M. (2019). *DZHW-Studienberechtigtenpanel 2012*. Deutsches Zentrum für Hochschul-und Wissenschaftsforschung (DZHW), Hannover, Germany.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blossfeld, H.-P., Roßbach, H.-G., and von Maurice, J. (2011). Education as a Lifelong Process–The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft [Special Issue]*, 14.

Blüthmann, I., Lepa, S., Thiel, F., et al. (2012). Überfordert, enttäuscht, verwählt oder strategisch? eine typologie vorzeitig exmatrikulierter bachelorstudierender. *Zeitschrift fĆ1/4r PĆÂ¤ dagogik*, 58(1):89–108.

BMBF (2020). Studienerfolg und Studienabbruch. https://www.wihoforschung.de/de/studienerfolg-und-studienabbruch-3166.php (2020-11-10).

Bortz, J. and Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler. Lehrbuch mit Online-Materialien*. Springer-Lehrbuch. Springer Berlin Heidelberg.

Boulesteix, A.-L., Bender, A., Lorenzo Bermejo, J., and Strobl, C. (2011). Random forest gini importance favours snps with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics*, 13(3):292–304.

Bourdieu, P. (1977). *Outline of a Theory of Practice*. Cambridge University Press.

Brandstätter, H., Grillich, L., and Farthofer, A. (2006). Prognose des Studienabbruchs. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 38(3):121–131.

Brandstfätter, H. and Farthofer, A. (2003). Einfluss von erwerbstfätigkeit auf den studienerfolg. *Zeitschrift fü Arbeits- u. Organisationspsychologie*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. and Cutler, A. (2004). Random forests. `http://www.math.usu.edu/~adele/forests/cc_home.htm`.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cabrera, A. F., Castaneda, M. B., Nora, A., and Hengstler, D. (1992). The convergence between two theories of college persistence. *The journal of higher education*, 63(2):143–164.

Cabrera, A. F., Nora, A., and Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The journal of higher education*, 64(2):123–139.

Chen, J. J., Tsai, C., Moon, H., Ahn, H., Young, J. J., and Chen, C.-H. (2006). Decision threshold adjustment in class prediction. *SAR and QSAR in Environmental Research*, 17(3):337–352.

Chen, R. (2012). Institutional characteristics and college student dropout risks: A multilevel event history analysis. *Research in Higher Education*, 53(5):487–505.

Clerici, R., Giraldo, A., and Meggiolaro, S. (2014). The determinants of academic outcomes in a competing risks approach: evidence from italy. *Studies in Higher Education*, 40(9):1535–1549.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Culp, M., Johnson, K., and Michailidis, G. (2006). ada: The R package ada for stochastic boosting. *Journal of Statistical Software*, 17(2):1–27.

da Silva, T. L., Zakzanis, K., Henderson, J., and Ravindran, A. V. (2017). Predictors of post-secondary academic outcomes among local-born, immigrant, and international students in canada: A retrospective analysis. *Canadian Journal of Education*, 40(4):543–575.

Darlington, R. B. and Hayes, A. F. (2016). *Regression analysis and linear models: Concepts, applications, and implementation.* Guilford Publications.

Dekker, G., Pechenizkiy, M., and Vleeshouwers, J. (2009). Predicting students drop out: A case study. *Proceedings of the 2nd International Conference on Educational Data Mining*, pages 41–50.

DESTATIS (2018). Gesellschaft & staat - bildung, forschung, kultur. https://www.destatis.de (03.03.2018).

DESTATIS (2019). Bildung und kultur - prüfungen an hochschulen. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Publikationen/Downloads-Hochschulen/pruefungen-hochschulen-2110420187004.pdf?__blob=publicationFile (08.01.2020).

Di Pietro, G. (2006). Regional labour market conditions and university dropout rates: Evidence from italy. *Regional Studies*, 40(6):617–630.

Di Pietro, G. and Cutillo, A. (2008). Degree flexibility and university drop-out: The Italian experience. *Economics of Education Review*, 27(5):546–555.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2008). Misc functions of the department of statistics (e1071), TU Wien. *R package.*

Dinov, I. D. (2018). *Data Science and Predictive Analytics: Biomedical and Health Applications using R.* Springer.

Durkheim, E. (1951). *Suicide : a study in sociology.* Glencoe: The Free Press.

Dustmann, C. (2004). Parental background, secondary school track choice, and wages. *Oxford Economic Papers*, 56(2):209–230.

Elkan, C. (2001). The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978.

Ethington, C. A. (1990). A psychological model of student persistence. *Research in higher Education*, 31(3):279–293.

Fahrmeir, L. and Tutz, G. (2013). *Multivariate statistical modelling based on generalized linear models.* Springer Science & Business Media.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Garciarena, U. and Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, 89:52–65.

Georg, W. (2009). Individuelle und institutionelle Faktoren der Bereitschaft zum Studienabbruch: eine Mehrebenenanalyse mit Daten des Konstanzer Studierendensurveys. *Zeitschrift für Soziologie der Erziehung und Sozialisation*, 28(2):191–206.

Ghignoni, E. (2017). Family background and university dropouts during the crisis: the case of Italy. *Higher Education*, 73(1):127–151.

Glaesser, J. (2006). Dropping out of further education: a fresh start? findings from a german longitudinal study. *Journal of Vocational Education and Training*, 58(1):83–97.

Glocker, D. (2011). The effect of student aid on the duration of study. *Economics of Education Review*, 30(1):177–190.

Gold, A. (1988). *Studienabbruch, Abbruchneigung und Studienerfolg: Vergleichende Bedingungsanalysen des Studienverlaufs*. Frankfurt a.M.: Peter Lang Verlag.

Griesbach, H., Lewin, K., Heublein, U., and Sommer, D. (1998). Studienabbruch – Typologie und Möglichkeiten der Abbruchquotenbestimmung. *Kurzinformation A5/98. Hannover: HIS*.

Gury, N. (2011). Dropping out of higher education in france: a micro-economic approach using survival analysis. *Education Economics*, 19(1):51–64.

Haas, C. and Hadjar, A. (2019). Students' trajectories through higher education: a review of quantitative research. *Higher Education*, pages 1–20.

Hadjar, A. and Becker, R. (2004). Warum einige studierende ihr soziologie-studium abbrechen wollen. studienwahlmotive, informationsdefizite und wahrgenommene berufsaussichten als determinanten der abbruchneigung. *Soziologie-Forum der Deutschen Gesellschaft für Soziologie*, 33(3):47–65.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier, 3 edition.

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186.

Handl, A. and Kuhlenkasper, T. (2017). *Multivariate Analysemethoden: Theorie und Praxis mit R*. Springer-Verlag.

Hansen, M. N. and Mastekaasa, A. (2006). Social origins and academic performance at university. *European Sociological Review*, 22(3):277–291.

Hapfelmeier, A., Hothorn, T., Ulm, K., and Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing*, 24(1):21–34.

Härterich, J., Dehling, H., Glasmachers, E., Griese, B., and Kallweit, M. (2014). MP2-Mathe/Plus/Praxis: Strategien zur Vorbeugung gegen Studienabbruch. *Zeitschrift für Hochschulentwicklung*, 9(4):39–56.

Hartung, J., Elpelt, B., and Klösener, K. (2009). *Statistik: Lehr- und Handbuch der angewandten Statistik ; [mit zahlreichen durchgerechneten Beispielen]*. Oldenbourg.

Hartung, J., Knapp, G., and Sinha, B. K. (2011). *Statistical meta-analysis with applications*, volume 738. John Wiley & Sons.

Hasan, M. M. and Dunn, P. K. (2011). Two tweedie distributions that are near-optimal for modelling monthly rainfall in australia. *International Journal of Climatology*, 31(9):1389–1397.

Hastie, T. and Qian, J. (2014). Glmnet vignette. *Retrieve from http://www. web. stanford. edu/˜ hastie/Papers/Glmnet_Vignette. pdf. Accessed March 2020*, 20.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. pages 475–492.

Hetze, P. (2011). *Nachhaltige Hochschulstrategien für mehr MINT-Absolventen*. Stifterverband für die Deutsche Wissenschaft.

Heublein, U. (2014a). *Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen: Berechnungen auf der Basis des Absolventenjahrgangs 2012*. Dt. Zentrum für Hochsch.-und Wiss.-Forschung.

Heublein, U. (2014b). Student drop-out from German higher education institutions. *European Journal of Education*, 49(4):497–513.

Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J., and Woisch, A. (2017). Zwischen Studienerwartungen und Studienwirklichkeit. Forum Hochschule 1/2017.

Heublein, U., Hutzsch, C., Schreiber, J., Sommer, D., and Besuch, G. (2010). Ursachen des Studienabbruchs in Bachelor-und in herkömmlichen Studiengängen - Ergebnisse einer bundesweiten Befragung von Exmatrikulierten des Studienjahres 2007/08. HIS: Forum Hochschule 2/2010.

Heublein, U., Richter, J., Schmelzer, R., and Sommer, D. (2014). Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen: Statistische Berechnungen auf der Basis des Absolventenjahrgangs 2012. Forum Hochschule 4/2014.

Heublein, U., Schmelzer, R., Sommer, D., and Wank, J. (2008). Die entwicklung der studienabbruch-und schwundquotenquoten an den deutschen universitäten und fachhochschulen. *HIS-Projektbericht. Hannover*.

Heublein, U., Schmelzer, R., Sommer, D., and Wank, J. (2012). Die entwicklung der schwund-und studienabbruchquoten an den deutschen hochschulen. In *HIS: Forum Hochschule*, volume 3, page 2012.

Heublein, U., Wolter, A., et al. (2011). Studienabbruch in deutschland. definition, häufigkeit, ursachen, maßnahmen. *Zeitschrift für Pädagogik*, 57(2):214–236.

Hoffait, A.-S. and Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, 101:1–11.

Horstschräer, J. and Sprietsma, M. (2015). The effects of the introduction of bachelor degrees on college enrollment and dropout rates. *Education Economics*, 23(3):296–317.

Hothorn, T., Hornik, K., Strobl, C., and Zeileis, A. (2018). Package 'party': A laboratory for recursive partytioning. *Package Reference Manual for Party Version 1.3-0*, 16.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hovdhaugen, E. (2015). Working while studying: the impact of term-time employment on dropout rates. *Journal of Education and Work*, 28(6):631–651.

Hovdhaugen, E. and Aamodt, P. O. (2009). Learning environment: Relevant or not to students' decision to leave university? *Quality in Higher Education*, 15(2):177–189.

HRK (2019). *German higher education system. German Rectors' Conference.* www.hrk.de/fileadmin/_migrated/content_uploads/GERMAN_HIGHER_ EDUCATION_SYSTEM.pdf (07.05.2020).

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). A practical guide to support vector classification.

in Zahlen, D. (2020). Öffentliche Bildungsausgaben in Prozent des BIP. https://www.deutschlandinzahlen.de/tab/deutschland/bildung/bildungsausgaben/oeffentliche-bildungsausgaben-in-prozent-des-bip (02.10.2020).

Isphording, I. and Wozny, F. (2018). Ursachen des Studienabbruchs – eine Analyse des Nationalen Bildungspanels. *IZA Research Report No. 82.*

Jadrić, M., Garača, Ž., and Čukušić, M. (2010). Student dropout analysis with application of data mining methods. *Management: Journal of Contemporary Management Issues*, 15(1):31–46.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning.* Springer.

Janitza, S., Strobl, C., and Boulesteix, A.-L. (2013). An auc-based permutation variable importance measure for random forests. *BMC bioinformatics*, 14(1):119.

Jeatrakul, P., Wong, K. W., and Fung, C. C. (2010). Classification of imbalanced data by combining the complementary neural network and smote algorithm. In *International Conference on Neural Information Processing*, pages 152–159. Springer.

Johnes, G. and McNabb, R. (2004). Never give up on the good times: student attrition in the uk. *Oxford Bulletin of Economics and Statistics*, 66(1):23–47.

Johnes, J. (1990). Determinants of student wastage in higher education. *Studies in Higher Education*, 15(1):87–99.

Johnes, J. and Taylor, J. (1989). Undergraduate non-completion rates: differences between UK universities. *Higher Education*, 18(2):209–225.

Jørgensen, B. and Paes De Souza, M. C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93.

Kemper, L., Vorhoff, G., and Wigger, B. U. (2019). Predicting student dropout: a machine learning approach. *European Journal of Higher Education, forthcoming*.

Kleinke, K., Reinecke, J., Salfrán, D., and Spiess, M. (2020). *Applied Multiple Imputation: Advantages, Pitfalls, New Developments and Applications in R*. Springer Nature.

Knoke, M. (2018). Übergänge gestalten - Studienerfolg verbessern. German Rectors' Conference. `https://www.hrk-nexus.de/projekt-nexus/aufgaben-und-ziele/`.

Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in wisconsin. *JEDM-Journal of Educational Data Mining*, 7(3):18–67.

Köhler, C., Pohl, S., and Carstensen, C. H. (2015). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling*, 57(4):499.

Korhonen, V. and Rautopuro, J. (2018). Identifying problematic study progression and "at-risk" students in higher education in Finland. *Scandinavian Journal of Educational Research*, 63(7):1056–1069.

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data. *Proceedings of Informing Science & IT Education Conference*, pages 647–665.

Krause, A. and Schüller, S. (2014). Evidence and persistence of education inequality in an early-tracking system: The german case.

Krawczyk, B., Woźniak, M., and Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14:554–562.

Kriegel, H.-P., Kröger, P., and Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1–57.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):10.

Larsen, M. R., Sommersel, H. B., and Larsen, M. S. (2013a). *Evidence on Dropout Phenomena at Universities*. Danish Clearinghouse for educational research.

Larsen, M. S., Kornbeck, K. P., Kristensen, R., Larsen, M. R., and Sommersel, H. B. (2013b). Dropout phenomena at universities: What is Dropout? Why does Dropout Occur? What Can be Done by the Universities to Prevent or Reduce it? Danish Clearinghouse for Educational Research - Research Series 15.

Larsen, M. S., Kornbeck, K. P., Kristensen, R., Larsen, M. R., and Sommersel, H. B. (2013c). Dropout phenomena at universities: What is dropout? why does dropout occur? what can be done by the universities to prevent or reduce it? Technical report, Danish Clearinghouse for educational research.

Lassibille, G. and Gómez, M. L. N. (2009). Tracking students' progress through the spanish university school sector. *Higher Education*, 58(6):821–839.

Lassibille, G. and Navarro Gómez, L. (2008). Why do higher education students drop out? evidence from Spain. *Education Economics*, 16(1):89–105.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3).

LIfBi (2017). Startkohorte 5: Studierende (SC5) - Studienübersicht Wellen 1 bis 9. Technical report, Leibniz Institut für Bildungsverläufe e.V.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Luan, J. (2002). Data mining and its applications in higher education. *New directions for institutional research*, 2002(113):17–36.

Majka, M. (2017). naivebayes: High performance implementation of the naive bayes algorithm. *R package*.

Mäkinen, J., Olkinuora, E., and Lonka, K. (2004). Students at risk: Students' general study orientations and abandoning/prolonging the course of studies. *Higher education*, 48(2):173–188.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., and Ziegler, A. (2012). Probability machines: Consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, 51(01):74–81.

Manhães, L. M. B., da Cruz, S. M. S., and Zimbrão, G. (2014). Evaluating performance and dropouts of undergraduates using educational data mining. In *Proceedings of the Twenty-Ninth Symposium on Applied Computing*.

Maron, M. E. (1961). Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 8(3):404–417.

Mastekaasa, A. and Smeby, J.-C. (2008). Educational choice and persistence in male-and female-dominated fields. *Higher Education*, 55(2):189–202.

Mayer, K. U., Müller, W., and Pollak, R. (2007). Germany: Institutional change and inequalities of access in higher education. *Stratification in higher education: A comparative study*, pages 241–265.

Mayra, A. and Mauricio, D. (2018). Factors to predict dropout at the universities: A case of study in ecuador. In *Global Engineering Education Conference (EDUCON), 2018 IEEE*, pages 1238–1242. IEEE.

Meeyai, S. (2016). Logistic regression with missing data: A comparison of handling methods, and effects of percent missing values. *Journal of Traffic and Logistics Engineering Vol*, 4(2).

Meggiolaro, S., Giraldo, A., and Clerici, R. (2015). A multilevel competing risks model for analysis of university students' careers in italy. *Studies in Higher Education*, 42(7):1259–1274.

Metzner, B. S. and Bean, J. P. (1987). The estimation of a conceptual model of nontraditional undergraduate student attrition. *Research in higher education*, 27(1):15–38.

Montmarquette, C., Mahseredjian, S., and Houle, R. (2001). The determinants of university dropouts: a bivariate probability model with sample selection. *Economics of Education Review*, 20(5):475–484.

Mora, T. (2008). Why higher graduated regret their field of studies? some evidence from catalonia. In *XV Encuentro de Economía Pública: políticas públicas y migración*, page 41.

Mortagy, Y., Boghikian-Whitby, S., and Helou, I. (2018). An analytical investigation of the characteristics of the dropout students in higher education. *Issues in Informing Science and Information Technology*, 15:249–278.

Müller, L. and Braun, E. (2018). Student Engagement - Ein Konzept für ein evidenzbasiertes Qualitätsmanagement an Hochschulen. *Zeitschrift für Erziehungswissenschaft*, 21(3):649–670.

Müller, S. and Schneider, T. (2013). Educational pathways and dropout from higher education in germany. *Longitudinal and Life Course Studies*, 4(3):218–241.

Neyt, B., Omey, E., Verhaest, D., and Baert, S. (2019). Does student work really affect educational outcomes? a review of the literature. *Journal of Economic Surveys*, 33(3):896–921.

Nordmann, E., Calder, C., Bishop, P., Irwin, A., and Comber, D. (2019). Turn up, tune in, don't drop out: The relationship between lecture attendance, use of lecture recordings, and achievement at different levels of study. *Higher Education*, 77(6):1065–1084.

OECD (2017). Education and training. http://stats.oecd.org/ (01.09.2017).

OECD (2020). Education and training. http://stats.oecd.org/ (07.09.2020).

Ozdemir, S. (2016). *Principles of Data Science*. Packt Publishing Ltd.

Pascarella, E. T. (1980). Student-faculty informal contact and college outcomes. *Review of educational research*, 50(4):545–595.

Pascarella, E. T. and Terenzini, P. T. (2005). *How college affects students*, volume 2. Jossey-Bass San Francisco, CA.

Pelánek, R. (2015). Metrics for evaluation of student models. *Journal of Educational Data Mining*, 7(2):1–19.

Plackett, R. L. (1983). Karl pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique*, 51:59–72.

Pochiraju, B. and Seshadri, S. (2018). *Essentials of Business Analytics: An Introduction to the Methodology and its Applications*, volume 264. Springer.

Prussog-Wagner, A., Weiß, T., Aust, F., and Turri, F. (2016). Methodenbericht: NEPS-Startkohorte 5 – CATI-Haupterhebung Sommer 2016 B112. Technical report, Leibniz Institut für Bildungsverläufe e.V.

Quadri, M. M. and Kalyankar, N. (2010). Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, 10(2).

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramaswami, M. and Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. *arXiv preprint arXiv:0912.3924*.

Reay, D., David, M., and Ball, S. (2001). Making a difference? institutional habituses and higher education choice. *Sociological Research Online*, 5(4).

Reinalda, B. and Kulesza-Mietkowski, E. (2005). *The Bologna process: Harmonizing Europe's higher education*. Barbara Budrich Farmington Hills, MI.

Reisel, L. and Brekke, I. (2009). Minority dropout in higher education: A comparison of the united states and norway using competing risk event history analysis. *European Sociological Review*, page jcp045.

Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.

Rios, G. et al. (2013). Predicting early students with high risk to drop out of university using a neural network-based approach. In *ICCGI 2013, The Eighth International Multi-Conference on Computing in the Global Information Technology*.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121.

Rodríguez-Gómez, D., Feixas, M., Gairín, J., and Muñoz, J. L. (2015). Understanding catalan university dropout from a cross-national approach. *Studies in Higher Education*, 40(4):690–703.

Rodriguez-Muñiz, L. J., Bernardo, A. B., Esteban, M., and Diaz, I. (2019). Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? *PLoS ONE*, 14(6):1–20.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Rovira, S., Puertas, E., and Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PloS one*, 12(2):e0171207.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94.

Saar-Tsechansky, M. and Provost, F. (2007). Handling missing values when applying classification models. *Journal of machine learning research*, 8(Jul):1623–1657.

Saarela, M. and Kärkkäinen, T. (2015). Analysing student performance using sparse data of core bachelor courses. *JEDM-Journal of Educational Data Mining*, 7(1):3–32.

Sales, A., Balby, L., and Cajueiro, A. (2017). Exploiting academic records for predicting student drop out: a case study in brazilian higher education. *Journal of Information and Data Management*, 7(2):166.

Sarcletti, A. and Müller, S. (2011). Zum Stand der Studienabbruchforschung. Theoretische Perspektiven, zentrale Ergebnisse und methodische Anforderungen an künftige Studien. *Zeitschrift für Bildungsforschung*, 1(3):235–248.

Schiefele, U., Streblow, L., and Brinkmann, J. (2007). Aussteigen oder Durchhalten. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 39(3):127–140.

Schnepf, S. V. (2003). Inequalities in secondary school attendance in germany.

Schnepf, S. V. (2014). Do tertiary dropout students really not succeed in european labour markets? *IZA Discussion Paper No. 8015*.

Schröder-Gronostay, M. (1999). Studienabbruch–zusammenfassung des forschungsstandes. *Studienerfolg und Studienabbruch*, pages 209–240.

Seaman, S. R., Farewell, D., and White, I. R. (2016). Linear increments with non-monotone missing data and measurement error. *Scandinavian Journal of Statistics*, 43(4):996–1018.

Sellar, S. and Lingard, B. (2014). The OECD and the expansion of PISA: New global modes of governance in education. *British Educational Research Journal*, 40(6):917–936.

Severiens, S. and Ten Dam, G. (2012). Leaving college: A gender comparison in male and female-dominated programs. *Research in Higher Education*, 53(4):453–470.

Shannaq, B., Rafael, Y., and Alexandro, V. (2010). Student relationship in higher education using data mining techniques. *Global Journal of Computer Science and Technology*, 10(11).

Sherman, J. (1979). Predicting mathematics performance in high school girls and boys. *Journal of Educational Psychology*, 71(2):242.

Shono, H. (2008). Application of the tweedie distribution to zero-catch data in cpue analysis. *Fisheries Research*, 93(1-2):154–162.

Sill, J., Takács, G., Mackey, L., and Lin, D. (2009). Feature-weighted linear stacking. *arXiv preprint arXiv:0911.0460*.

Singell, L. D. and Waddell, G. R. (2010). Modeling retention at a large public university: Can at-risk students be identified early enough to treat? *Research in Higher Education*, 51(6):546–572.

Singer, J. (2019). Shaping the arc of educational research. Hedges lecture at the SREE Spring 2019 Conference, Washington, D.C. https://www.sree.org/2019-video.

Siri, A. (2015). Predicting students' dropout at university using artificial neural networks. *Italian Journal of Sociology of Education*, 7(2):225–247.

Smith, J. P. and Naylor, R. A. (2001). Dropping out of university: a statistical analysis of the probability of withdrawal for UK university students. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(2):389–405.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1):64–85.

Spady, W. G. (1971). Dropouts from higher education: Toward an empirical model. *Interchange*, 2(3):38–62.

Spangenberg, H. and Quast, H. (2016). Bildungsentscheidungen und Umorientierungen im nachschulischen Verlauf - Dritte Befragung der Studienberechtigten 2010 viereinhalb Jahre nach Schulabschluss. *Forum Hochschule 5/2016, Hannover: DZHW.*

Statistisches Bundesamt (2011). Bildung und kultur - studierende an hochschulen wintersemester 2010/2011 fachserie 11 reihe 4.1. Technical report, Statistisches Bundesamt.

Stinebrickner, R. and Stinebrickner, T. (2014). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*, 32(3):601–644.

Stinebrickner, T. R. and Stinebrickner, R. (2008). The effect of credit constraints on the college drop-out decision a direct approach using a new panel study. *American Economic Review*, 98(5):2163–2184.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., and Abreu, R. (2015). A comparative study of classification and regression algorithms for modelling students' academic performance. *International Educational Data Mining Society.*

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307).

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25).

Suhlmann, M., Sassenberg, K., Nagengast, B., and Trautwein, U. (2018). Belonging mediates effects of student-university fit on well-being, motivation, and dropout intention. *Social Psychology*.

Suhre, C. J., Jasen, E. P., and Harskamp, E. G. (2007). Impact of degree program satisfaction on the persistence of college students. *Higher Education*, 54(2):207–226.

Superby, J.-F., Vandamme, J., and Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Workshop on Educational Data Mining*, volume 32, page 234.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958.

Tan, P.-N., Steinbach, M., and Kumar, V. (2007). *Introduction to data mining*. Pearson Education India.

Theune, K. (2015). The working status of students and time to degree at german universities. *Higher Education*, 70(4):725–752.

Thomas, L. (2002). Student retention in higher education: the role of institutional habitus. *Journal of Education Policy*, 17(4):423–442.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125.

Tinto, V. (1988). Stages of student departure: Reflections on the longitudinal character of student leaving. *The Journal of Higher Education*, 59(4):438–455.

Tinto, V. (1993). *Leaving College: Rethinking the causes and cures of student attrition*. Chicago: Chicago University Press, 2 edition.

Torgo, L. (2010). *Data Mining with R, learning with case studies*. Chapman and Hall/CRC.

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 23(5):373–405.

Tweedie, M. C. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, volume 579, pages 579–604.

Ulriksen, L., Madsen, L. M., and Holmegaard, H. T. (2010). What do we know about explanations for drop out/opt out among young people from STM higher education programmes? *Studies in Science Education*, 46(2):209–244.

Van Bragt, C. A., Bakx, A. W., Bergen, T. C., and Croon, M. A. (2011a). Looking for students personal characteristics predicting study outcome. *Higher Education*, 61(1):59–75.

Van Bragt, C. A., Bakx, A. W., Teune, P. J., Bergen, T. C., and Croon, M. A. (2011b). Why students withdraw or continue their educational careers: a closer look at differences in study approaches and personal reasons. *Journal of Vocational Education and Training*, 63(2):217–233.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

van Gennep, A. (1960). *The Rites of Passage*. Chicago: The University of Chicago Press.

Vandamme, J.-P., Meskens, N., and Superby, J.-F. (2007). Predicting academic performance by data mining methods. *Education Economics*, 15(4):405–419.

Vandecasteele, L. and Debels, A. (2007). Attrition in panel data: the effectiveness of weighting. *European Sociological Review*, 23(1):81–97.

Velasco, M. S. et al. (2012). More than just good grades: candidates' perceptions about the skills and attributes employers seek in new graduates. *Journal of Business Economics and Management*, 13(3):499–517.

Vink, G., Frank, L. E., Pannekoek, J., and Van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90.

Voelkle, M. C. and Sander, N. (2008). University dropout: A structural equation approach to discrete-time survival analysis. *Journal of Individual Differences*, 29(3):134–147.

Vogler-Ludwig, K., Düll, N., and Kriechel, B. (2016). Arbeitsmarkt 2030 - wirtschaft und arbeitsmarkt im digitalen zeitalter. W. Bertelsmann Verlag.

Vossensteyn, H., Stensaker, B., Kottmann, A., Hovdhaugen, E., Jongbloed, B., Wollscheid, S., Kaiser, F., and Cremonini, L. (2015). Dropout and completion in higher education in Europe. Luxembourg: Publications Office of the European Union.

Watermann, R., Daniel, A., and Maaz, K. (2014). Primäre und sekundäre Disparitäten des Hochschulzugangs: Erklärungsmodelle, Datengrundlagen und Entwicklungen. *Zeitschrift für Erziehungswissenschaft*, 17(2):233–261.

Weerasinghe, I. S., Lalitha, R., and Fernando, S. (2017). Students' satisfaction in higher education literature review. *American Journal of Educational Research*, 5(5):533–539.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Würbach, A. (2020). Samples, weights, and nonresponse: Neps starting cohort 5 — first-year students - from higher education to the labor market (wave 14). Technical report, Bamberg, Germany: Leibniz Instiute for Educational Trajectories, National Educational Panel Study.

Yathongchai, W., Yathongchai, C., Kerdprasop, K., and Kerdprasop, N. (2012). Factor analysis with data mining technique in higher educational student drop out. *Latest Advances in Educational Technologies*, pages 111–116.

Yukselturk, E., Ozekes, S., and Türel, Y. K. (2014). Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and E-learning*, 17(1):118–133.

Zinn, S. (2019). Samples, weights, and nonresponse: Neps starting cohort 5 — first-year students - from higher education to the labor market (wave 12). Technical report,

Bamberg, Germany: Leibniz Instiute for Educational Trajectories, National Educational Panel Study.

Zinn, S., Steinhauer, H. W., and Aßmann, C. (2017). Samples, Weights, and Nonresponse: the Student Sample of the National Educational Panel Study (Wave 1 to 8) (NEPS Survey Paper No. 18). Technical report, Leibniz Institut für Bildungsverläufe e.V.

10 Attachments

Überblick über die Einzelbeiträge der vorliegenden Dissertation:

- Andreas Behr, Marco Giese, Hervé D. Teguim K., Katja Theune (2020): "Dropping out of university: a literature review". Diese Studie wurde veröffentlicht im *Review of Education*, 8(2), 614-652.

- Andreas Behr, Marco Giese, Hervé D. Teguim K., Katja Theune (2020): "Dropping out from higher education in Germany - an empirical evaluation of determinants for Bachelor students". Diese Studie wurde veröffentlicht in *Open Education Studies*, 2(1), 126-148.

- Andreas Behr, Marco Giese, Hervé D. Teguim K., Katja Theune (2020): "Early prediction of university dropouts - a random forest approach". Diese Studie wurde veröffentlicht in *Jahrbücher für Nationalökonomie und Statistik* (Druck folgt in Kürze).

- Andreas Behr, Marco Giese, Hervé D. Teguim K., Katja Theune: "Predicting Dropout from Higher Education - A Comparison of Machine Learning Algorithms". Diese Studie wurde bereits mehrfach eingereicht, aktuell eingereicht in *Education and Information Technologies*, und basierend auf den Vorschlägen vorheriger Gutachter überarbeitet.

- Andreas Behr, Marco Giese, Hervé D. Teguim K., Katja Theune (2020): "Motives for dropping out from higher education - An analysis for Bachelor students in Germany". Diese Studie wurde veröffentlicht in *European Journal of Education* (Druck folgt in Kürze).

- Marco Giese (2020): "Predicting Higher Education Grades using Strategies Correcting for Panel Attrition". Diese Studie wurde veröffentlicht in *Open Education Studies*, 2(1), 180-201.

- Marco Giese: "Prediction of Time Dependent Dropout and Graduation Rates in Higher Education under the Presence of Panel Attrition". Das Abstract dieser Studie wurde eingereicht und akzeptiert für das DZHW Sonderband "Survey Methoden in der Hochschulforschung".

Marco Giese
Tiegelstr. 25
45141 Essen

Ich gebe folgende eidesstattliche Erklärung ab zu meiner Dissertation mit dem Titel:
“An analysis of dropout students in the German higher education system using modern data mining techniques”

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig ohne unzulässige Hilfe Dritter verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen unter der Angabe der Quelle als solche gekennzeichnet habe.

Die Grundsätze für die Sicherung guter wissenschaftlicher Praxis an der Universität Duisburg-Essen sind beachtet worden.

Ich habe die Arbeit keiner anderen Stelle zu Prüfungszwecken vorgelegt

Ort, Datum	Marco Giese

Erklärung zur Koautorenschaft

Grundsätzlich hat sich jeder der vier Autoren mit den gesamten gemeinsamen Artikeln aus Kapitel 2, 3, 4, 5 und 6 auseinandergesetzt und einzelne Stellen überarbeitet. Die hauptverantwortlichen für die jeweiligen Abschnitte werden hier detailliert beschrieben.

Kapitel 2 der vorliegenden Dissertation - "Dropping out of university: a literature review" - ist eine gemeinsame Forschungsarbeit mit Prof. Dr. Andreas Behr, Dr. Katja Theune und Hervé Donald Teguim Kamdjou.

Mein eigener hauptverantwortlicher Beitrag betrifft die Einleitung (Kapitel 2.1), Literaturrecherche mit dem Fokus auf internationale Studien, und Ausformulierung der Ergebnisse in Kapitel 2.3 und 2.4 (mit dem Fokus auf der internationalen Literatur). Herrn Teguims Beitrag lag im Erstellen der Abbildung 2.1 und Kapitel 2.3.1 zur Erklärung der Selektion der Papiere. Zusätzlich erstellte er die Übersichtstabelle 2.3. Mit dem Fokus auf Deutsche Literatur war er maßgeblich an den Kapiteln 2.3 und 2.4 beteiligt. Dr. Theunes Beitrag lag im Kapitel 2.2 zum theoretischen Hintergrund und der Zusammenfassung in Kapitel 2.5. Sie war außerdem als Projektleiterin verantwortlich für die inhaltliche Überarbeitung. Prof. Dr. Behrs Beitrag lag in Vorschlägen und Hinweisen zur methodischen Herangehensweise, und einer detaillierten sprachlichen und formellen Überarbeitung.

Kapitel 3 der vorliegenden Dissertation - "Dropping out from higher education in Germany - an empirical evaluation of determinants for Bachelor students" - ist eine gemeinsame Forschungsarbeit mit Prof. Dr. Andreas Behr, Dr. Katja Theune und Hervé Donald Teguim Kamdjou.

Mein eigener hauptverantwortlicher Beitrag betrifft die Einleitung (Kapitel 3.1), sowie die Programmierung und Ausformulierung der bivariaten Analyse (Kapitel 3.4). Herrn Teguims Beitrag war die Datenbeschreibung (Kapitel 3.3) und die Programmierung und Ausformulierung des multivariaten Modells, sowie dessen Ergebnisformulierung (Kapitel 3.5 und 3.6). Dr. Theunes hauptverantwortlicher Beitrag lag in der Zusammenfassung der Literatur und Einordnung in den Forschungskontext (Kapitel 3.2), sowie der Zusammenfassung der Ergebnisse (Kapitel 3.7). Sie war außerdem als Projektleiterin verantwortlich für die inhaltliche Überarbeitung. Prof. Dr. Behrs Beitrag lag in Vorschlägen und Hinweisen zur methodischen Herangehensweise, und einer detaillierten sprachlichen

und formellen Überarbeitung.

Kapitel 4 der vorliegenden Dissertation - "Early prediction of university dropouts - a random forest approach" - ist eine gemeinsame Forschungsarbeit mit Prof. Dr. Andreas Behr, Dr. Katja Theune und Hervé Donald Teguim Kamdjou.

Mein eigener hauptverantwortlicher Beitrag betrifft die Beschreibung des Datensatzes und der ausgewählten Variablen (Kapitel 4.3). Außerdem war ich für die Methodenbeschreibung verantwortlich (Kapitel 4.4), sowie der Argumentation, warum unser Modell nicht stark von Panel Attrition betroffen ist (Unterkapitel 4.5.5). Herrn Teguims Beitrag lag in der Programmierung und Beschreibung der Ergebnisse (Kapitel 4.5) und in der Ausformulierung der Einleitung (Kapitel 4.1). Dr. Theune war maßgeblich verantwortlich für die Literaturübersicht (Kapitel 4.2) und die Diskussion und Zusammenfassung der Ergebnisse (Kapitel 4.6). Sie war außerdem als Projektleiterin verantwortlich für die inhaltliche Überarbeitung. Prof. Dr. Behrs Beitrag lag in Vorschlägen und Hinweisen zur methodischen Herangehensweise, und einer detaillierten sprachlichen und formellen Überarbeitung.

Kapitel 5 der vorliegenden Dissertation - "Predicting Dropout from Higher Education - A Comparison of Machine Learning Algorithms" - ist eine gemeinsame Forschungsarbeit mit Prof. Dr. Andreas Behr, Dr. Katja Theune und Hervé Donald Teguim Kamdjou.

Mein eigener hauptverantwortlicher Beitrag betrifft die Literaturbeschreibung und Einordnung in den aktuellen Forschungsstand der Arbeit (Kapitel 5.2), sowie die Programmierung, Beschreibung, Interpretation und Diskussion der Ergebnisse (Kapitel 5.5). Herrn Teguims Beitrag lag in der Datenbeschreibung (Kapitel 5.3) und der Beschreibung der verwendeten Modelle (Kapitel 5.4). Dr. Theunes Beitrag lag in der Einleitung (Kapitel 5.1) und in der Zusammenfassung der Ergebnisse (Kapitel 5.6). Sie war außerdem als Projektleiterin verantwortlich für die inhaltliche Überarbeitung. Prof. Dr. Behrs Beitrag lag in Vorschlägen und Hinweisen zur methodischen Herangehensweise, und einer detaillierten sprachlichen und formellen Überarbeitung.

Kapitel 6 der vorliegenden Dissertation - "Motives for dropping out from higher education - An analysis for Bachelor students in Germany" - ist eine gemeinsame Forschungs-

arbeit mit Prof. Dr. Andreas Behr, Dr. Katja Theune und Hervé Donald Teguim Kamdjou.

Mein eigener hauptverantwortlicher Beitrag betrifft die Beschreibung der Methoden (Kapitel 6.4). Außerdem war ich verantwortlich für die Ergebnisse der Clusteranalyse und der Zusammenfassung der Variablen mittels Hauptkomponentenanalyse (Kapitel 6.5.1, 6.5.4 und 6.5.5). Herrn Teguims Beitrag lag in der Datenbeschreibung (Kapitel 6.3) und in der deskriptiven Auswertung (Kapitel 6.5.2 und 6.5.3). Dr. Theune war maßgeblich verantwortlich für die Einleitung (Kapitel 6.1), die Literaturübersicht (Kapitel 6.2) und die Zusammenfassung der Ergebnisse (Kapitel 6.6). Sie war außerdem als Projektleiterin verantwortlich für die inhaltliche Überarbeitung. Prof. Dr. Behrs Beitrag lag in Vorschlägen und Hinweisen zur methodischen Herangehensweise, und einer detaillierten sprachlichen und formellen Überarbeitung.

Hiermit wird die Richtigkeit der obigen Angaben zur Koautorenschaft bestätigt:

Ort, Datum	Marco Giese
Ort, Datum	Prof. Dr. Andreas Behr
Ort, Datum	Dr. Katja Theune
Ort, Datum	Hervé Donald Teguim Kamdjou