Research Article

Andreas Behr*, Marco Giese, Hervé D. Teguim K., Katja Theune

Dropping out from Higher Education in Germany an Empirical Evaluation of Determinants for Bachelor Students

<https://doi.org/10.1515/edu-2020-0104>

received April 22, 2020; accepted July 12, 2020.

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.

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, 2020). 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.

*Corresponding author: **Andreas Behr**, University Duisburg-Essen, Essen, Germany, E-mail: Andreas.Behr@vwl.uni-due.de
Marco Giese, Hervé D. Teguim K., Katja Theune, University Duisburg-Essen, Essen, Germany

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.

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., 2013b). 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 "reselection" (Larsen et al., 2013b) or "institutional departure" (Tinto, 1993, p. 36), the latter "de-selection" (Larsen et al., 2013b) 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, 2014). 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.

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 relationship 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 (2010) 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.

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.

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). 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.

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 (2019) 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.

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 (2019) 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).

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. (2013b), 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 Sarclotti 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 The National Educational Panel Study

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 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 A2 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. (2020b) 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

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-Universität Bamberg in cooperation with a Germany-wide network.

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 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 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%).

3.2 Predictor variables included in the study

Since the NEPS contains more than 3,000 variables in total, a variable preselection 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 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.

Table 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%

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 Identifying dropouts

According to Larsen et al. (2013a), 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.

Table 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%

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.

4 Bivariate analysis of dropout determinants

Let Y be the status variable and $\mathbf{X} = \{X_1, \dots, X_k\}$ a set of k determinants, that are potentially related to the status Y . In section 5 (multivariate analysis), the focus lies on $P(Y = 1 | X_1 = x_1, \dots, X_p = x_p)$, 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), $j = 1, \dots, k$.

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 , $j = 1, \dots, k$ (Bortz and Schuster, 2010), calculated by

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

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.

with sample sizes $n_{1,j}$ (dropout) and $n_{0,j}$ (graduate) in the two groups, S_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 A3 in the appendix. Table A3 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 standard deviation

$$S = \frac{(n_{1,j} - 1)S_{1,j} + (n_{2,j} - 1)S_{2,j}}{n_{1,j} + n_{2,j}}$$

and $S_{1,j}$ and $S_{2,j}$ are the weighted standard deviations in the two groups for variable j . The interest is more on the absolute value of Cohen's d , i.e. $|\text{Cohen's } d|$ to give a ranking of variables with comparable huge differences in the two groups.

Table 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

4.1 Demographic and family background

According to Table 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, 2010). 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.

4.2 Financial situation

Strongly related to the family background of students is their financial situation and off-study work. Dropouts more often receive financial aid (BAföG: Bundes Ausbildungsförderungs Gesetz, financial support for students with a poor socioeconomic 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.

4.3 Prior education

Educational achievements up to secondary education generally influence the higher education performance (Sarclotti and Müller, 2011). According to Table 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

³ For all calculations in this article the statistical software R (R Core Team, 2018) is used.

Figure 1: Distribution of subjects in the dropout and graduate group.

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.

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 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).

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 (2010) 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: Bivariate analysis of determinants and student status.

Variable	Dropouts group		Graduates group		Effect size	
	Mean	Std. Err.	Mean	Std. Err.	Cohen's <i>d</i>	<i>r</i> _{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
Financial aid	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 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 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

Continued **Table 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_{gh} $
Institutional determinants						
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

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.

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 $\mathbf{X} = (X_1, \dots, X_d)$ (here, $d = 52$).

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

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

$$P(Y = 0|\mathbf{X}) = 1 - P(Y = 1|\mathbf{X}) = 1 - F(\beta_0 + \mathbf{X} \boldsymbol{\beta}). \quad (2)$$

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

$$P(Y = 1|\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 \cdot X_1 + \dots + \beta_d \cdot X_d)}{1 + \exp(\beta_0 + \beta_0 + \beta_1 \cdot X_1 + \dots + \beta_d \cdot X_d)}. \quad (3)$$

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

Additionally to the main effects $X_1, \dots, X_d$ considered in Equation (3), interaction effects and effects in quadratic order are sometimes included in the logit model. This leads to:

$$P(Y = 1|\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 \cdot X_1 + \dots + \beta_d \cdot X_d + \sum_{i=1, j=1, i \leq j}^d \gamma_{i,j} X_i X_j)}{1 + \exp(\beta_0 + \beta_1 \cdot X_1 + \dots + \beta_d \cdot X_d + \sum_{i=1, j=1, i \leq j}^d \gamma_{i,j} X_i X_j)}. \quad (4)$$

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

5.2 Best subset model

As the number of predictor variables can rapidly increase (as in Equation (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):

$$\min_{\beta_0, \boldsymbol{\beta} \in \mathbb{R}^{d+1}} \left[\frac{1}{n} \sum_{i=1}^n y_i \cdot (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}) - \log(1 + \exp[\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}]) \right] + \lambda[(1 - \alpha)\|\boldsymbol{\beta}\|_2^2/2 + \alpha\|\boldsymbol{\beta}\|_1], \quad (5)$$

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.

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.

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. Garcarena 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).

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) 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 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 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 2.

As noted in Table 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 (BAföG) 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 counter-

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.

Table 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 budget term break: Employment	-0.227	0.050	-4.583	0.000***
Study costs	-0.091	0.044	-2.074	0.038**
Financial aid	-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 (Mathematics)	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

		predicted_value	
		dropout	graduate
real_value	dropout	700	243
	graduate	698	1,927

Accuracy = 73.35%
Recall= 74.23%
Threshold= 0.265

Table 5: Confusion matrix.

intuitive 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.

Figure 2: ROC-curve.

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 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.

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.

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.

7.3 Financial situation

The financial situation of students, for instance in form of financial aid (BAföG 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.

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, 2014). 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.

References

Aina, C. (2013). Parental background and university dropout in Italy. *Higher Education*, 65(4), 437–456.

- Batista, G. E. and Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6), 519–533.
- Beerens, 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., Giese, M., Tegum K., H. D., and Theune, K. (2020a). Early prediction of university dropouts—a random forest approach. *Jahrbücher für Nationalökonomie und Statistik (forthcoming)*. <https://doi.org/101515/jbnst-2019-0006>.
- Behr, A., Giese, M., Tegum K., H. D., and Theune, K. (2020b). Dropping out of university: A literature review. *Review of Education, forthcoming*.
- Belloc, F., Maruotti, A., and Petrella, L. (2010). University drop-out: an Italian experience. *Higher Education*, 60(2), 127–138.
- Blossfeld, H.-P., Roßbach, H.-G., and von Maurice, J. E. (2011). Education as a lifelong process-The German National Educational Panel Study (NEPS). [Special Issue] *Zeitschrift für Erziehungswissenschaft*, 14.
- Bortz, J. and Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler. Lehrbuch mit Online-Materialien*. Springer-Lehrbuch. Springer Berlin Heidelberg.
- Brandstätter, H., Grillich, L., and Farthofer, A. (2006). Prognose des Studienabbruchs. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 38(3), 121–131.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Darlington, R. B. and Hayes, A. F. (2016). *Regression analysis and linear models: Concepts, applications, and implementation*. Guilford Publications.
- DESTATIS (2020). Gesellschaft & Staat - Bildung, Forschung, Kultur. [https://www.destatis.de\(03.03.2018\)](https://www.destatis.de(03.03.2018)).
- 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 dropout: the Italian experience. *Economics of Education Review*, 27(5), 546–555.
- Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.
- 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.
- Garciaarena, 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.
- Glocker, D. (2011). The effect of student aid on the duration of study. *Economics of Education Review*, 30(1), 177–190.
- Gury, N. (2011). Dropping out of higher education in France: a micro-economic approach using survival analysis. *Education Economics*, 19(1), 51–64.
- Hartung, J., Knapp, G., and Sinha, B. K. (2011). *Statistical meta-analysis with applications*, volume 738. John Wiley & Sons.

- Hastie, T. and Qian, J. (2014). Glmnet vignette. page 2016. http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer.
- Hetze, P. (2011). *Nachhaltige Hochschulstrategien für mehr MINT-Absolventen*. Stifterverband für die Deutsche Wissenschaft.
- Heublein, U. (2014). 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 Studierenerwartungen und Studienwirklichkeit. *Forum Hochschule* 1/2017.
- Heublein, U., Schmelzer, R., Sommer, D., and Wank, J. (2008). Die Entwicklung der Studienabbruch- und Schwundquoten an den deutschen Universitäten und Fachhochschulen. *HIS-Projektbericht*. Hannover.
- 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.
- Kleinke, K., Reinecke, J., Salfrán, D., and Spiess, M. (2020). *Applied Multiple Imputation: Advantages, Pitfalls, New Developments and Applications in R*. Springer Nature.
- Korhonen, V. and Rautopuro, J. (2019). Identifying problematic study progression and “at-risk” students in higher education in Finland. *Scandinavian Journal of Educational Research*, 63(7), 1056–1069.
- Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Crossvalidation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(1), 1–15.
- 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.
- Lassibille, G. and Navarro Gómez, L. (2008). Why do higher education students drop out? Evidence from Spain. *Education Economics*, 16(1), 89–105.
- 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. In *Stratification in Higher Education*, pp. 240–265. Stanford University Press Stanford, CA.
- 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*, 4(2), 128–134.
- Mora, T. (2010). Why higher graduated regret their field of studies? Some evidence from Catalonia, Spain. *Education Economics*, 18(1), 93–109.
- Müller, S. and Schneider, T. (2013). Educational pathways and dropout from higher education in Germany. *Longitudinal and Life Course Studies*, 4(3), 218–241.
- 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.
- R Core Team (2018). R: A language and environment for statistical computing. R foundation for statistical computing. <http://www.R-project.org>.
- Reisel, L. and Brekke, I. (2010). Minority dropout in higher education: a comparison of the United States and Norway using competing risk event history analysis. *European Sociological Review*, 26(6), 691–712.
- Rodríguez-Gómez, D., Feixas, M., Gairín, J., and Munoz, J. L. (2015). Understanding Catalan university dropout from a cross-national approach. *Studies in Higher Education*, 40(4), 690–703.
- Sarclotti, 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. (2014). Do tertiary dropout students really not succeed in European labour markets? *IZA Discussion Paper No. 8015*.
- 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.
- 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.
- Statistisches Bundesamt (2011). Bildung und Kultur - Studierende an Hochschulen Wintersemester 2010/2011 Fachserie 11 Reihe 4.1. Technical report.
- 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.
- 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*, 49(1), 16–28.
- 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.
- 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. (1993). *Leaving College: rethinking the causes and cures of student attrition*. Chicago: Chicago University Press, 2nd edition.
- Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 23(5), 373–405.

- 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.
- Vossensteyn, H., Stensaker, B., Kottmann, A., Hovdhaugen, E., Jongbloed, B., Wolscheid, S., Kaiser, F., and Cremonini, L. (2015). Dropout and completion in higher education in Europe. Luxembourg: Publications Office of the European Union.
- 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.
- 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.

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 A1.⁵ 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 A2 shows the results.

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 A2 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 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 + $22^2 = 231$ interactions of the second order). As λ varies, values of MSE and AUC and the number of nonzero coefficients are recorded (Figure A1).

As displayed in Figure A1, 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 $\alpha = 0.268$. 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 A1. 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 A1, 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.

⁵ Models are fitted using the function *glmnet*.

⁶ The deviance is defined as $2 \cdot (\log\text{like sat} - \log\text{like})$, where $\log\text{like sat}$ is the loglikelihood for the saturated model (Friedman et al., 2010).

⁷ The function *cv.glmnet* is used for that.

Figure A1: 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.

Figure A2: Selection of the best λ parameter based on MSE and AUC.

Figure A3: Models including main effects, interactions between the predictors and curvilinear effects.

Table A1: 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 A2: 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

Table A3: 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

Continued **Table A3:** Attribute description.

Attribute	Description (Data type)
Financial situation	
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 repeated 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

Continued **Table A3:** Attribute description.

Attribute	Description (Data type)
Motivation and satisfaction with study	
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 interesting (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 match	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