Berens, Johannes; Schneider, Kerstin; Görtz, Simon; Oster, Simon; Burghoff, Julian

Working Paper

Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods

CESifo Working Paper, No. 7259

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Berens, Johannes; Schneider, Kerstin; Görtz, Simon; Oster, Simon; Burghoff, Julian (2018) : Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods, CESifo Working Paper, No. 7259, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/185457

CESifo WORKING PAPERS

7259
2018

September 2018

Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods

Johannes Berens, Kerstin Schneider, Simon Görtz, Simon Oster, Julian Burghoff

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editors: Clemens Fuest, Oliver Falck, Jasmin Gröschl
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7259*
Category 5: Economics of Education

Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods

Abstract

To successfully reduce student attrition, it is imperative to understand what the underlying determinants of attrition are and which students are at risk of dropping out. We develop an early detection system (EDS) using administrative student data from a state and a private university to predict student success as a basis for a targeted intervention. The EDS uses regression analysis, neural networks, decision trees, and the AdaBoost algorithm to identify student characteristics which distinguish potential dropouts from graduates. Prediction accuracy at the end of the first semester is 79% for the state university and 85% for the private university of applied sciences. After the fourth semester, the accuracy improves to 90% for the state university and 95% for the private university of applied sciences.

JEL-Codes: I230, H420, C450.

Keywords: student attrition, machine learning, administrative student data, AdaBoost.

Johannes Berens
WIB, University of Wuppertal
Wuppertal / Germany
berens@wiwi.uni-wuppertal.de

*Kerstin Schneider**
WIB, University of Wuppertal
Wuppertal / Germany
schneider@wiwi.uni-wuppertal.de

Simon Görtz
University of Wuppertal
Wuppertal / Germany
sgoertz@uni-wuppertal.de

Simon Oster
WIB, University of Wuppertal
Wuppertal / Germany
oster@wiwi.uni-wuppertal.de

Julian Burghoff
WIB, University of Wuppertal
Wuppertal / Germany
burghoff@wiwi.uni-wuppertal.de

*corresponding author

September, 2018
We gratefully acknowledge funding for the project “FragSte” by the Federal Ministry of Education and Research (BMBF Förderrichtlinie “Studienerfolg und Studienabbruch”). We also thank J. Michael Perdue for his support.

Die Zahl der Kröten und Kraniche ist in Deutschland
besser erfasst als die Zahl der Studienabbrecher.

The number of toads and cranes in Germany is better understood
than the number of university dropouts.

***Dr. Angela Merkel** - 09.12.2016*

1. INTRODUCTION

Student attrition at universities has a negative impact on all parties involved: students, institutions involved, and the general public (Bowen et al., 2009; Bound et al., 2010). Notwithstanding the educational gain of a student prior to dropping out, university attrition represents a misuse of public and private resources. In addition to monetary losses, dropping out can cause feelings of inadequacy and lead to one being socially stigmatized (Larsen et al., 2013). The importance of academic performance and informational frictions for explaining attrition has been stressed in recent literature (Stinebrickner & Stinebrickner, 2008; 2012; 2013; 2014; Arcidiacono et al., 2016). But despite the importance of the topic, there is still much that is unknown about the underlying determinants of attrition and effective means for reducing it.

Facing high attrition rates and an ever greater demand for a workforce qualified in STEM related subjects, education policy makers are increasing their efforts to reduce the number of student dropouts (Gaebel et al., 2012). An appropriate initiative to combat student attrition must 1) be cost efficient and 2) should target students in danger of dropping out. First, students at risk need to be identified using available information (administrative data); second, at-risk students need to come into contact with a relevant outreach program and, finally, the intervention needs to be evaluated.

The present paper contributes to the first point. We present a self-adjusting early detection system (EDS) that can be implemented at any point in time within a student's career. The EDS uses student administrative data that is regularly updated and maintained by legal mandate at all German universities. Not only does this ensure low maintenance and startup costs when implementing an EDS, it also allows the EDS to be readily implemented at every (type) of university throughout Germany; and with minor adjustments, the EDS can be transferred to most institutes of higher education worldwide. At the end of each semester, the EDS is updated to reflect the most recent student performance data and changes in the composition of the student body and study programs; hence, the EDS is self-adjusting. Once implemented, the system is not constrained to a sample of students but can make use of the longitudinal census of all students. This precludes the need for costly student surveys which would otherwise need to be performed regularly for the whole student body and would depend on voluntary participation of the students. Furthermore, an EDS that uses readily available administrative data can be implemented and run without involvement from university staff, thus, considerably easing legal concerns for data protection. The system can be used to monitor individual student groups, study programs, entire student cohorts, and, if desired, even individual students. Thus, the EDS provides a good starting point for research on student attrition, it offers important insights for university administration, and it can serve as a basis for interventions. It can therefore support

the strategic, tactical, and operational decision making processes of universities. For example, the EDS allows for studying the effects of changes in study programs and courses, the influence of entry barriers on enrollment, e.g., study fees, and it can monitor the efficiency of intervention measures and aid programs.

The EDS can also be useful in the efficient allocation of support and intervention measures to reach at-risk students. As a general rule, there are a large number of preventative measures taken at a university to reduce the number of student dropouts. Unfortunately, these programs currently do not help in identifying at-risk students and are, thus, offered to the general student body. Accordingly, in order for at-risk students to benefit from them, they have to self-select into a program. Hence, due to a matching problem, individual support networks and assistance programs may go underutilized.

We develop and test the EDS using two medium-sized universities in the federal state of North Rhine-Westphalia: a state university (SU) with about 23,000 students and 90 different bachelor programs and a private university of applied science (PUAS) with about 6,700 students and 26 bachelor programs. The state university is tuition free, while tuition at the private university is about 400 Euros per month.

Instead of relying on only one method for prediction purposes, we present a selection of methods starting with regression models, followed by different machine learning methods, and finally combining all of the approaches in a boosting algorithm. Our results indicate that 74% of SU and 72% of PUAS dropouts are correctly identified at the end of the first semester; furthermore, the accuracy of the EDS increases as new student performance data becomes available at the end of each consecutive semester: after the fourth semester, the EDS correctly predicts 80% of the SU and 83% of the PUAS student dropouts. Confirming earlier studies, performance data, in particular at the early stages, is important for predicting student attrition, while demographic data has limited predictive value once performance data is made available.

This paper is organized as follows. Section 2 reviews related literature. Section 3 offers a description of the data. Section 4 explains the empirical strategy. In Section 5 we present the results. Section 6 concludes.

2. RELATED LITERATURE ON STUDENT ATTRITION

Stinebrickner and Stinebrickner (2008; 2012; 2014) contribute to understanding the determinants of student attrition by using data from the Berea Panel Study, which includes two cohorts of students who entered Berea College in 2000 and 2001. The main insights from those studies are that financial factors do not play a significant role in explaining college attrition. Instead, academic performance is the most important factor. Moreover, they stress the importance of learning how academic performance affects the decision to drop out. The

importance of informational frictions for college attrition is confirmed by Arcidiacono et al. (2016)

Another strand of the literature presents a sociological approach to the topic (Larsen et al., 2013). Tinto's (1975) "student integration model" established the central importance of the social and academic integration of the student. Pascarella and Terenzini (1979) adopt the idea of integration and extend the model by distinguishing between forced and voluntary attrition. Bean (1983), on the other hand, presents the importance of integration as a main predictor of attrition and adds student satisfaction as a central variable.

Other papers focus on the importance of individual characteristics like minority status or the chosen field of study. Arcidiacono (2016) used data from the University of California to argue that minorities are less likely to be accepted into STEM related graduate programs, because they often lack the necessary academic proficiency to successfully compete in the programs with their (non-minority) peers.

The quality of empirical research on student attrition depends on the availability of good data. There are two types of data that have been exploited in the literature: administrative data and survey data. In German universities, in contrast to, for instance, the British research approach, data on student attrition are mainly gathered from surveys (Larsen et al., 2013) due to the lack of available administrative data. However, student surveys have significant limitations when investigating the causes of attrition. In ex-ante interviews, the dependent variable, student attrition, must be replaced with the intention of dropping out. Using the intention to drop out as a predictor for actually dropping out is, however, controversial in the literature as it assumes that the intention is not exaggerated or otherwise subjected to self-adjustment.

Using student administrative data, Arulampalam et al. (2005) and Danilowicz-Gösele et al. (2014) show for Great Britain and Germany, respectively, that the probability of dropping out can be determined from the analysis of student data. The academic performance of the student and the performance of the student's peer group are both relevant for predicting student attrition.

Improvements in data mining and machine learning methods have facilitated the use of automated methods to forecast student attrition. A relatively recent discipline of educational data mining has emerged (Dekker et al., 2009) that addresses, in particular, study courses with high attrition rates—e.g., distance learning courses. Kotsiantis et al. (2003) analyze demographic and performance data using machine learning methods to identify successful students. They correctly predicted more than 70% of successful students using various methods such as decision trees, neural networks, a naive Bayes method, logistic regression analysis, support vector machines, and instant learning algorithms. Subsequent studies have largely followed a similar structure and methodology. Examples are Xenos (2004), Minaei-Bidgoli et al. (2004), Nghe et al. (2007), Dekker et al. (2009), Zhang et al. (2010), Bayer et al. (2012), Er

(2012) and Yukselturk et al. (2014), Sara et al. (2015), Santana et al. (2015) and Kemper et al. (2018). While the studies are not easily comparable due to differences in sample size, variable settings, research methods, and research questions the different methods employed within a given study resulted in only marginal differences in predictive accuracy. Significant differences in between-study results primarily reflect the predictive quality of the data, i.e., the power of the data to predict study outcomes.

However, besides student achievement data, additional influential factors exist that can increase the accuracy of prediction. For example, the research work mentioned above by Kotsiantis et al. (2003) used, in addition to demographic and performance data, data from optional face-to-face consultations with university staff. Zhang et al. (2010) base their data selection on Tinto's integration model (1975) and collect information best describing the social and academic integration of the student. For this purpose, performance data, registration in online learning platforms, use of the university library, reading behavior from the online library as well as online activity were all evaluated. In particular, learning behavior and student-teacher interactions were easier to observe and collect data on. Other components of Tinto's integration model, such as personal development of the student, interest in the subject matter, and social integration could not be observed. Bayer et al. (2012) address social integration in the university environment. They evaluated the behavior and social connections of 775 students in social networks. It turned out that more active and cross-linked—integrated—students were more successful. Furthermore, after adding the social network data, the prognostic accuracy of first semester data increased by 5 percentage points to 72%.

3. ADMINISTRATIVE STUDENT DATA USED IN THE EDS

The EDS developed in this paper uses student administration and performance data to predict whether a student will drop out of his/her program. Using historical student data from dropouts and graduates, our system identifies the demographic and performance characteristics of students who are at risk of dropping out. The current analysis restricts itself to bachelor level degree programs; however, the method can be easily applied to master level programs, as well.

As already mentioned, the EDS was developed and tested at two medium-sized universities in the federal state of North Rhine-Westphalia: a state university (SU) with about 23,000 students and 90 different bachelor programs and a private university of applied sciences (PUAS) with about 6,700 students and 26 undergraduate programs. The machine learning process was performed using administrative data from former bachelor students, who either dropped out or graduated between 2007 and 2017. The forecasting system was then tested using student data that had not been included in the training data. Data from students who matriculated in 2012

and 2010 were chosen for the test data of PUAS and the SU, respectively.[1] The training data of former SU students who matriculated in 2007-2009 and 2011-2017 comprised a total of 12,730 observations; the 2010 data used for testing consisted of 1,766 bachelor students. PUAS training data from former students who matriculated in 2007-2011 and 2013-2017 included a total of 6,297 observations; the test data from 2012 comprised 1,303 bachelor students.

The EDS is designed in such a way that it can be introduced and operationally maintained at low cost in German state and private universities as well as in universities of applied sciences. Provided that the administrative data requirement is met, the implementation is, of course, not limited to Germany. For ease of implementation, however, it is necessary that only standardized data—data which is necessarily collected by law at all universities—be required for implementing the system.

The standardized and nationally available student data used in the EDS is collected and stored by mandate of the Higher Education Statistics Act (HStatG). The HStatG established a nationwide standard for collecting specific student data. Furthermore, §3 HStatG, which is relevant to the present analysis, was last modified in 1997 (BGBI I, 1997, p. 3158). According to §3 HStatG, both public and state-recognized private universities have to collect, store, and regularly report the student data outlined in Table 1.

[1] We choose the year 2010 as our test cohort at the SU because, while the duration of the bachelor program is 6 semesters—similar to the programs in the PUAS, the actual observed duration of studies is longer at the SU as compared to the PUAS.

Table 1: Data collected according to the Higher Education Statistics Act

		Data collected according to the Higher Education Statistics Act	Variables	Values
Demographic data	Personal	Year of birth	Age at enrolment	Age in years
		Gender	Gender	1=male; 0=female
		Place of birth	Federal state of birth	16 German federal states
		Nationality	Nationality	1=foreign; 0=German
			Region and country of birth	11 regions and 5 countries
		First and last name	Migration background of students	Probability in percent
		Health insurance company	Health insurance (private / state)	1=private; 0=public
	Previous education	Type of university entrance qualification	Type of entrance degree (AHR, FHR, fgHR, foreign)	1 to 4
		City where university entrance qualification was earned	City where university entrance degree was earned	1 = City of university; 0 = else
		Grade of university entrance qualification	Grade of entrance degree	1.00 to 4.00
		No. of semesters in previously enrolled study programs	Lateral entrants	1=yes; 0=no
		Number of study programs previously enrolled in at this university	Number of previous semesters	0 to max
			Number of previous courses of study at this university	0 to max
	Study	Course of study	Course of study or number of simultaneous enrolled programs	1 to max
		Type of study program	Study form (full time / part time / dual)	1 to 3
Academic Performance data		Name of exam	No. of important successfully completed exams	1 to 9
			No. of other successfully completed exams	0 to max
		Exam grade	Average grade per semester	1.00 to 4.00
		Date of exam		
		Result of enrolled exams (pass/fail/withdrawn/no-show)	No. of failed exams per semester	0 to max
			No. of exams per semesters not participated in	0 to max
			No. of no-show exams per semester	0 to max
Outcome		Ex-matriculation date	Graduate or drop out	1=drop out, 0=graduate
		Reason for ex-matriculation		

Notes:
Nationality: Citizenship and place of birth distinguishes foreign from non-foreign students, students without a migration background, and students that are first generation immigrants.
Immigration background: Name based imputation of immigration background distinguishes between students that are second generation immigrants and those that are not.
Type of entrance degree: AHR = university entrance degree, FHR = university of applied science entrance degree, fgHR = restricted subject-specific entrance degree, foreign = foreign entrance degree.
Average grade: Failed exams have to be rewritten; thus, they don't lower the GPA.
No. of exams per semesters not participated in: When available, some universities register when a student has withdrawn from an exam, others don't. Furthermore, some universities register non-participation—when a student neither withdraws nor presents a medical excuse—as a "no-show", others as a "not-pass". The latter can't be distinguished from failed exams.

In the event that additional relevant student data are collected at universities, the EDS can be expanded to accommodate additional variables. For example, the university entrance qualification grade is, according to prevailing opinion, a well-suited predictor of study outcomes (Danilowicz-Gösele et al., 2014; Trapmann et al., 2007; Brandstätter & Farthofer, 2002).

Using the standardized student data referenced in Table 1 has advantages, but it certainly limits the dimensions of the EDS in explaining and predicting dropouts. Some of the reasons cited in the literature for dropping out are not captured by the student data collected at universities. In the literature reviewed above, it is agreed that the determinants of attrition are multi- and not monodimensional and include the student's self-concept (Burrus et al., 2013; Larsen et al., 2013, p. 47). With regard to German universities, Heublein, et al. (2014) identified seven causes for attrition: performance requirements, finances, exam failure, lack of motivation, study conditions, professional reorientation, and illness. Wiers-Jenssen et al. (2002), on the other hand, states that student satisfaction is a key factor for student success, although it is unclear whether a lack of satisfaction leads to dropping out or whether dropping out leads to lack of satisfaction or if both are mutually dependent.

While possibly important for explaining student dropout rates, the following data are not available for every student at every university and, even if available, data protection laws render it difficult if not impossible to use in an EDS: information on student satisfaction, financial circumstances, family situation, personal motivation, individual fit of the institutional framework, diligence while choosing the course of study, professional interest in the subject of study, professional inclination, academic or social student integration, and the student's state of health. Thus the EDS is based on student demographic and academic achievement data that are collected according to §3 HStatG. Moreover, the central importance of academic achievement as a predictor for dropping out is emphasized again and again in the literature (Larsen et al., 2013). The extent to which data limitations impede the efficacy of the EDS depends on whether and how quickly the above-mentioned factors influence academic performance before leading to student attrition.

Table 1 shows how the §3 HStatG student data are transformed into the variables used in the EDS. In summary, the demographic variables consist of the following information:

- Personal: age, gender, address, place of birth, immigration background
- Previous education: type and place of university entrance qualification, previous academic experience
- Study: course of study, type of enrollment (i.e., full- or part-time)

Additional information for students with an immigration background includes nationality, domestic or foreign university entrance qualification, and whether the student is a first or second generation immigrant (cf. Section 3.1. below).

In addition to the demographic data, student performance data are also made available at the end of each subsequent semester. The student performance data collected at the end of each completed semester include the average semester grade, average semester credit points earned, the number of registered but unattended exams, and the number of attempted but failed exams. In addition, it is determined how many of the "most important" exams were passed in a given semester. An exam is determined to be most important when its successful completion is highly correlated with the successful completion of the degree. Finally, in order to fit our model, former students are classified as dropouts or graduates.

3.1. EXPANDING THE DATA BY IMPUTING INFORMATION ON MIGRATION STATUS AND ADDITIONAL INFORMATION

In addition to the student data collected pursuant to §3 HStatG, the EDS is able to utilize additional student data which may already exist or which can be imputed from the available data. If known, students' home addresses can be used to gather socioeconomic data on the student, the university entrance grade can provide information about previous academic performance, and the first and last name can provide information on student immigration background.

German universities distinguish between a semester address and a home address. Accordingly, it is possible to determine whether the student has moved from her home for the purpose of studying, is commuting over long distances, or is studying in her home town (Dekker et al., 2009). Using the home address postal code, the median income from the student's postal code area can be used as a proxy for income (Danilowicz-Gösele et al., 2014). Another variable which can act as a proxy for socioeconomic background is health insurance type. Here one can distinguish between private and publicly insured students (Danilowicz-Gösele et al., 2014). Students with private health insurance are primarily children of parents who are self-employed, civil servants, or employees with an income above a certain threshold (in 2017 €57,600 per year). Thus, students with private health insurance are typically from families with a higher socioeconomic background.

Immigration background (with or without German citizenship) has been shown to be particularly helpful for predicting educational success in Germany. As a rule, however, institutions of higher education typically only know a student's citizenship, place of university entrance qualification, and place of birth. Thus, international students can be included in the group of foreign educated students and students that have been partly or wholly educated in Germany—but do not hold German citizenship. Non-German citizens born abroad are considered first generation immigrants. However, second generation immigrants with German citizenship cannot be directly identified from university administration data.

Since it is known, however, that second and third generation immigrants underperform in the German educational system, it is important to be able to identify them. For this reason, first names and family names of students are examined to determine their ethnic origin. Germans born in Germany, whose first and surnames reveal an immigration background, are considered migrants of the second or third generation. The method of Humpert and Schneiderheinze (2002) is a common method for determining a subject's country and region of origin from the combination of first and surnames (Berger et al., 2004). Based on the methodology of Humpert and Schneiderheinze (2002), a name-database containing around 200,000 first names and another database containing around 600,000 surnames (Michael, 2007; Michael, 2016) are used. There is a probability for each country-name combination (including a total of 145 countries) that indicates the likelihood that the person in question migrated from the given country. For gender-specific names, the information of the gender is also included; gender-neutral names, as well as names for which the gender-specificity depends upon the country, are marked as well.

Using the information in the database, the probability of an immigration background is determined from the distribution of first and surnames in represented countries; the region/country of origin is determined in a second step. Since most names are common in more than one country, the 145 countries are aggregated into 11 regions. In accounting for the main countries and regions of origin for immigrants into Germany, we distinguish between the following 11 regions (Statistisches Bundesamt, 2015):

- North America
- Central and South America
- Northern and Western Europe
- Southern Europe
- Eastern Europe
- North Africa
- Rest of Africa
- Western Asia
- Eastern and South-Eastern Asia
- Southern Asia
- Australia, New Zealand, and Melanesia

Some of the regions above, such as the Americas, are uncommon regions of origin for foreign students in Germany. Thus, even though the countries in those regions are very heterogeneous, the high level of aggregation does not present a problem for the analysis of German student data. In Germany, the most frequently represented countries among students with an immigration background are Turkey, Italy, Croatia, Russia, and China (Statistisches Bundesamt, 2015; Heublein & Burkhart, 2013, p. 23). For this reason, in addition to the regions given above, these countries will be considered separately.

The validity of the imputation was checked in two different ways. Firstly, the group of non-German students with a known citizenship was used. Of the 4,004 foreign citizens, more than 94% of the first and surname combinations were correctly assigned. Secondly, the imputed immigration background from 1,598 first names was compared with the migration information in the German Socio Economic Panel (GSOEP). In the questionnaire the respondents report their first name and, if applicable, immigration background. Applying our imputation method to the GSOEP information, we correctly label 82% of the existing and non-existing immigration backgrounds. Note that in the second test—using the GSOEP data—only the subject's first name was used which is expected to lower the accuracy of the imputation. Excluding the subject's surname lowered the imputation's accuracy in the first test from 94% to 88%.

The imputed immigration data for both universities is summarized in Table 2. At both universities, 29% of the students are first or second generation immigrants, and the distribution of countries of origin is similar at both universities. The only difference is that the proportion of Chinese and Turkish students is higher at the PUAS.

Table 2: Ethnic composition of student population

	State University				Private University of Applied Sciences			
			N	26,686	16,192			
			not identified	234	147			
			Germans	18,574	11,510			
			MigBackg	7,721	4,684			
			MigRate	**28.93%**	**28.93%**			
Region	Students with foreign nationality	Domestic students with migration background	Migration background	Proportion of student body	Students with foreign nationality	Domestic students with immigration background	Migration background	Proportion of student body
North America	8	46	54	0.20%	0	41	41	0.25%
Central & South America	27	133	160	0.60%	15	88	103	0.64%
Northern & Western Europe	103	1,102	1,205	4.52%	88	778	866	5.35%
Southern Europe	615	324	939	3.52%	341	255	596	3.68%
Eastern Europe	433	419	852	3.19%	87	322	409	2.53%
North Africa	296	137	433	1.62%	90	113	203	1.25%
Rest of Africa	136	116	252	0.94%	39	61	100	0.62%
Western Asia	748	697	1,445	**5.41%**	812	1,008	1,820	**11.24%**
Eastern & Southeast Asia	392	323	715	2.68%	142	65	207	1.28%
Southern Asia	116	165	281	1.05%	40	201	241	1.49%
Australia/New Zealand/Melanesia	3	2	5	0.02%	0	1	1	0.01%
Countries								
Italy	174	153	327	1.23%	102	103	205	1.27%
Russia	93	154	247	0.93%	33	143	176	1.09%
Turkey	620	641	1,261	**4.73%**	761	1,000	1,761	**10.88%**
China	278	274	552	**2.07%**	123	26	149	**0.92%**
Germany	23,757	0	-	71.07%	14,537	0		71.07%

Notes:

N: No. of undergraduate SU (PUAS) students between 2000 and 2017 (2007 and 2017).

not identified: First and second name not in the database.

Germans: Students with German citizenship and no apparent immigration background.

MigBackg: Students with foreign nationality, foreign place of birth, or, most likely, a foreign name.

3.2. DATA DESCRIPTION

Tables 3a and 3b show a summary of the data for both universities. In each of the columns, the data is summarized with respect to year of enrollment. Thus the descriptive statistics in column (6) refers to students who enrolled in 2012 and either dropped out or graduated by 2017 or earlier. First, looking at the SU, women are overrepresented in most of the years, which is most likely explained by a large education department at the SU (cf. Table 3a). Age at enrollment is between 21 and 22.6 years. Between 24% and 29% of the students do not have an immigration background. The percentage of foreign born students is between 7% and 11%. There does not appear to be a time trend with regard to migration. The vast majority of students live in a city other than the home city of the university in question. The average grade for the university entrance exam is between 2.6 and 2.9. Between 5% and 8% of the students have private health insurance and the average number of failed exams is between 0.44 and 0.75.

Comparing the descriptive statistics for the PUAS in Table 3b to the descriptive statistics for the SU in Table 3a, it turns out that there are some substantial differences. Male students are overrepresented at the PUAS, the age of enrollment is higher, and there are more foreign students. Fewer students have a regular university entrance degree. There is no information about the grade of the entrance degree, nor do we have data on the type of health insurance. The average number of failed exams ranges between 0.17 and 0.62, and is thus lower than at the SU.

In the absence of performance data, the EDS forecasts are based solely on student demographic data. The demographic data available at the two universities differs. For instance, the number of students enrolled at an SU is usually substantially higher than the number enrolled at a PUAS. Moreover, enrollment at a PUAS is limited to only one study program. At the SU, however, of the 20,707 enrolled students between 2007 and 2017, 11,193 students were enrolled in two or more study programs, 10,467 in three or more programs, and 2,770 in four or more programs. Thus, at the SU, students might be counted more than once if they enroll in different programs; an example illustrates this. Students, who plan to become school teachers, study two majors, e.g., German and Mathematics. Consequently, they are enrolled in two different departments and would be counted twice. For this reason, type of study program is used as a predictor at the PUAS and not at the SU.

Furthermore, there are also differences regarding university entrance requirements. Generally, the prerequisites for studying at a PUASs are less restrictive than at a university; this is true for both the grade of the university entrance qualification (for instance, there might not be a numerus clausus) and type of university entrance qualification. As a result, the composition of the student body is different (cf. Tables 3a and 3b). As the institutions are different, the variables are likely to have a different impact on the prediction outcome. This does not only

apply to the demographic variables but also to the performance data, which has the highest explanatory power and is available after the first completed semester. Of particular importance are earned credit points per semester, average score of successfully completed exams, the number of successfully completed exams, and the successful completion of exams deemed most important for the student's respective study program.

Table 3a: Summary statistics: SU (mean and standard deviation)

Cohort	(1) 2007	(2) 2008	(3) 2009	(4) 2010	(5) 2011	(6) 2012
Gender (1=male; 0=female)	0.34	0.43	0.40	0.43	0.51	0.46
Age at enrollment	21.24 (3.15)	21.84 (3.75)	21.86 (3.56)	21.93 (3.72)	22.28 (4.38)	22.60 (4.67)
First generation immigrant (1=yes; 0=no)	0.08	0.11	0.10	0.09	0.07	0.08
Second generation immigrant (1=yes; 0=no)	0.19	0.17	0.19	0.18	0.17	0.19
City of entrance qualification (1= city of university; 0=else)	0.14	0.19	0.18	0.18	0.21	0.22
General university entrance qualification (1=yes; 0=no)	0.97	0.95	0.95	0.94	0.94	0.94
University of applied sciences entrance qualification (1=yes; 0=no)	0.00	0.01	0.00	0.01	0.01	0.01
Restricted university entrance qualification (1=yes; 0=no)	0.01	0.01	0.01	0.02	0.01	0.01
Foreign university entrance qualification (1=yes; 0=no)	0.03	0.04	0.03	0.03	0.04	0.03
Grade of university entrance qualification	2.87 (0.82)	2.85 (1.00)	2.79 (0.97)	2.71 (0.87)	2.68 (0.92)	2.61 (0.89)
Health insurance (1=private; 0=public)	0.06	0.08	0.06	0.05	0.07	0.06
# of enrolled study programs	3.06 (1.85)	2.62 (1.87)	2.67 (1.76)	2.71 (1.97)	2.32 (1.68)	2.20 (1.51)
Lateral entrants (1=yes; 0=no)	0.17	0.25	0.30	0.39	0.38	0.44
# of semesters at prev. university	1.63 (4.45)	2.47 (5.32)	2.65 (5.01)	3.15 (5.09)	2.63 (4.50)	3.02 (5.13)
Average grade per semester	2.46 (0.55)	2.49 (0.59)	2.45 (0.56)	2.50 (0.56)	2.51 (0.58)	2.49 (0.58)
Average CPs per semester	12.91 (16.44)	17.18 (26.92)	18.33 (29.29)	19.80 (30.12)	15.38 (22.38)	15.22 (23.87)
No exam taken	0.19	0.20	0.22	0.19	0.24	0.31
# of exams per semester not participated in	0.18 (0.62)	0.40 (1.56)	0.38 (1.44)	0.44 (1.28)	0.43 (1.16)	0.45 (1.35)
# of failed exams per semester	0.44 (1.02)	0.63 (1.78)	0.62 (1.88)	0.75 (1.90)	0.59 (1.24)	0.64 (1.77)
Obs.	2,637	1,846	2,215	2,170	2,860	2,674

Note: Performance data refers to data from the first semester.

Table 3b: Summary statistics: PUAS (mean and standard deviation)

Cohort	(1) 2007	(2) 2008	(3) 2009	(4) 2010	(5) 2011	(6) 2012
Gender (1=male; 0=female)	0.66	0.68	0.65	0.64	0.65	0.62
Age at enrollment	22.27 (2.99)	23.95 (3.33)	24.40 (3.22)	24.36 (3.04)	24.01 (2.76)	23.97 (2.39)
First generation immigrant (1=yes; 0=no)	0.13	0.15	0.13	0.10	0.08	0.09
Second generation immigrant (1=yes; 0=no)	0.20	0.15	0.18	0.17	0.18	0.18
City of entrance qualification (1= city of university; 0=else)	0.26	0.34	0.34	0.30	0.30	0.32
General university entrance qualification	0.56	0.49	0.48	0.49	0.51	0.49
University of applied sciences entrance qualification (1=yes; 0=no)	0.40	0.44	0.48	0.46	0.44	0.44
Restricted university entrance qualification (1=yes; 0=no)	0.01	0.00	0.01	0.02	0.02	0.05
Foreign university entrance qualification (1=yes; 0=no)	0.04	0.06	0.03	0.02	0.02	0.02
Lateral entrants (1=yes; 0=no)	0.32	0.34	0.34	0.34	0.34	0.34
Average grade per semester	2.37 (0.55)	2.32 (0.51)	2.32 (0.53)	2.28 (0.53)	2.24 (0.51)	2.28 (0.53)
Average CPs per semester	12.78 (11.15)	16.25 (10.84)	18.77 (10.71)	19.11 (11.17)	19.98 (11.67)	19.69 (11.63)
No exam taken	0.35 (0.00)	0.18 (0.00)	0.11 (0.00)	0.09 (0.00)	0.09 (0.00)	0.09 (0.00)
# of exams per semester not participated in	0.40 (0.57)	0.50 (0.80)	0.21 (0.42)	0.25 (0.46)	0.25 (0.43)	0.23 (0.49)
# of failed exams per semester	0.17 (0.34)	0.30 (0.45)	0.62 (0.79)	0.56 (0.69)	0.50 (0.62)	0.54 (0.68)
Obs.	193	1,175	1,423	1,343	1,358	1,563

Note: Performance data refers to data from the first semester.

4. EMPIRICAL STRATEGY

We now present the empirical strategy behind development of the EDS. Instead of relying on a single method, the EDS model is composed of multiple evaluation methods (classifiers). The methods are used alongside each other to evaluate their respective predictive powers. We combine the methods by means of the AdaBoost algorithm (Schapire & Freund, 1997; Schapire & Freund, 2012). The methods used for the analysis are the logit regression models, the neural network model, and decision tree algorithms.

First, a prediction model (parameters, weights, rules, and point estimates) is developed using the training data. The aim of the model is to identify potential dropouts as early as possible by classifying student observations as graduates or dropouts and then checking the precision of the prediction. Subsequently, the results of the individual methods are merged using the boosting algorithm first developed by Schapire and Freund (1997; 2012).

4.1. LOGIT MODEL

As a starting point for our analysis, we estimate a logit model

$$P(y_{it} = 1|x_i, z_{it}) = \Lambda(\beta_0 + \beta_1 x_i + \beta_2 z_{it}),$$

with *i* and *t* denoting student and semester, respectively and Λ representing the logistic function. The dependent variable y_{it} is a binary variable indicating graduate (0) and dropout (1). Demographic information, x_i, is time invariant, while the performance data, z_{it}, varies over time. Section 5 discusses the results of the logit model using student performance and demographic data from the time of enrollment up to the sixth and fourth semester for the SU and the PUAS, respectively. The logit model affords some advantages in that the coefficients are easier to interpret making it easier to understand the importance and magnitude of the explanatory variables on the likelihood of dropping out.

4.2. NEURAL NETWORK

The backpropagation algorithm is used for the multilayer perceptron (MLP). In summary, the architecture of the MLP can be described by about 31 neurons (depending on semester and university) in the input layer, 16 (8) neurons in the first (second) hidden fully-connected layer and one neuron in the output layer. We select the logistic function as the activation function for all neurons. The training process is briefly described below (Mucherino et al., 2009).

The neurons of the input layer become initialized with the training data set, which consists of the external inputs (determinant variables) and the actual outcome y_{it} (dropout or graduate).

All other neurons existing in the hidden layers are set randomly between minus one and one. In the supervised learning process, the network predicts student outcomes from the training data. The network then uses the assigned prediction weights and probability estimates to forecast student outcomes $\tilde{y}_{it}$. An advantage of supervised learning is that the prediction algorithm is assigned an error term, e_t, that is the difference between the actual study outcome explained by the training data and the predicted outcome from the neural network. The error or loss function is the sum of squared errors.

$$\boldsymbol{e_t} = \sum_{\boldsymbol{i}} (\tilde{y}_{it} - y_{it})^{\mathbf{2}}$$

The error function has the advantage that it is continuously differentiable and, thus, simplifies the weight adjustment process during the training phase. Backpropagation optimizes the weights such that the neural network can learn how to correctly assign inputs to outputs by minimizing the error function at every step.

4.3. DECISION TREE

Predictions for the outcome variable across observations are determined by decision tree algorithms. An overview of the most frequently used algorithms can be found in Schapire and Freund (2012) and Sammut and Web (2017). In the present paper, we use the C4.5 algorithm for decision trees (Hall et al., 2009). The C4.5 recursively performs the process of tree building, using information gain to guide the attribute selection process. In addition, this algorithm uses an enhancement of the attribute selection and branching.

Since decisions trees are a very flexible nonparametric machine learning algorithm, they tend to overfit the data. To decrease the variance and to improve the precision of the estimates, we use the bagging (bootstrap aggregation) meta-learning algorithm. Random forest is a method for generating multiple versions of the tree by bootstrapping on the training sample and averaging these to get an improved classifier (Breimann, 1996; 2001). While bagging constructs a large number of (possibly similar) trees with bootstrap samples, the random forest algorithm additionally chooses a random subset of predicting variables before each node is split. This will lead to different, uncorrelated trees from each sample.[2] We applied bagging on the test data before estimating a random forest, therefore bagging with random forest (BRF).

[2] From all tested decision trees (i.a. C4.5, M5p, CART, Decision Stump, RepTree) with all tested meta-learning algorithms (i.a. bagging, random subspace, random committee, classification via regression, random forest), the BRF and C4.5 perform best. Results are available upon request.

4.4. META-ALGORITHM ADABOOST

To combine the predictive powers of the neural network, regression model, and BRF, we use a boosting algorithm. Boosting algorithms evaluate the influence of the individual methods (weak classifiers) and merges the results into a single (strong) classifier. Here the adaptive boosting (AdaBoost) algorithm developed by Freund and Schapire (1997) is applied. The AdaBoost algorithm was originally used to solve character recognition problems, but it also achieved good results solving various classification problems. It is a general method for improving the classification accuracy. The basic idea is to combine the results obtained from various methods into an efficient decision-making rule, so that in our application dropout behavior can be forecasted with better accuracy. On the basis of the calculated forecasts, these methods (described above) are initially weighted equally. In each repetition of the algorithm, the individual weights are adapted according to the distribution in such a way that the resulting classifier has the smallest possible error value. The prediction of the AdaBoost is calculated as the sum of the weighted predictions and improves the prediction accuracy as compared to using any single method. Moreover, since the proposed EDS should be applicable to any German university for more than one time period, the AdaBoost avoids the need to choose a single best working method.

4.5. CHOICE OF IDENTIFICATION THRESHOLD

Each forecasting method estimates a dropout-probability for each student that is between 0 (graduate) and 1 (dropout). Thus, the EDS needs a threshold beyond which, based on the results from the forecast, potential dropouts are defined to be at risk. The choice of the threshold is important for the potential use of an EDS. An EDS has little value in itself unless it is used as a basis for interventions aimed at for instance lowering dropout rates. Thus, the EDS could be used to inform students about the risk of failure. In practical terms, university administration assigns a threshold delineating students at-risk and thus in need of intervention. The lower the threshold, the higher is the rate of correctly predicted dropouts; but at the same time, the rate of correctly identified students decreases, as many students who will not drop out are treated as potential dropouts. This may have a negative impact on the student body's acceptance of the EDS.

One possible solution is to assign the threshold using the average dropout rate of students enrolled in previous terms. The graduate-dropout cohort informs on the previous year's study-success rate thus assigning a benchmark to adapt the EDS to.. As the dropout rate varies between the cohorts, this will result in a margin of error separating probable dropouts from actual dropouts. To test the predictive power of our EDS on previous cohorts, we set this threshold such that the number of identified dropouts coincides with the known number of dropouts in the

test cohort for each semester. This allows distinguishing between the two causes of deviations from the true dropout rate, namely inadequate data and forecasting error resulting from the chosen method. If the EDS is used on current students, the threshold should be based on the average dropout rate of previous cohorts. While the threshold can be modified to either reach more potential dropouts at the cost of a higher misidentification rate (e.g., lower the threshold) or to increase the accuracy of identifications and having at-risk students left unidentified (e.g., increase the threshold). In Section 5 we present the results using two variants of the threshold. The "true" threshold is based on the actual number of dropouts, so that the number of identified at-risk students matches the number of dropouts. The "average" threshold is based on the average dropout rate of the 2010-2012 cohorts. Using later cohorts would bias the threshold, as students might still be enrolled and have the chance to graduate.

4.6. PERFORMANCE

The performance of a machine learning method can be described by its forecasting accuracy, specificity, recall, and precision (Ting, 2011; Powers, 2011). Similar to binary or binomial classification, the task is to classify elements of a given set into two groups. These can be arranged into a 2x2 contingency table or confusion matrix as seen below:

Confusion matrix

	Prediction is dropout	Prediction is graduate
Student is dropout	True positive (t_p)	False negative (f_n)
Student is graduate	False positive (f_p)	True negative (t_n)

For our purposes, a correctly predicted graduate is a student which is correctly rejected as an at-risk student, i.e., a true negative. Consequently, a correctly predicted dropout is correctly identified as an at-risk student, i.e., a true positive. Derived from the confusion matrix, we define our measures of forecasting quality as follows:

Accuracy: $\frac{t_p+t_n}{t_p+f_p+f_n+t_n}$

Precision: $\frac{t_p}{t_p+f_p}$

Recall (sensitivity or true positive rate): $\frac{t_p}{t_p+f_n}$

Specificity (true negative rate): $\frac{t_p}{t_n+f_p}$

Since the aim of the EDS is to identify students at risk, in the present study, besides the accuracy, i.e. the rate of correct predictions, both recall and precision are of particular relevance. Specificity on the other hand measures how accurate graduates are identified and is therefore not as informative for the purpose of the present study—it is not the purpose of the EDS to support graduates but to optimize the allocation of help to those at risk.

Recall, also known as sensitivity or true positive rate, measures how many of the at-risk students are identified, while the precision, also known as positive predictive value, measures how many of the identified students are in fact at risk. Since the true identification threshold is set such that the predicted dropout rate equals the known dropout rate in the test cohort, it follows that the number of false negatives equals the number of false positives, thus $f_p = f_n$. As a result, precision and recall are identical in this study. In the following we focus on accuracy and recall only.

We further illustrate the diagnostic quality of our classifiers by plotting the Receiver Operating Characteristics (ROC) curve. The ROC curve represents specificity and recall in a coordinate system, where recall is plotted on the y-axis and one minus the specificity on the x-axis. Hence the ROC curve depicts relative trade-offs between true positive and false positives. For example the best possible prediction method would yield the point, $(x, y) = (0,1)$, representing 100% recall (no false negatives) and 100% specificity (no false positives). A random guess is on the 45°-line (50% false negatives and 50% false positives).

5. Results: Forecasting student dropout

5.1. Regression results

In order to get a better understanding of the data and the factors that are related to dropout, we show and discuss the results of the pooled sample logit model. The models discussed describe correlations only and they cannot be interpreted as causal effects. Table 4a shows the results of the logit models using the student performance and demographic data from the first four semesters of the SU (cf. Table 4a, columns (1) to (5)). Note that we only use the training data and report the odds ratios; we keep the specifications of the models simple, as we want to point out correlations in the data between the dependent and explanatory variables so as to find good predictors for student dropout. More sophisticated modelling to identify causal effects is beyond the scope of the current paper. Here, the goal of the paper is to combine various methods and to build a self-adjusting turnkey application.

The binary dependent variable has a value of 0 for graduation and has a value of 1 for dropout. Recall, that “dropouts” are students who leave the university without a degree, regardless of whether the student continues her studies at another university immediately after dropping out

or at a later date. The latter information is not available. The number of observations in Table 4a drops by 47% from the first (12,728) semester to the fourth (6,693) due to students dropping out. It follows that the coefficients in the columns are not directly comparable, as the size and composition of the sample changes every semester.

We look first at the fit of the regression model as described by the AIC. Using only the demographic information available at the time of enrolment, the AIC is 15,658 (Table 4a, column (1)). Incorporating the performance data from the first semester reduces the AIC to 11,988 (Table 4a, column (2)). The AIC drops to 7,448 in the second semester and 4,178 in the fourth semester. Thus as expected, the model fit improves in later semesters.

Note that the estimates in column (1) only reflect the demographic variables, i.e., information that is available at the time of enrollment. At the time of enrollment, being male increases the odds of dropping out by 60%. Age at enrollment is also positively correlated with dropping out. Immigrants have a higher dropout risk as compared to native students (baseline category), and first generation immigrants have a higher dropout risk than second generation immigrants. The odds at the time of enrollment are 22% higher for first generation immigrants and 46% for second generation immigrants. Students with a high school degree allowing them entrance to a university of applied sciences (Fachhochschulreife) are less likely to graduate as compared to students with a general university entrance qualification (Allgemeine Hochschulreife). The effect from the grade of the high school degree (Abiturnote) is negative and statistically significant.[3] The coefficient on the dummy variable for private health insurance is only marginally significant. And, it is significantly more likely that lateral entrants graduate at a SU.

Some, if not most of the demographic variables lose statistical significance when controlling for the performance data available after the first semester. Gender for instance has a sizable and significant effect in column (1), the effect becomes smaller over time and in column (5), using the information in the 4th semester, the effect is substantially diminished. In semesters 5 and 6 (not reported here), the effect is even smaller and insignificant. Most notably, immigration status is no longer significant once achievement data becomes available; the magnitude of deviation in precision caused by imprecisely estimated effect is also reduced. Thus, the rich student data available at the time of enrollment is only valuable for identifying at-risk students at the very beginning of their studies. Even as early as after the first semester, performance data picks up the most relevant information (Stinebrickner & Stinebrickner, 2012; 2014). One exception is the dummy variable for private health insurance, where a significant correlation is still estimated in the 4th semester. Controlling for academic performance, students who have private health insurance are more likely to graduate than those who have public health insurance. As stated

[3] In the German grading system (school and tertiary education), the grading scale ascends from highest to lowest in achievement, i.e., a 1 is excellent and a 5 indicates failure.

above, students with private insurance are more likely to come from high-income families or have parents who are civil servants. Thus, even controlling for academic performance, family background partly explains dropping out even in later semesters. The performance variables (Average Grade, No Exam, Not Participated, and Failed Exam) are negatively associated with study success (columns 2-7). Not surprisingly, failed exams and non-participation in exams are good predictors for dropouts. Note that the explanatory power of the performance indicators is decreasing in successive semesters, but academic performance in previous semesters continues to have explanatory power in later semesters (results are not reported in the Tables). This is even true for performance variables in the first semester. Thus, students who do not drop out after having performed poorly in the first semester, still face a higher probability of not finishing their studies. In addition, the number of credit points (CP) is also a statistically significant predictor of the dependent variable.

Table 4a: Effects of performance and demographic variables on dropout prediction (SU)

Dependent variable: student drops out (1=yes; 0=no); Logistic Regression (Odds Ratio)					
	(1) Enrollment	(2) 1st Semester	(3) 2nd Semester	(4) 3rd Semester	(5) 4th Semester
Gender (1=male; 0=female)	1.612** (0.000)	1.309** (0.000)	1.313** (0.000)	1.288** (0.000)	1.200* (0.035)
Age at enrollment	1.076** (0.000)	1.048** (0.000)	1.070** (0.000)	1.046** (0.000)	1.077** (0.000)
First generation immigrant (1=yes; 0=no)	1.462** (0.000)	1.083 (0.222)	1.054 (0.300)	1.113 (0.959)	1.072 (0.860)
Second generation immigrant (1=yes; 0=no)	1.222** (0.000)	1.074 (0.410)	1.082 (0.661)	1.004 (0.429)	1.019 (0.661)
City of entrance qualification (1= city of university; 0=else)	1.337** (0.000)	1.084 (0.170)	1.143+ (0.076)	1.199* (0.036)	1.332** (0.005)
Univ. of Appl. Sciences entrance qualification (1=yes; 0=no)	2.033** (0.003)	1.611+ (0.085)	1.132 (0.683)	0.878 (0.714)	0.888 (0.776)
Restricted university entrance qualification (1=yes; 0=no)	0.822 (0.432)	1.173 (0.612)	0.891 (0.738)	0.897 (0.786)	0.903 (0.828)
Foreign university entrance qualification (1=yes; 0=no)	1.055 (0.695)	0.869 (0.387)	0.944 (0.769)	0.887 (0.589)	0.968 (0.898)
Grade of university entrance Qualification	1.368** (0.000)	1.056+ (0.081)	1.000 (0.996)	0.947 (0.220)	0.944 (0.281)
Health insurance (1=private; 0=public)	0.875+ (0.097)	0.694** (0.000)	0.775* (0.042)	0.670** (0.004)	0.718* (0.044)
# of enrolled study programs	0.879** (0.000)	0.934** (0.000)	0.978 (0.292)	1.048+ (0.070)	1.087** (0.005)
Lateral entrants	0.399** (0.000)	0.570** (0.000)	0.474** (0.000)	0.498** (0.000)	0.462** (0.000)
# of semesters at prev. university	1.089** (0.000)	1.061** (0.000)	1.059** (0.000)	1.047** (0.000)	1.030+ (0.052)
Average grade current semester		1.658** (0.000)	1.381** (0.000)	1.193** (0.006)	1.352** (0.000)
Average CPs current semester		0.948** (0.000)	0.932** (0.000)	0.936** (0.000)	0.942** (0.000)
No exam taken current semester		17.142** (0.000)	5.988** (0.000)	3.470** (0.000)	4.180** (0.000)
# of exams current semester not participated in		1.464** (0.000)	1.272** (0.000)	1.183** (0.000)	0.993 (0.892)
# of failed exams current semester		1.380** (0.000)	1.252** (0.000)	1.175** (0.000)	1.274** (0.000)
Constant	0.231** (0.000)	0.177** (0.000)	0.120** (0.000)	0.262** (0.000)	0.175** (0.000)
Performance previous semesters:					
Previous average grades			YES	YES	YES
Previous average CPs			YES	YES	YES
Previous # of exams			YES	YES	YES
Prev. # of not participated. exams			YES	YES	YES
Previous # of failed exams			YES	YES	YES
Important Exams		YES	YES	YES	YES
AIC	15,657.54	11,988.42	7,448.05	5,757.76	4,178.26
N	12,728	12,728	9,228	8,015	6,693

*Notes: +p<0.1. *p<0.05. **p<0.01. Standard errors in parentheses.*

Table 4b shows the results of the logit estimation using the student performance and the demographic data from the first to fourth semester at the PUAS. The number of observations drops from 7,077 in the first semester to 5,448 students in the fourth semester. Similar to the SU, the model fit improves between the first and fourth semesters. A presumption is that the tuition fees—that are absent at the SU—accelerate the decision to drop out. The results are comparable with the results from the SU—especially with regard to the strength and direction of the coefficients on the performance-related data.

Table 4b: Effects of performance and demographic variables on dropout prediction (Private University of Applied Sciences)

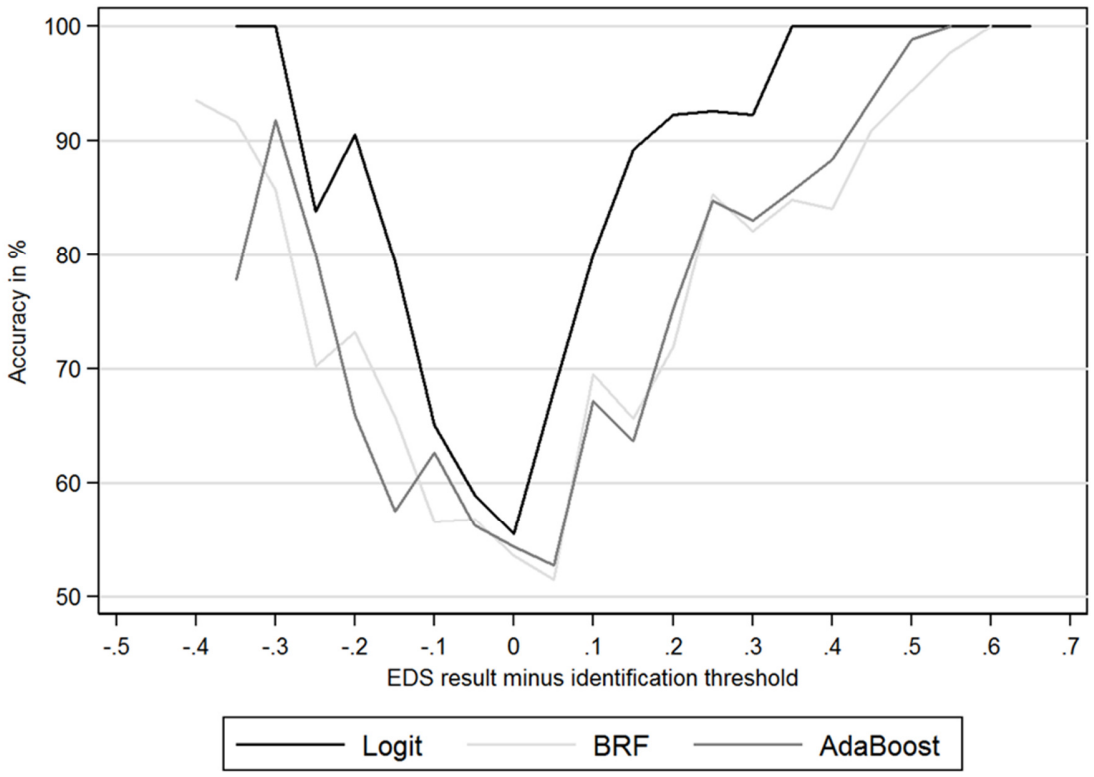
Dependent variable: student drops out (1=yes;0=no); Logistic Regression (Odds Ratio)					
	(1) Enrollment	(2) 1st Semester	(3) 2nd Semester	(4) 3rd Semester	(5) 4th Semester
Gender (1=male; 0=female)	1.576** (0.000)	1.279** (0.002)	1.064 (0.556)	0.968 (0.808)	0.865 (0.387)
Age at enrollment	1.045** (0.000)	1.030** (0.001)	1.031** (0.006)	1.035* (0.016)	1.072** (0.000)
First generation immigrant (1=yes; 0=no)	1.602** (0.976)	0.904 (0.002)	0.795 (0.001)	0.827 (0.000)	0.763 (0.045)
Second generation immigrant (1=yes; 0=no)	1.003 (0.001)	0.777** (0.387)	0.692** (0.139)	0.569** (0.323)	0.712* (0.248)
City of entrance qualification (1= city of university; 0=else)	1.197* (0.045)	1.006 (0.932)	0.868 (0.138)	0.892 (0.342)	0.856 (0.306)
Univ. of appl. sciences entrance qualification (1=yes; 0=no)	2.136** (0.000)	1.314** (0.000)	1.122 (0.228)	1.194 (0.139)	1.040 (0.795)
Restricted university entrance qualification (1=yes; 0=no)	3.148** (0.000)	2.143** (0.000)	1.865* (0.035)	1.980+ (0.083)	1.922 (0.199)
Foreign university entrance qualification (1=yes; 0=no)	5.365** (0.000)	3.388** (0.000)	1.414 (0.280)	0.788 (0.544)	0.320* (0.017)
Lateral entrants	1.468** (0.000)	1.345** (0.000)	1.212** (0.000)	1.142* (0.046)	1.110 (0.202)
Average grade current semester		2.141** (0.000)	1.357** (0.001)	1.352** (0.010)	0.999 (0.995)
Average CPs current semester		0.974** (0.008)	0.901** (0.000)	0.893** (0.000)	0.915** (0.000)
No exam taken current semester		17.281** (0.000)	10.948** (0.000)	3.620** (0.002)	2.645* (0.038)
# of exams current semester not participated in		1.333** (0.000)	1.281** (0.000)	1.140* (0.027)	1.330** (0.000)
# of failed exams current semester		1.339** (0.000)	1.082* (0.040)	1.108* (0.024)	1.230** (0.000)
Constant	-0.055 (0.108)	-0.065+ (0.070)	0.461** (0.000)	0.670** (0.000)	0.704** (0.000)
Type of study program	YES				
Previous performance:					
Previous average grades			YES	YES	YES
Previous average CPs			YES	YES	YES
Previous without exam			YES	YES	YES
Prev. # of exams not participated in			YES	YES	YES
Previous # of failed exams		YES	YES	YES	YES
Important exams			YES	YES	YES
AIC	3,983.58	6,108.78	3,719.80	2,525.38	1,703.21
N	7,077	7,077	6,329	5,847	5,448

*Notes: +p<0.1. *p<0.05. **p<0.01. Standard errors in parentheses.*

5.2. ACCURACY OF CLASSIFIERS

Before we describe the results for the different classifiers, Figure 1 shows the forecast accuracy of the logit, BRF, and AdaBoost models for the SU. The results for the private university are very similar and not reported. Each method estimates a dropout probability for each student between 0 (graduate) and 1 (dropout). Forecasted dropouts with probabilities close to 0 or 1 are accurate. Forecasts close to the identification threshold are uncertain. Figure 1 illustrates the accuracy. As expected, close to the true threshold, the proportion of correct predictions (among all predictions) is lowest. This is true for all classifiers; however, when comparing, the AdaBoost outperforms the logit and the random forest, albeit not over the entire range of observations.

Figure 1: Accuracy of the EDS



Furthermore, the predictive accuracy is tied to the underlying dropout rate, which determines the threshold. As explained above, we work with two thresholds. First, to test the accuracy of our procedure, we use the rate of actual dropouts per cohort in each semester to define the true threshold. Second, to simulate the performance of the EDS in a more realistic setting, we use the number of dropouts in the previous cohorts to define the average threshold. Whether the threshold is set using the number of rate of dropouts per cohort in each semester or the average dropout rate of previous cohorts (which would be done, when implementing the EDS at

universities), is of consequence. With a dropout rate of 0% or 100% the accuracy is 100%, regardless of the chosen threshold. This is different with dropout rates of 50%. Accuracy is increasing as the dropout rates increase or decrease, as it becomes easier to classify observations correctly. Thus, the accuracy of the EDS depends on the dropout rate of the university and the cohort. Since dropout rates tend to decrease over time—dropouts leave university earlier than graduates—accuracy of prediction is also expected to increase. And in fact, accuracy increases with successive semesters at both universities.

5.2.1. Logit model

Table 5a summarizes the forecasting quality measures of the logit model. As expected, quality of prediction increases over time. This applies to all quality measures. For instance, recall (how many of the at-risk students are identified) at the SU increases from about 71% in the first semester to 80% in the fourth semester. At the PUAS, recall for the 1st and 4th semesters was 69% and 78%, respectively.

Table 5a: Performance of the logit model based on the dropout rate of the test cohort

Logit	State University					Private University of Applied Sciences				
	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.
Accuracy[a]	63.47	76.15	81.80	86.55	89.56	66.53	82.67	88.47	91.36	93.76
Recall[a]	61.67	70.80	74.05	78.71	79.94	48.95	68.68	74.30	76.68	78.26
No. of graduates	1,112	1,039	1,027	1,015	9,92	976	976	974	969	961
No. of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	724	830	883	916	922	733	860	902	918	926
Incorrectly predicted dropouts	388	209	144	99	70	243	116	72	51	35
Correctly predicted dropouts	626	514	411	377	279	233	261	211	171	126
Incorrectly predicted graduates	389	212	144	102	70	243	119	73	52	35
Correctly predicted graduates[a]	65.11	79.88	85.98	90.25	92.94	75.10	88.11	92.61	94.74	96.36
Incorrectly predicted dropouts[a]	34.89	20.12	14.02	9.75	7.06	24.90	11.89	7.39	5.26	3.64
Correctly predicted dropouts[a]	61.67	70.80	74.05	78.71	79.94	48.95	68.68	74.30	76.68	78.26
Incorrectly predicted graduates[a]	38.33	29.20	25.95	21.29	20.06	51.05	31.32	25.70	23.32	21.74

Notes: [a] In percent

As explained above, both accuracy and recall are based on an identification threshold that in this paper matches the actual rate of dropouts in the test cohort (true threshold). Therefore, the values reported here are the optimal values of accuracy and recall, conditional on the given variables and chosen method.

If the true number of dropouts is unknown—as is the case in real applications of the EDS—the identification thresholds have to be based on the dropout rates of previous cohorts (average threshold, cf. Table 5b). This is expected to reduce the forecast performance. In our case, the accuracy of the SU is between 73% in the first semester and 82% in the fourth semester. The recall is between 84% in the first semester and 92% in the fourth semester. There are high recall rates because the number of dropouts in the test cohort is below average. However, the high fraction of correctly identified dropouts (recall) comes at the cost of too many false-positives, i.e. incorrectly predicted-dropouts (accuracy).

Table 5b: Performance of the logit model based on the average dropout rate

Logit	State University					Private University of Applied Sciences				
	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.
Accuracy[a]	63.19	72.75	76.36	81.33	81.73	66.87	80.53	86.96	91.11	93.05
Recall[a]	69.06	83.61	87.57	90.81	91.69	48.95	72.37	78.17	79.37	81.37
No. of graduates	1,112	1,039	1,027	1,015	992	976	976	974	969	961
No. of dropouts	1,015	726	555	479	349	476	380	284	223	161
Correctly predicted graduates	688	699	736	785	779	738	817	872	909	913
Incorrectly predicted dropouts	424	340	291	230	213	238	159	102	60	48
Correctly predicted dropouts	743	631	499	442	323	233	275	222	177	131
Incorrectly predicted graduates	272	95	56	37	26	243	105	62	46	30
Correctly predicted graduates[a]	61.87	67.28	71.67	77.34	78.53	75.61	83.71	89.53	93.81	95.01
Incorrectly predicted dropouts[a]	38.13	32.72	28.33	22.66	21.47	24.39	16.29	10.47	6.19	4.99
Correctly predicted dropouts[a]	73.20	86.91	89.91	92.28	92.55	48.95	72.37	78.17	79.37	81.37
Incorrectly predicted graduates[a]	26.80	13.09	10.09	7.72	7.45	51.05	27.63	21.83	20.63	18.63

Notes: [a] In percent

5.2.2. Bagging with random forest and neural network

Next we base the prediction on machine learning methods. In line with similar analyses found in the literature, there is not much difference in the forecast accuracy between the methods tested: regression model, neural net, and random forest. Furthermore, we also confirm the superior performance of BRF. This method outperformed the others in terms of forecasting accuracy by 0.88 - 2.93% (SU) and 0.88 - 1.03% (PUAS) (Tables 6a and 6b)[4].

[4] The results of all tested methods are available on request. For the sake of brevity, we only report accuracy and recall.

Table 6a: Performance of the BRF

BRF	State University					Private University of Applied Sciences				
	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.
Accuracy (true threshold)	65.02	78.41	82.43	86.95	88.96	63.36	82.52	88.79	91.36	93.58
Accuracy (average threshold)	65.35	74.84	77.24	80.79	81.43	63.71	82.45	87.52	91.44	92.87
Recall (true threshold)	63.55	73.83	75.14	79.96	79.08	44.54	68.95	75.35	77.13	78.26
Recall (average threshold)	71.43	86.36	89.01	90.40	91.40	44.33	76.05	80.28	80.72	81.37

Notes: All results in percent.

Table 6b: Performance of the neural net

Neural net	State University					Private University of Applied Sciences				
	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.
Accuracy (true threshold)	62.53	72.75	81.54	85.27	86.35	66.67	82.49	88.70	80.00	94.76
Accuracy (average threshold)	63.38	73.14	76.86	78.78	79.34	66.74	81.34	87.54	82.44	92.51
Recall (true threshold)	60.69	47.80	70.09	73.28	72.78	49.13	68.51	74.44	92.79	80.42
Recall (average threshold)	69.06	80.99	84.86	87.06	83.67	48.69	74.03	80.45	91.65	81.82

Notes: All results in percent.

Using the average identification threshold, accuracy of the BRF is 75% in the first semester and rises to 81% in the fourth semester. The accuracy of the neural net is 73% in the first semester and 79% in the fourth semester. Since the dropout rate in the test cohort is below average, we expect recall to be high. And in fact, recall is between 86% in the first semester and 91% in the fourth semester when using the BRF and between 81% and 83% when using the neural net.

To further illustrate the diagnostic quality of our classifiers, we plot the ROC curves in Figures 2a and 2b. First, all methods perform substantially better than a random guess. Second, prediction power improves with more information in higher semesters (the area below the ROC curve increases). In addition, the ranking of the methods differs by university and semester. This

is our motivation for combining the predictive power of neural networks, BRF, and logit model using the AdaBoost algorithm in Section 5.2.3.

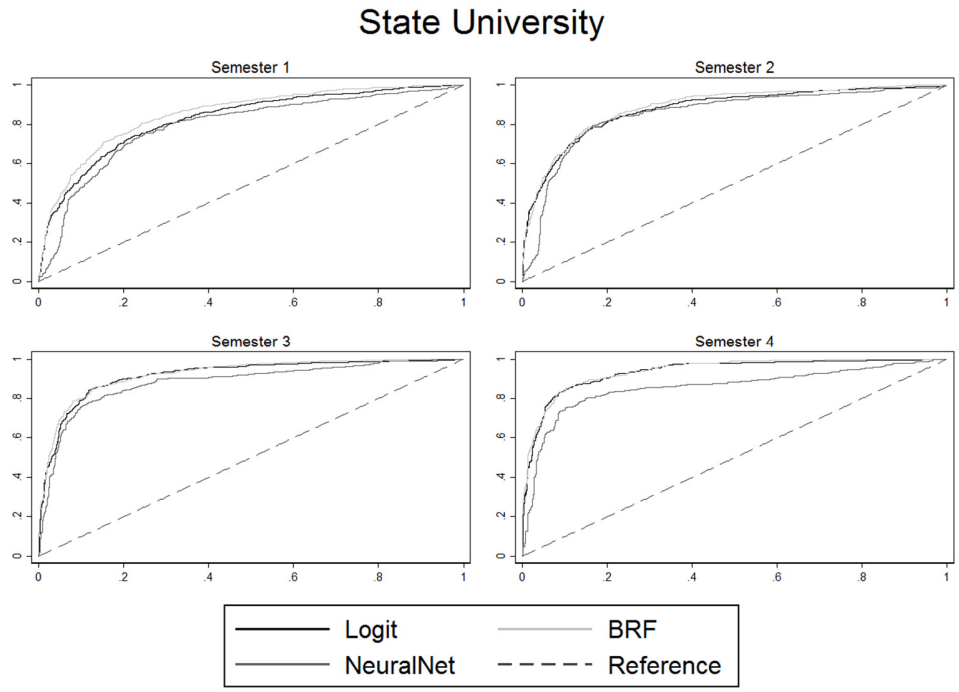


Figure 2a: ROC curves—State University

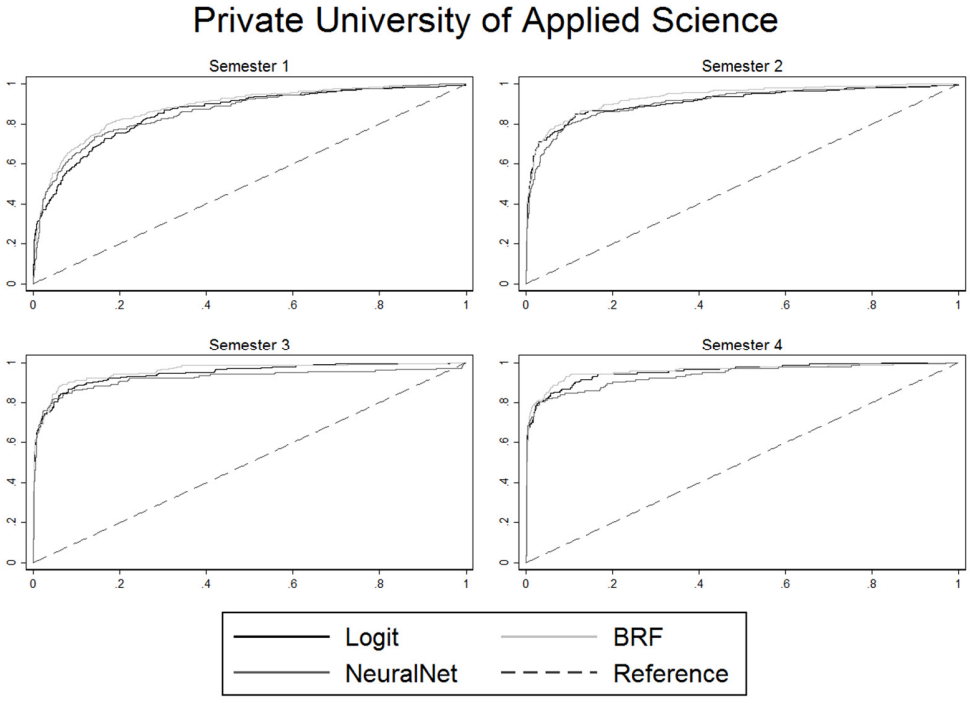


Figure 2b: ROC curves—Private University of Applied Science

While not the focus of our study, we use the values from the information gain in the random forest, using data from the first semester, to assess the relative importance of the input variables. In Figure 3, we differentiate between demographic variables (right) and performance variables (left) and between the two universities.

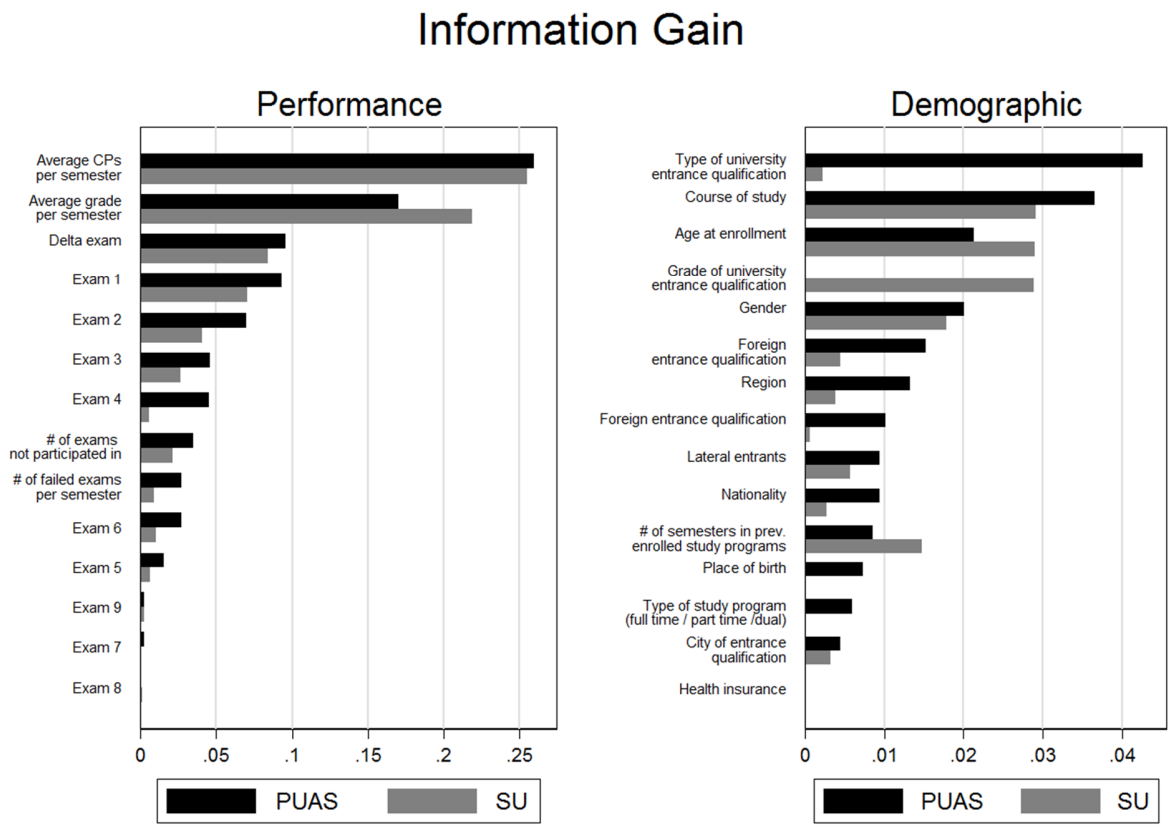


Figure 3: Information gain: BRF, first semester

It is apparent that performance data better predicts dropouts than demographic data at both universities. This confirms the results from the logit model. In particular, the pace of study (avg. CP/semester), the average grade (avg. Grade/semester) as well as the most important exam all have a high degree of explanatory power. Comparing SU and PUAS, the five most important predictor variables are identical for both universities and the difference in information gain is small.

A substantial yet expected difference between the two universities is that the variable "type of entrance degree" is almost irrelevant at the SU with a value of 0.008, while it is the most important demographic variable at the PUAS with an information gain of 0.043.

5.2.3. AdaBoost

Table 7 summarizes the forecast accuracy of the AdaBoost, our preferred classifier. It shows the results for the SU and the PUAS; there are noticeable differences in the levels of forecast accuracy, recall, and precision between the two institutions. However, for both institutions, prediction accuracy increases as early dropouts leave the university. Thus, not surprisingly, regular updates from end-of-semester performance data improve the prediction results.

Table 7: Performance of the AdaBoost

AdaBoost	State University					Private University of Applied Sciences				
	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.
Accuracy (true threshold)	67.65	78.53	82.43	87.62	89.63	67.17	84.49	89.70	81.95	95.51
Accuracy (average threshold)	67.28	75.35	78.07	82.13	82.18	67.10	83.18	88.95	85.85	93.45
Recall (true threshold)	65.81	73.83	74.95	80.58	79.94	49.78	72.10	76.69	93.50	83.22
Recall (average threshold)	73.20	86.91	89.91	92.28	92.55	49.34	77.35	83.83	92.88	85.31

Notes: All results in percent.

If the identification threshold for the SU test cohort is determined based on the average dropout rate instead of the true dropout rate, the accuracy of the AdaBoost is 75% in the first semester and 82% in the fourth semester. The recall is 87 % in the first semester and 93% in the fourth semester. Comparing the performance measures with the true threshold and the average threshold, it turns out that, as before, accuracy is lower and recall is higher when using the average threshold; because the dropout rate of the test cohort is below average. That leads to over identification of potential dropouts.

When implementing the EDS at universities, the selection of relevant information is an important issue. While machine learning can certainly handle large data sets, data cleaning is a resource intensive task. Thus for reasons of efficiency, it is worthwhile to think about restricting the required variables. Moreover, data privacy rules might aggravate the implementation, in particular when the system uses demographic information. While this issue cannot be fully resolved in this paper, we nevertheless briefly address it.

First, we focus on the relevance of using information collected at the time of enrollment—namely, demographic data. In the first semester, before having taking any exams, about 21% of all dropouts in the sample left the PUAS and 28.5% left the SU (cf. Table 7). The forecast

accuracy is about 68% for both institutions but with distinct differences in the dropout detection rate. At the PUAS, successful students are better predicted than at-risk students, while at-risk students are better predicted than successful students at the SU (this pattern of results is consistent throughout all semesters).
Alternatively, one could use performance data only (cf. Table 8).

Table 8: Performance of the AdaBoost using only academic performance data

AdaBoost	State university				Private university of applied sciences			
	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.
Accuracy	76.60	81.42	86.61	88.52	83.64	90.45	92.44	94.76
Recall	71.49	73.51	79.12	77.94	69.89	78.20	79.02	80.42

Notes: All results in percent.

At both universities, forecasts are only marginally enhanced when using both demographic and performance data as opposed to just performance data. Thus, the use of student demographic data is only beneficial if no performance data are available, as performance data and the demographic data are correlated. Forecasts using performance data from the end of the first semester are only marginally enhanced by the addition of demographic data. The additional forecast accuracy gained from the demographic data is reduced with each new update from student performance data following the end of a semester. This is important information when planning, for instance, interventions based on the forecasting system. Only if successful interventions take place right at the beginning of the student's career, before students take the first exams, is demographic data an important source of information. Once performance data is available after the first semester, rich demographic data adds little additional information to the forecasting model. After the first semester, the percentage of correctly predicted dropouts at the SU is 71% when using academic performance data only and 74% when using demographic and achievement data.

5.2.4. Robustness of performance with respect to assignment of test data

As described above, data from students who matriculated in 2010 (SU) and 2012 (PUAS) was used as testing data. We selected these years to ensure that almost all students have completed their studies or dropped out. In later years, the proportion of the students who are still enrolled increases, as Table 9 shows. The differences between the SU and the PUAS are remarkable. Students at the SU are enrolled for much longer than students at the PUAS. The proportion of students who are enrolled in a (3 year) bachelor program in the 2010 -academic year and who are still enrolled in the winter term 2017/18 is less than 3% at the private university

and almost 12% at the SU. The difference is at least partly explained by student fees. While student fees are in place for most study programs at the PUAS, there have been no fees at SUs since the winter term of 2011/12. Hence aside from administration fees, enrollment is free and offers some benefits for students.

Table 9: Proportion of students in an academic year who are still enrolled in the winter term of 2017/2018

Year	SU	PUAS
2009	8.23%	2.63%
2010	11.74%	2.42%
2011	16.26%	4.76%
2012	20.56%	6.94%
2013	30.57%	16.07%
2014	48.38%	35.72%
2015	58.79%	64.12%
2016	76.96%	83.39%
2017	97.72%	95.27%

In Table 10 we summarize our performance indicators accuracy and recall for different test data with BRF. Testing with data from later years tends to improve performance, in particular for early semesters at SU and in particular recall. This is explained by the composition of cohorts from later years. Recall that only dropouts and graduates are included in the data. In later cohorts, the graduates and dropouts are early dropouts and fast graduates. Other students of that cohort are still matriculated and not included in the data. Thus, our predictor performs well, as it only has to predict students at the tails of the distribution: early dropouts and fast graduates.

Table 10: Performance of BRF—sensitivity to test data

BRF	State University					Private University of Applied Sciences				
	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.	Enroll-ment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.
Accuracy (2010)[a]	65.02	78.41	82.43	86.95	88.96	61.75	82.03	88.00	90.28	92.89
Accuracy (2011)	59.35	81.07	83.99	87.63	88.71	62.31	80.46	87.45	90.37	93.92
Accuracy (2012)	69.07	83.02	86.90	89.25	90.19	63.36	82.52	88.79	91.36	93.58
Accuracy (2013)	74.41	86.53	88.65	89.57	89.86	64.80	82.78	88.93	92.99	95.98
Recall (2010)	63.55	73.83	75.14	79.96	79.08	42.02	65.69	69.08	67.97	69.91
Recall (2011)	66.88	81.73	82.01	84.80	83.22	39.90	61.54	69.23	70.93	77.94
Recall (2012)	76.79	86.18	87.39	88.54	87.06	44.54	68.95	75.35	77.13	78.26
Recall (2013)	84.09	91.15	91.33	91.28	89.50	50.32	71.07	75.00	80.43	84.92

Notes: [a] The year denotes the year of the matriculation of the test cohort. All results in percent.

6. CONCLUSIONS AND OUTLOOK

University attrition is an important issue for education policy. Student attrition is costly for all involved parties; resources spent on educating students and the effort and time spent by the student in the university system are both of limited economic value when not accompanied by a graduating certificate. Thus, it is in everybody's interest to optimize (prevent or speed up) student attrition through diagnosis and intervention. This paper develops and tests a forecasting system for the early detection of university dropouts. The forecasting system is based on administrative data available at the universities; it is self-adjusting and can be used to identify students at risk and to allocate students into support interventions at universities.

In addition to using traditional regression analyses to predict dropouts, we also employ machine learning algorithms which do not rely on complex model building and are self-adjusting whenever new data becomes available. Instead of relying on a single method, we use the AdaBoost algorithm to combine the various methods employed. This reduces the disadvantages inherent in using any single method and caused by the heterogeneity of study programs and student body compositions at the different universities.

In the present paper, we use data from a state and a private university to develop and test the model. The predictive power of our preferred method, the AdaBoost, is strong. The accuracy of the results improve with increasing semesters and are only marginally improved by the additional use of demographic data. Using only demographic data available at the time of enrollment, our early detection system already correctly predicts 67% of dropouts at the SU; prediction accuracy increases to 80% in the fourth semester. The corresponding numbers for the PUAS are only 50% at the time of enrollment and 83% in the fourth semester. Moreover, using the rich demographic data available, does not substantially improve the performance accuracy once performance data becomes available.

The advantage of the EDS presented is that after having identified students at risk, it can serve as a basis for an early intervention system to either prevent dropouts or to even speed up a student's decision to drop out. In this way, the public and private costs associated with attrition can be reduced by implementation of an EDS as a starting point for allocating intervention support to students at-risk and for testing the effectiveness of student intervention.

REFERENCES

Arcidiacono, P., Aucejo, E., Maurel, A. & Ransom, T. (2016) College Attrition and the Dynamics of Information Revelation. *NBER Working Papers - National Bureau of Economic Research*.

Arulampalam, W., Naylor, R.A. & Smith, J.P. (2005) Effects of in-class variation and student rank on the probability of withdrawal: cross-section and time-series analysis for UK university students. *Economics of Education Review*, 24, 251-62.

Bayer, J. et al. (2012) Predicting Drop-Out from Social Behaviour of Students. *International Educational Data Mining Society*.

Bean, J.P. (1983) The Application of a Model of Turnover in Working Organizations to the Student Attrition Process. *The Review of Higher Education*, 6, 129-48.

Berger, M., Galonska, C. & Koopmans, R. (2004) Political Integration by a Detour? Ethnic Communities and Social Capital of Migrants in Berlin. *Journal of Ethnic and Migration Studies*, 30, 491-507.

Bound, J., Lovenheim, M.F. & Turner, S. (2010) Why Have College Completion Rates Declined? An Analysis of Changing Student Preparation and Collegiate Resources. *American Economic Journal: Applied Economics*, 2, 129-57.

Bowen, W., Chingos, M. & McPherson, M. (2009) *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton: Princeton University Press.

Brandstätter, H. & Farthofer, A. (2002) Studienerfolgsprognose – konfigurativ oder linear additiv? [Predicting student success – Configurational or linear additive?]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 23, 381-91.

Breimann, L. (1996) Bagging Predictors. *Machine Learning*, 24(2), 123-40.

Breimann, L. (2001) Random Forests. *Machine Learning*, 45(1), 5-32.

Burrus, J., Elliott, D., Brennemann, M. & Markle, R. (2013) Putting and Keeping Students on Track: Toward a Comprehensive Model of College Persistence and Goal Attainment. *ETS Research Report Series*.

Danilowicz-Gösele, K., Meya, J., Schwager, R. & Suntheim, K. (2014) Determinants of students success at university. *discussion papers*.

Dekker, G.W., Pchenenizkiy, M. & Vleeshouwers, J.M. (2009) Predicting Students Drop out: A Case Staudy. *Educational Data Mining 2009*.

Er, E. (2012) Identifying At-Risk Students Using Machine Learning - Techniques: A Case Study with IS 100. *International Journal of Machine Learning and Computing, Vol 2, No 4*, 476-80.

Gaebel, M., Hauschildt, K., Mühleck, K. & Smidt, H. (2012) Tracking Learners' and Graduates' Progression Paths. TRACKIT. *EUA Publications*.

Hall, M. et al. (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Heublein, U. (2014) Student Drop-out from German Higher Education Institutions.. *European Journal of Educationc. Research, Development and Policy*, 49(4), 497-513.

Heublein, U. & Burkhart, S. (2013) Bildungsinländer 2011 - Daten und Fakten zur Situain von von ausländischen Studierenden [Educational natives 2011 – Data and facts describing the situation of international students]. Bonn.

Humpert, A. & Schneiderheinze, K. (2002) *Stichprobenziehung für telefonische Zuwandererumfragen. Praktische Erfahrungen und Erweiterung der Auswahlgrundlage. [Sampling for telephone surveys of immigrants. Experience and broadening of the sampling frame]*. Münster: Waxmann.

Kemper, L., Vorhoff, G. & Wigger, B.U., 2018. Predicting Student Dropout: a Machine Learning Approach. *working paper*.

Kotsiantis, S.B., Pierrakeas, C.J. & Pintelas, P.E. (2003) Preventing Student Dropout in Distance Learning - Using Machine Learning Techniques. *KES 2003*, 267-74.

Larsen, M.L. et al. (2013) Dropout Phenomena at Universities: What is Dropout? Why does Dropout Occur? What Can be Done by the Universities to Prevent or Reduce it? A systematic review. *Danish Clearinghouse for Educational Research*.

Michael, J. (2007) Anredebestimmung anhand des Vornamens. *c't*, 17/2007, 182-83.

Michael, J. (2016) Name Quality Pro (to be published). *(available from the author; mail to: namequality.pro@gmail.com)*.

Minaei-Bidgoli, B., Kortemeyer, G. & Punch, W.F. (2004) Enhancing Online Learning performance: An Application of Data Mining Methods. *Proceedings of the Seventh IASTED International Conference on Computers and Advanced Technology in Education*.

Minsky, M. & Papert, S. (1969) Perceptrons: An Introduction to Computational Geometry. *Institute of Technology: Massachusetts.*.

Mucherino, A., Papajorgji, P.J. & Pardalos, P.M. (2009) k-Nearest Neighbor Classification. *Data Mining in Agriculture. Springer Optimization and Its Applications*, 34, 109-13.

Nghe, N.T., Janecek, P. & Haddaway, P. (2007) A Comparative analysis of techniques for predicting academic performance. *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, FIE´0, 37th Annual IEEE*.

Pascarella, E.T. & Terenzini, P.T. (1979) Interaction Effects in Spady's and Tinto's Conceptual Models of College Dropout. *Sociology of Education*, 52, 197-210.

Powers, D.M.W. (2011) Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning* , 2(1), 37-63.

Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning 1*, 81-106.

Sammut, C. & Webb, G. (2017) *Encyclopedia of Machine Learning and Data Mining*. New York: Springer US.

Santana, M. et al. (2015) A Predicitive Model for Identifying Students with Dropout Profiles in Online Courses. *working paper*.

Sara, N.-B., Halland, R., Igel, C. & Alstrup, S. (2015) High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence*, 319-24.

Schapire, E. & Freund, Y. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Science*, 55, 119-39.

Schapire, R.E. & Freund, Y. (2012) *Boosting - Foundations and Algorithms*. Massachusetts: Institute of Technology.

Statistisches Bundesamt, DZHW-Berechnungen. (2015) Studierendenstatistik [DZHW student statistics].

Statistisches Bundesamt. (2015) Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund – Ergebnisse des Mikrozensus 2015 [Population and employment. Immigrants – Results from the Mikrozensus].

Stinebrickner, T. & 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, 2163-84.

Stinebrickner, T. & Stinebrickner, R. (2012) Learning about Academic Ability and the College Dropout Decision. *Journal of Labor Economics*, 32, 707-48.

Stinebrickner, T. & Stinebrickner, R. (2013) A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout. *Review of Economic Studies*, 81, 426-72.

Stinebrickner, T. & Stinebrickner, R. (2014) Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model. *Journal of Labor Economics*, 32, 601-44.

Ting, K.M. (2011) Precision and Recall. In C. Sammut, Webb & G., eds. *Encyclopedia of Machine Learning*. Springer US. 781 & 901.

Tinto, V. (1975) Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45, 89-125.

Trapmann, S., Hell, B., Weigand, S. & Schuler, H. (2007) Die Validität von Schulnoten zur Vorhersage des Studienerfolgs - eine Metaanalyse. *Zeitschrift für pädagogische Psychologie*, 21, 11-27.

Werbos, P. (1974) *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Cambridge, MA: Harvard University.

Wiers-Jenssen, J., Stensaker, B. & Grogaard, J.B. (2002) Student satisfaction: towards an empirical deconstruction of the concept. *Quality in Higher Education*, 8, 183-95.

Xenos, M. (2004) Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education Journal*, 345-59.

Yukselturk, E., Ozekes, S. & 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*, 118-33.

Zhang, Y., Oussena, S., Clark, T. & Kim, H. (2010) Use Data Mining to Improve Student Retention in Higher Education - a Case Study. *Proceedings of the 12th International Conference on Enterprise Information Systems, Volume 1, DISI, Funchal, Madeira, Portugal, June 8 - 12, 2010*.