Schumpeter School
of Business and Economics

SCHUMPETER DISCUSSION PAPERS

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

Johannes Berens
Simon Oster
Kerstin Schneider
Julian Burghoff

The Schumpeter Discussion Papers are a publication of the Schumpeter School of Business and Economics, University of Wuppertal, Germany
For editorial correspondence please contact
SSBEEditor@wiwi.uni-wuppertal.de

SDP 2018-006
ISSN 1867-5352

Impressum
Bergische Universiät Wuppertal
Gaußstraße 20
42119 Wuppertal
www.uni-wuppertal.de

BERGISCHE
UNIVERSITÄT
WUPPERTAL

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

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

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

Kerstin Schneider
WIB, University of Wuppertal and CESifo
schneider@wiwi.uni-wuppertal.de

Julian Burghoff
University of Düsseldorf
jubur104@uni-duesseldorf.de

Discussion Paper
February, 2018

To successfully reduce student attrition, it is imperative to understand which students are at risk of dropping out. We develop an early detection system (EDS) to predict student success in tertiary education 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. The developed method can be implemented in every German university, as it uses student performance and demographic data collected and maintained by legal mandate. Therefore the EDS self-adjusts to the university where it is employed. The EDS is tested and applied on a state university and a private university of applied sciences. Both institutes of higher education differ considerably in their organization, tuition fees and student-teacher ratios. Our results indicate a prediction accuracy at the end of the first semester of 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.

1. INTRODUCTION

Student attrition at universities has a negative impact on all parties involved: the students, the institutions, 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 may create feelings of inadequacy and lead to one being socially stigmatized (Larsen et al., 2013). The importance of academic performance and informational frictions for explaining dropout has been stressed in the recent literature (Stinebrickner & Stinebrickner, 2008; 2012; 2013; 2014; Arcidiacono et al., 2016). But despite the importance of the topic there is still much unknown about the underlying determinants and consequences of dropout and also about effective means to reduce student attrition.

Facing high attrition rates and an increasing demand for a qualified workforce in particular in the STEM, education policy makers are increasing their efforts to reduce the number of student dropouts (Gaebel et al., 2012). Actions to be taken have to satisfy the following criteria: They need to be cost efficient and targeted at students in need. First, the students at risk need to be identified, ideally using information that is available (administrative data). Second, students and student support have to be matched and the interventions need to be evaluated. Third, the system should be dynamic and self-adjusting.

The present paper contributes to the first and third point. We present a dynamic and self-adjusting early detection system (EDS) that can be implemented at any point in time within a student's career. The EDS uses student data that all German universities are legally required to maintain and regularly update; thus, implying that it can be readily implemented at every type of university throughout Germany. However, as these data are common administrative demographic and performance data, the EDS is with minor adjustments transferable to most institutes of higher education worldwide. At the end of each semester, the EDS is updated with the most recent student performance data and it reflects current changes in the composition of the student body and study programs; hence, the EDS is self-adjusting. To enable implementation of the early detection system with minimal costs, only student data which universities—in Germany—are already required to collect, store, and maintain are used for the implementation and regular updating of the EDS. Once implemented, the system is not constrained to a sample of students but the longitudinal census of all students. This precludes the need for costly student surveys which would otherwise need to be performed repeatedly 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 the involvement of university staff, thus, considerably easing the legal requirements with regard to data protection laws. 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 of dropouts and it offers important insights for university administration and can serve as 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.

Besides that, the EDS can also be useful for the administration of universities, e.g., 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.

The EDS is 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 science (PUAS) with about 6,700 students and 26 undergraduate programs. The state university charges no fees, while of the study programs at the private university charge fees of 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 using rich demographic and performance data; furthermore, the accuracy of the EDS increases as new student performance data becomes available at the end of each semester: after the fourth semester, the EDS correctly predicts 80% of the state university and 83% of the private university of applied sciences student dropouts. Confirming earlier studies, performance data, in particular at the early stages, is important for predicting dropout. Demographic information has only limited predictive value, once performance data is 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

The work by Stinebrickner and Stinebrickner (2008; 2012; 2014) contributes to the understanding of the determinants of student dropouts 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 major role for explaining college dropouts. Instead, academic performance is the most important factor. Moreover, they stress the importance of learning about academic performance and how this dynamic learning process affects the dropout decision. The importance of informational frictions for college attrition is confirmed by (Arcidiacono et al., 2016).

Another strand of the literature represent 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.

Another line of literature focusses on the importance of individual characteristics like the minority status or the chosen field of study. Arcidiacono (2016) use data from the University of California to argue that minorities are less likely to graduate in STEM because of being less prepared when entering university. Thus, the matching as well as the reduction in pre-college disparities ought to be focused on.

The quality of empirical work related to 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 based on surveys (Larsen et al., 2013) because of a 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 administrative student 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 dropouts.

The improvement in data mining and machine learning methods has seen an increase in 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—for instance, distance learning courses. Kotsiantis et al. (2003) analyze demographic and performance data using machine learning methods to determine whether it would be a good predictor for student success. 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), and Santana et al. (2015). While the studies are not easily comparable due to differences in sample size, variable settings, research methods, and research questions, it turned out the different methods employed within a study resulted in only marginal differences of the predictive accuracy. The more significant differences in between study results depends primarily on the predictive ability of the data and less on the method of prediction.

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 data and performance results 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 data from the online library as well as online activity level were all evaluated. In particular, learning behavior and student-teacher interactions could be observed in this way. 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 into 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 administrative data to predict whether a student will drop out from 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. In our current analysis we restrain ourselves to bachelor's degree programs; however, the method can be easily applied to master level programs, as well.

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 between 2007 and 2017. The forecasting system was then tested at both universities on student data that was not included in the training data. For testing the system, data from the winter and summer semesters of 2012 and 2010 were chosen for the PUAS and the SU, respectively[1]. These data included the most recent academic performance data for a large number of the students enrolled. SU training data of former students from the years 2007-2009 and 2011-2017 comprised a total of 12,730 observations; the 2010 data used for testing included a total of 1,766 bachelor students. PUAS training data of former students from the years 2007-2011 and 2013-2017 included a total of 6,297 observations; the test data from 2012 comprised 1,303 bachelor students.

Since the development of an EDS is not only interesting from a scientific point of view but also necessary for the governance and operation of institutions of higher education as well as the design and implementation of education policy, 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 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 the collection of 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 an land of origin	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 = other district; 0 = City of university
		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 Date of exam	Average grade per semester	1.00 to 4.00
		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 Reason for ex-matriculation	Graduate or drop out	1 = drop out, 0 = graduate

Notes:

Nationality: Citizenship and place of birth distinguishes between foreign students, students without an immigration background, and students that are first generation immigrants.

Migration background: Name based imputation of migration 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 student variables. For example, the university entrance qualification grade is, according to prevailing opinion, a well-suited predictor (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. (2011) 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 lack of satisfaction leads to dropping out or dropping out leads to lack of satisfaction or both are mutually dependent.

While possibly important for explaining student dropout rates, 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 state of health of the student are not available at universities. They are not available for every student at every university and, even if available, data protection laws render it impossible to use that information in an EDS. 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 early detection system 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 raw 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-time or part-time)

Additional information for students with a migration background includes the nationality, domestic or foreign university entrance qualification, and whether the student is a first or second generation immigrant.

In addition to the demographic data, student performance data are also available and can be used and or updated into the system as soon as they are made available at the end of each subsequent semester. The student performance data collected at the end of each completed semester includes the average semester grade, average semester credit points earned, the number of registered but unattended exams, and the number of exams that were taken but not passed. 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 result is amongst the exams in a study program most highly correlated with the successful completion of the degree (for previous cohorts). 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, the students' home address can be used to get some (limited) information about the socioeconomic background, the university entrance grade can provide information about previous academic performance, and the first and last name can provide information on student migration 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 Euro per year). Thus, students with private health insurance are typically from families with a higher socioeconomic background.

Migration 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 partly or wholly educated in Germany but with non-German citizenship. Non-German citizens born abroad are considered first generation migrants. However, second generation immigrants with German citizenship cannot be directly identified from the collected 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 immigrants. 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 a migration 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 a migration 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 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, DZHW-Berechnungen, 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 migration 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, migration background. Applying our imputation method to the GSOEP information, we correctly label 82% of the immigrants. 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 the proportion of Chinese and Turkish students, which 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 found	234	147
Germans	18,574	11,510
MigBackg	7,721	4,684
MigRate	**28.93%**	**28.93%**

	State University				**Private University of Applied Sciences**			
Region	Students with foreign nationality	Domestic students with migration background	Migration background	Proportion of student body	Students with foreign nationality	Domestic students with migration 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%
All other African regions	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 & South EasternAsia	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%
Special 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:	SU: undergraduate students between 2000 and 2017; PUAS: undergraduate students between 2007 and 2017.
Not found:	First and second name not in the database.
Germans:	Students with German citizenship and no apparent immigration background.
MigBackg:	Students with foreign nationality, 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. 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). The age at enrollment is between 21 and 22.6 years. Between 24% and 29% of the students do not have a migration background. The percentage of foreign born students is between 7% and 11%. The vast majority of students have a home address belonging to a city other than the city hosting the university in question. And, 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 quite some 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 as compared to the SU.

In the absence of performance data, the EDS forecasts are based solely on student demographic data. Demographic data available at the two universities differs. For instance, the number of students enrolled at a SU is usually substantially higher than 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 Math. Consequently, they are enrolled in two different departments. For this reason, type of study program is used as a predictor at the PUAS and not the SU.

Furthermore, there are also differences regarding university entrance requirements. Generally, the prerequisites for studying at universities of applied sciences are less restrictive than at universities; this is true for the grade of university entrance qualification (for instance, there might not be a numerus clausus) and the 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[2] 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.

[2] Performance data in table 3a and 3b is based on the data of the first semester

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

Cohort	(1) 2007	(2) 2008	(3) 2009	(4) 2010	(5) 2011	(6) 2012
Gender	0.66	0.57	0.60	0.57	0.49	0.54
(0=male; 1=female)	(0.47)	(0.50)	(0.49)	(0.50)	(0.50)	(0.50)
Age at enrollment	21.24	21.84	21.86	21.93	22.28	22.60
	(3.15)	(3.75)	(3.56)	(3.72)	(4.38)	(4.67)
Student without immigration background	0.73	0.72	0.71	0.72	0.76	0.73
(1=yes; 0=no)						
Second generation immigrant	0.19	0.17	0.19	0.18	0.17	0.19
(1=yes; 0=no)						
First generation immigrant	0.08	0.11	0.10	0.09	0.07	0.08
(1=yes; 0=no)						
City of entrance qualification	0.86	0.81	0.82	0.82	0.79	0.78
(1 = other district; 0 = city of university)	(0.34)	(0.39)	(0.38)	(0.39)	(0.40)	(0.41)
General University entrance qualification	0.97	0.95	0.95	0.94	0.94	0.94
(1=yes; 0=no)						
University of Applied Sciences	0.00	0.01	0.00	0.01	0.01	0.01
entrance qualification (1=yes; 0=no)						
Restricted university entrance qualification	0.01	0.01	0.01	0.02	0.01	0.01
(1=yes; 0=no)						
Foreign university entrance qualification	0.03	0.04	0.03	0.03	0.04	0.03
(1=yes; 0=no)						
Grade of university entrance qualification	2.87	2.85	2.79	2.71	2.68	2.61
	(0.82)	(1.00)	(0.97)	(0.87)	(0.92)	(0.89)
Health insurance	0.06	0.08	0.06	0.05	0.07	0.06
(1=private; 0=public)	(0.23)	(0.27)	(0.24)	(0.21)	(0.25)	(0.25)
# of enrolled study programs	3.06	2.62	2.67	2.71	2.32	2.20
	(1.85)	(1.87)	(1.76)	(1.97)	(1.68)	(1.51)
Lateral entrants	0.17	0.25	0.30	0.39	0.38	0.44
(1=yes; 0=no)	(0.38)	(0.43)	(0.46)	(0.49)	(0.49)	(0.50)
# of semesters at prev. university	1.63	2.47	2.65	3.15	2.63	3.02
	(4.45)	(5.32)	(5.01)	(5.09)	(4.50)	(5.13)
Average grade per semester	2.46	2.49	2.45	2.50	2.51	2.49
	(0.55)	(0.59)	(0.56)	(0.56)	(0.58)	(0.58)
Average CPs per semester	12.91	17.18	18.33	19.80	15.38	15.22
	(16.44)	(26.92)	(29.29)	(30.12)	(22.38)	(23.87)
No exam taken	0.19	0.20	0.22	0.19	0.24	0.31
# of exams per semesters not participated in	0.18	0.40	0.38	0.44	0.43	0.45
	(0.62)	(1.56)	(1.44)	(1.28)	(1.16)	(1.35)
# of failed exams per semesters	0.44	0.63	0.62	0.75	0.59	0.64
	(1.02)	(1.78)	(1.88)	(1.90)	(1.24)	(1.77)
Obs.	2637	1846	2215	2170	2860	2674

Table 3b: Summary statistics: Private University of Applied Sciences (mean and standard deviation)

	(1)	(2)	(3)	(4)	(5)	(6)
Cohort	2007	2008	2009	2010	2011	2012
Gender	0.34	0.43	0.40	0.43	0.51	0.46
(0=male; 1=female)	(0.46)	(0.46)	(0.47)	(0.48)	(0.48)	(0.48)
Age at enrollment	22.27	23.95	24.40	24.36	24.01	23.97
	(2.99)	(3.33)	(3.22)	(3.04)	(2.76)	(2.39)
Student without immigration background	0.67	0.69	0.69	0.73	0.73	0.73
Second generation immigrant	0.20	0.15	0.18	0.17	0.18	0.18
First generation immigrant	0.13	0.15	0.13	0.10	0.08	0.09
City of entrance qualification	0.74	0.66	0.66	0.70	0.70	0.68
(1 = other district; 0 = city of university)						
General University entrance qualification	0.56	0.49	0.48	0.49	0.51	0.49
University of Applied Sciences	0.40	0.44	0.48	0.46	0.44	0.44
entrance qualification (1=yes; 0=no)						
Restricted university entrance qualification (1=yes; 0=no)	0.01	0.00	0.01	0.02	0.02	0.05
Foreign university entrance qualification	0.04	0.06	0.03	0.02	0.02	0.02
(1=yes; 0=no)						
Lateral entrants	0.32	0.34	0.34	0.34	0.34	0.34
(1=yes; 0=no)	(0.50)	(0.48)	(0.48)	(0.47)	(0.47)	(0.48)
Average grade per semester	2.37	2.32	2.32	2.28	2.24	2.28
	(0.55)	(0.51)	(0.53)	(0.53)	(0.51)	(0.53)
Average CPs per semester	12.78	16.25	18.77	19.11	19.98	19.69
	(11.15)	(10.84)	(10.71)	(11.17)	(11.67)	(11.63)
No exam taken	0.35	0.18	0.11	0.09	0.09	0.09
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
# of exams per semesters not participated in	0.40	0.50	0.21	0.25	0.25	0.23
	(0.57)	(0.80)	(0.42)	(0.46)	(0.43)	(0.49)
# of failed exams per semesters	0.17)	0.30)	0.62)	0.56)	0.50)	0.54)
	(0.34)	(0.45)	(0.79)	(0.69)	(0.62)	(0.68
Obs.	193	1175	1423	1343	1358	1563

4. EMPIRICAL STRATEGY

We now present the empirical strategy for building 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. Additionally, we combine the methods by means of the AdaBoost algorithm (Schapire & Freund, 1997; Schapire & Freund, 2012). The methods used for the analysis are the OLS and probit regression models, the neural network model, as well as decision tree algorithms.

In the first step, 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. LINEAR REGRESSION MODEL

The classic approach for measuring a statistical relationship between a dependent variable and several independent variables is the regression method. In our setting, the basic multivariate linear regression model is

$$y_{it} = \beta_0 + \beta_1 x_i + \beta_2 z_{it} + \varepsilon_{it},$$

with i and t denoting student and semester, respectively. The dependent variable, y_{it}, is a binary variable representing graduate (0) and dropout (1). Demographic information, x_i, is time invariant and the performance data, z_{it}, varies over time. Section 5 discusses the results of the linear probability model using student performance and demographic data from the time of enrollment up to the sixth semester and the fourth semester for the SU and the PUAS, respectively. An advantage of the linear probability model is that it is easy to interpret and it affords a better understanding of the importance and magnitude of the explanatory variables on the likelihood of dropping out. A disadvantage of the linear regression model—in estimating probabilities—is that it allows for predicting values of the dependent variable that are less than zero and greater than one. Therefore, the probit model is used to predict student dropouts. While the result of the probit model is very similar to the OLS model, the probit model estimates better forecast probabilities for binary selection problems. As expected, the results are very similar to the linear probability model and, thus, we refrain from reporting them here. However, of the two, only the probit model will be used in formulating the AdaBoost meta-algorithm.

4.2. NEURAL NETWORK

Behavioral simulation methods are a focus of "artificial intelligence" (AI) research. AI is inspired by brain research, and since the beginning of the 1950s it has been attempting to model the structure and functioning of the human brain. Minsky and Papert showed in (1969) that training algorithms could calculate linear-separable functions. It wasn't until 1986 that Rummelhart et al. developed the algorithm for error-back propagation (backpropagation), which was used for the first time in neural networks by Werbos (1974) and revitalized the stagnation of artificial intelligence. Today, higher-dimensional neural networks (multilayer perceptrons

MLP) are being developed which, in addition to pattern recognition, image processing, and speech recognition, are used to optimize processes, control systems, and to diagnose and predict various outcomes.

In general, the MLP is learned by adjusting the weights, edges, and parameters of the selected activation function, here the logistic function of all connections. The algorithm used for this MLP is the backpropagation algorithm. In summary, the architecture of the MLP can be described by 31 neurons in the input layer, 16 (8) neurons in the first (second) hidden fully-connected layer, and one neuron in the output layer. 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 the student outcomes from the training data described above. 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 e_t, that is the difference between the actual study outcome from the training data and the predicted outcome from the neural network. The error or loss function is the sum of squares of the errors.

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

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. Therefore, backpropagation uses the error values to calculate the gradient of the loss function for finding the minimum of the error function e.

The resulting weights from a neural network are analogous to the coefficients in a linear regression model. However, the number of weights compared to the number of coefficients is excessively high, making it challenging to interpret the weights in a neural network.

4.3. Decision Tree

A decision tree assigns objects (students) to one or more predetermined classes (dropout/graduate) of the target variable using rules derived from an existing data set (the training data). A decision tree defines itself by selecting attributes of an observation as nodes and creating branches from each possible attribute value, repeating this process recursively.

The principles of entropy and information gain are used to guide the attribute selection process. The decision tree algorithm successively selects observation attributes in a top-down approach beginning with the attribute that offers the highest degree of information gain. This attribute offers the best predictive power of the final outcome and is known as the root node. The root- and successive nodes—in order of predictive power—split the observations into smaller and smaller data subsets until all observations in a subset are of homogenous outcome: a pure subset. Entropy measures the homogeneity of outcomes in a subset of the data with zero entropy corresponding to a purely homogenous subset and entropy of one corresponding to a subset with equal shares of all outcomes.

Predictions for the outcome variable across observations are determined by respective decision tree algorithms. An overview of the most frequently used algorithms can be found in Schapire and Freund (2012) or Sammut and Web (2017). In the present paper we use the C4.5

algorithm for decision trees (Hall et al., 2009)[3], which is an extension of the popular ID3 algorithm by Quinlan (1986). It removes the restriction of the ID3 (complete and error-free data, no discrete variables). The C4.5 recursively performs the process of tree building using information gain. 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 meta learning algorithm bagging (bootstrap aggregation). 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.[4] We applied bagging on the test dataset before estimating a random forest, therefore bagging with random forest (BRF).

4.4. AdaBoost

To combine the predictive powers of the neural network, regression model, and bagging with random forest, 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. 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. Note that AdaBoost is valid under the assumption that each method applied solves the decision problem better than would a random decision.

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 lower the chosen 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. We set this threshold such that the number of identified dropouts coincides with the known number of dropouts in the test cohort.

[3] We used also four different classification algorithms (PART, REPTree, M5 and Decision Stump) with our training data. All algorithms are outperformed by the bagging with random forest algorithm. We refrain from reporting the results here. They are, however, available upon request from the authors.

[4] From all tested decision trees (i.a. C4.5, M5p, CART, decision stump, RepTree) with all tested meta-learning algorithm (i.a. Bagging, random subspace, random committee, AdaBoost, classification via regression) the bagging with random forest performs best. The results are available on request.

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 proportion of correct predictions among all predictions, both recall and precision are of particular relevance. 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 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. Therefore, 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. ACCURACY OF CLASSIFIERS

Before we describe the results for the different classifiers, Figure 1 shows the accuracy of the forecast for the probit model, bagging with random forest, and AdaBoost. 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 threshold, the proportion of correct predictions among all predictions is lowest. This is true for all classifiers, however, compared to the probit and the random forest, AdaBoost performs better, albeit not over the entire range of observations. In particular the accuracy of the probit is better for students with risks slightly above the threshold.

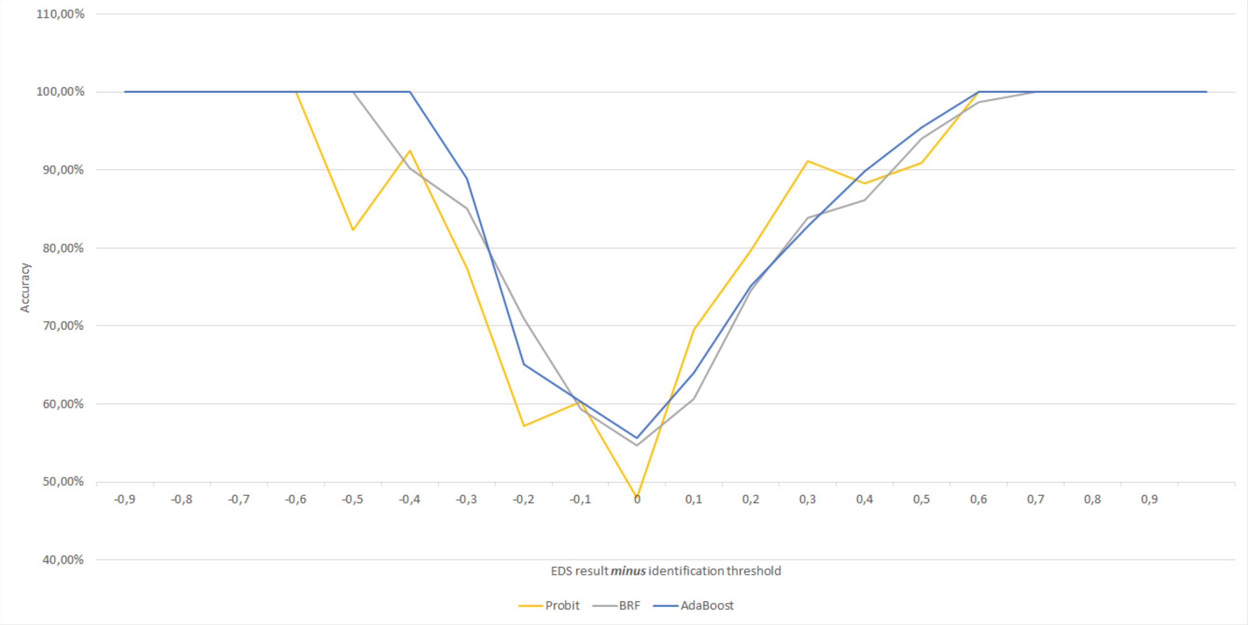


Figure 1: Accuracy of the EDS

5.2. REGRESSION RESULTS

Table 4a shows the results of the linear probability models using the student performance and demographic data from the first six semesters of the SU (cf. Table 4a, columns (1) to (7)). Note that the specification of the regression models is very simple as we want to point out correlations between the dependent and the explanatory variables in the data. More sophisticated modelling, that is targeted at a particular university, might improve the forecasting quality of the regression model. However, since the goal of the paper is to combine various methods and to build a self-adjusting tool, we refrain from putting a lot of effort into estimating parametric models.

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 number of observations in Table 4a drops by 65% from the first (11,860) to the sixth (4,149) semester due to students dropping out. It follows that the coefficients in the columns are not directly comparable, as the sample is different in every semester.

We look first at the fit of the regression model as described by the R^2. Using only the demographic information available at the time of enrolment, the R^2 is 0.1 (Table 4a, column (1)). Incorporating the performance data from the first semester increases the R^2 to 0.34 (Table 4a, column (2)). The R^2 jumps to 0.49 in the second semester and reaches 0.56 in the sixth semester.

Note that the estimates in column (1) only reflect the demographic variables, i.e. information that is available at time of enrollment. At the time of enrollment, it is more likely that a male student drops out as compared to a female student, and the probability of dropping out is increasing with age at enrollment. 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. Students with a university of applied sciences entrance qualification high school degree (Fachhochschulreife) are less likely to finish their studies as compared to students with a general university entrance qualification (Allgemeine Hochschulreife). The coefficient on the high school grade (Abiturnote) is negative and statistically significant.[5] The coefficient on the dummy variable for private health insurance is not statistically significant. And, lateral entrants graduate more often at SU.

Most of the demographic variables lose statistical significance when controlling for the performance data available after the first semester. Thus, the rich student data available at the time of enrollment is only valuable, if the EDS tries to identify students at risk right at the beginning of their studies. Even as early 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. Controlling for academic performance, students who have private health insurance are more likely to graduate than those who have public health insurance. As stated 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 dropout. However, over time, the coefficient on this variable becomes smaller and loses statistical significance. The regression coefficients of the performance variables (Average Grade, No Exam, Not Participated, and Failed Exam) have a negative impact on study success (columns 2-7). Not surprisingly, failed exams and non-participation in exams are good predictors for dropouts. Note, however, that in higher semesters, performance indicators become less informative and less significant when predicting dropouts. And, first semester results from the performance variables continue to have explanatory power in later semesters. Thus, students who do not drop out after having performed poorly in the first semester, c.p. 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.

[5] In the German grading system (school and tertiary education) low grades on the scale from 1 to 5 are associated with high performance, whereas a 5 indicates failure.

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

Dependent variable: student graduates (0=yes; 1=no); OLS							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Enrollment	1st Semester	2nd Semester	3rd Semester	4th Semester	5th Semester	6th Semester
Gender	0.097**	0.047**	0.043**	0.039**	0.021*	0.008	-0.004
(0=male; 1=female)	(0.000)	(0.000)	(0.000)	(0.000)	(0.019)	(0.345)	(0.725)
Age at enrollment	0.009**	0.005**	0.009**	0.007**	0.009**	0.007**	0.008**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Second generation immigrant	0.038**	0.014	0.012	-0.001	-0.002	0.000	0.000
(1=yes; 0=no)	(0.000)	(0.124)	(0.220)	(0.936)	(0.857)	(0.983)	(0.986)
First generation immigrant	0.067**	0.021	0.018	0.025	0.022	0.005	0.005
(1=yes; 0=no)	(0.000)	(0.202)	(0.306)	(0.172)	(0.238)	(0.790)	(0.794)
City of entrance qualification	-0.069**	-0.021*	-0.022*	-0.025*	-0.031**	-0.038**	-0.053**
(1=other district; 0=city of uni.)	(0.000)	(0.019)	(0.027)	(0.012)	(0.003)	(0.000)	(0.000)
Uni. of Appl. Sciences entrance	0.147**	0.068*	0.038	0.006	0.013	-0.033	-0.004
qualification (1=yes; 0=no)	(0.001)	(0.072)	(0.364)	(0.900)	(0.803)	(0.565)	(0.951)
Restricted university entrance	-0.142^{+}	0.070	-0.065	-0.031	-0.035	0.090	0.078
qualification (1=yes; 0=no)	(0.074)	(0.301)	(0.348)	(0.672)	(0.661)	(0.337)	(0.422)
Foreign university entrance	-0.130	-0.084	-0.143*	-0.139*	-0.147*	-0.052	-0.058
qualification (1=yes; 0=no)	(0.154)	(0.278)	(0.070)	(0.063)	(0.057)	(0.501)	(0.514)
Grade of university entrance	0.120**	0.017*	-0.010	-0.024**	-0.018*	-0.017*	-0.014
Qualification	(0.000)	(0.013)	(0.188)	(0.002)	(0.022)	(0.031)	(0.129)
Health insurance	-0.028	-0.057**	-0.033*	-0.047**	-0.032*	-0.023	-0.017
(1=private; 0=public)	(0.108)	(0.000)	(0.037)	(0.004)	(0.054)	(0.149)	(0.374)
# of enrolled study programs	-0.031**	-0.011**	-0.002	0.005*	0.008**	0.012**	0.016**
	(0.000)	(0.000)	(0.489)	(0.089)	(0.009)	(0.000)	(0.000)
Lateral entrants	-0.174**	-0.081**	-0.083**	-0.061**	-0.056**	-0.065**	-0.079**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
# of semesters at prev.	0.016**	0.008**	0.006**	0.004**	0.002	0.002^{+}	0.001
university	(0.000)	(0.000)	(0.000)	(0.009)	(0.213)	(0.092)	(0.463)
Average grade current semester		0.114**	0.062**	0.029**	0.032**	0.015*	-0.018*
		(0.000)	(0.000)	(0.000)	(0.000)	(0.067)	(0.049)
Average CPs current semester		-0.010**	-0.011**	-0.008**	-0.005**	-0.004**	-0.004**
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
No exam taken current semester		0.542**	0.349**	0.0254**	0.289**	-0.236**	0.107*
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
# of exams current semester not		0.048**	0.035**	0.023**	0.004	0.019**	0.011
participated in		(0.000)	(0.000)	(0.000)	(0.554)	(0.002)	(0.138)
# of failed exams current semester		0.045**	0.037**	0.028**	0.037**	0.022**	0.031**
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.456**	0.211**	0.161**	0.264**	0.244**	0.360**	0.423**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Performance previous semesters:							
Previous average grades			YES	YES	YES	YES	YES
Previous average CPs			YES	YES	YES	YES	YES
Previous # of exams			YES	YES	YES	YES	YES
Prev. # of not participated. exams			YES	YES	YES	YES	YES
Previous # of failed exams			YES	YES	YES	YES	YES
R^2	0.096	0.344	0.486	0.537	0.558	0.576	0.558
F	96.731	221.432	243.043	225.389	180.742	158.574	97.723
N	11860	11860	8509	7409	6212	5655	4149

Table 4b shows the results of the OLS estimation using the student performance data and the demographic data from semesters one through four of the PUAS. The number of observations drops from 6,296 in the first semester to 4,822 students in the fourth semester. In the first semester, similar to the SU, 33.4% of the variance in the dependent variable is explained by the model, this increases to 61% in the fourth semester. A presumption is that the tuition fees—that are not paid at the SU—lead to the decision to drop out sooner.

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. The aim of the forecasting system is to successfully predict future graduates and dropouts from the student observations. In this section it was shown that a regression model using the demographic and performance variables fits the data quite well and produces good predictions of graduates and dropouts.

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

Dependent variable: student graduates (0=yes; 1=no); OLS					
	(1) Enrollment	(2) 1st 0.ester	(3) 2nd Semester	(4) 3rd Semester	(5) 4th Semester
Gender	0.059**	0.029*	-0.007	-0.013	-0.014*
(0=male; 1=female)	(0.000)	(0.012)	(0.444)	(0.140)	(0.054)
Age at enrollment	0.005**	0.005**	0.003**	0.003**	0.003**
	(0.001)	(0.001)	(0.005)	(0.009)	(0.000)
Second generation immigrant	0.036*	-0.007	-0.018*	-0.023*	-0.010
(1=yes; 0=no)	(0.013)	(0.561)	(0.092)	(0.015)	(0.218)
First generation immigrant	0.091**	-0.001	-0.027*	-0.011	-0.015
(1=yes; 0=no)	(0.000)	(0.940)	(0.078)	(0.441)	(0.215)
City of entrance qualification	-0.031*	0.003	0.022*	0.016*	0.006
(1=other district; 0=city of uni)	(0.012)	(0.807)	(0.013)	(0.044)	(0.418)
Uni. of Appl. Sciences entrance	0.147**	0.038**	0.007	0.010	0.003
qualification (1=yes; 0=no)	(0.000)	(0.000)	(0.455)	(0.201)	(0.718)
Restricted university entrance	0.259**	0.148**	0.074*	0.063*	0.044
qualification (1=yes; 0=no)	(0.000)	(0.000)	(0.025)	(0.040)	(0.113)
Foreign university entrance	0.356**	0.156**	0.021	-0.019	-0.065*
qualification (1=yes; 0=no)	(0.000)	(0.000)	(0.487)	(0.490)	(0.013)
Lateral entrants	0.050**	0.045**	0.026**	0.016**	0.008*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.046)
Average grade current semester		0.115**	0.033**	0.016*	-0.021*
		(0.000)	(0.000)	(0.069)	(0.010)
Average CPs current semester		-0.002**	-0.009**	-0.007**	-0.004**
		(0.005)	(0.000)	(0.000)	(0.000)
No exam taken current semester		0.546**	0.385**	0.216**	0.188**
		(0.000)	(0.000)	(0.000)	(0.000)
# of exams current semester not		0.048**	0.026**	0.003	-0.022**
participated in		(0.000)	(0.000)	(0.566)	(0.000)
# of failed exams current semester		0.057**	0.009*	0.003	0.008*
		(0.000)	(0.046)	(0.470)	(0.020)
Constant	-1.015**	-1.038**	0.515**	0.359**	0.286**
	(0.000)	(0.000)	(0.000)	(0.000)	(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
R^2	0.119	0.352	0.511	0.580	0.610
F	46.969	102.986	153.482	164.447	155.439
N	6296	6296	5611	5173	4822

As noted earlier, we do not report the results of the probit model. Instead, Table 5 shows the forecasting quality measures. As expected, quality of the prediction increases over time. This applies to all quality measures. For instance, the recall (how many of the at-risk students are identified), rises from about 71% in the first semester at the SU to 79% in the fourth semester. At the PUAS, recall for the 1st and 4th semesters was 66% and 80%, respectively.

Table 5: Forecasting quality of the probit model

Probit	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	64.74%	75.75%	81.29%	86.08%	89.04%	67.67%	81.26%	89.53%	92.27%	94.76%
Recall	63.45%	70.39%	73.33%	78.29%	78.80%	50.87%	66.30%	76.32%	78.54%	80.42%
No. of graduates	1112	1039	1027	1015	992	940	940	938	933	925
No. of dropouts	1015	726	555	479	349	458	362	266	205	143
Correctly predicted-graduates	733	826	879	911	919	713	818	875	889	897
Incorrectly predicted-graduates	371	215	148	104	74	225	122	63	44	28
Correctly predicted-dropouts	644	511	407	375	275	233	240	203	161	115
Incorrectly predicted-dropouts	379	213	148	104	73	227	122	63	44	28
Correctly predicted-graduates	65.92%	79.50%	85.59%	89.75%	92.64%	75.85%	87.02%	93.28%	95.28%	96.97%
Incorrectly predicted-graduates	33.36%	20.69%	14.41%	10.25%	7.46%	23.94%	12.98%	6.72%	4.72%	3.03%
Correctly predicted-dropouts	63.45%	70.39%	73.33%	78.29%	78.80%	50.87%	66.30%	76.32%	78.54%	80.42%
Incorrectly predicted-dropouts	37.34%	29.34%	26.67%	21.71%	20.92%	49.56%	33.70%	23.68%	21.46%	19.58%

5.3. RESULTS FROM BAGGING WITH RANDOM FOREST AND NEURAL NETWORK

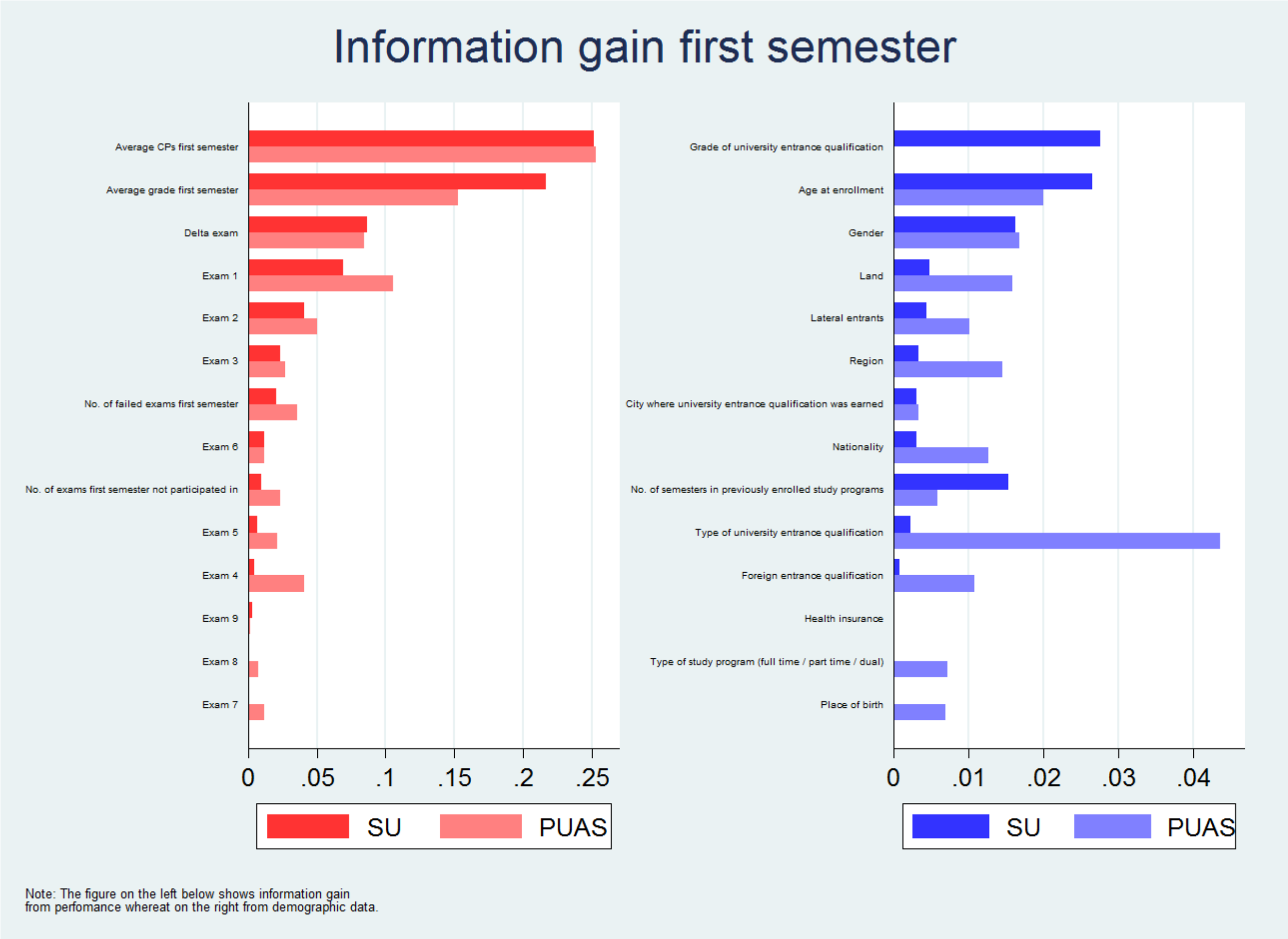
Next, we base the prediction on machine learning methods. In line with similar analyses found in the literature, there is not much difference in the forecast accuracy between the tested methods, the regression model, the neural net and the random forest. Furthermore, we also confirm the superior performance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

[6] The results of all tested methods are available on request.

Table 6: Accuracy of forecasting methods compared to the Bagging Random Forest

Average difference between forecasting performance measures compared to BRF	**SU**		**PUAS**	
	Probit	Neural Net	Probit	Neural Net
Accuracy	-0,88 %	-2,93 %	-1.03 %	-0.88 %
Recall	-1,38 %	-9,88 %	-2.53 %	-2.58 %
Correctly predicted-graduates	-0,62 %	1,10 %	-0.61 %	-0.45 %
Incorrectly predicted-graduates	0,82 %	5,51 %	0.76 %	0.69 %
Correctly predicted-dropouts	-1,38 %	-9,88 %	-2.53 %	-2.58 %
Incorrectly predicted-dropouts	0,94 %	-0,46 %	1.94 %	1.63 %

Figure 2 shows the value of the information gain in the random forest using data from the first semester. In Figure 2, we differentiate between demographic variables (blue) and performance variables (red) as well as between the two universities.


Information gain first semester

Note: The figure on the left below shows information gain from perfomance whereat on the right from demographic data.

Figure 2: Bagging with random forest (information gain); First semester

It is apparent that for both universities, the performance data is a much better predictor of dropouts than the demographic data. In particular, the pace of study (avg. CP/semester), the average grade (avg. Grade/semester) as well as the most important exam have a high degree of explanatory power. Comparing SU and PUAS, the five most important predictor variables are identical for both universities and the information gain differ only slightly from each other.
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.

The ROC curve supports these results. First, all methods perform substantially better than a random guess. Second, prediction power improves with more information (the area below the ROC curve increases). In addition, the ranking of the methods differs slightly by university and semester. This is our motivation for combining the predictive power of neural networks, bagging with random forest, and probit model using the AdaBoost algorithm.

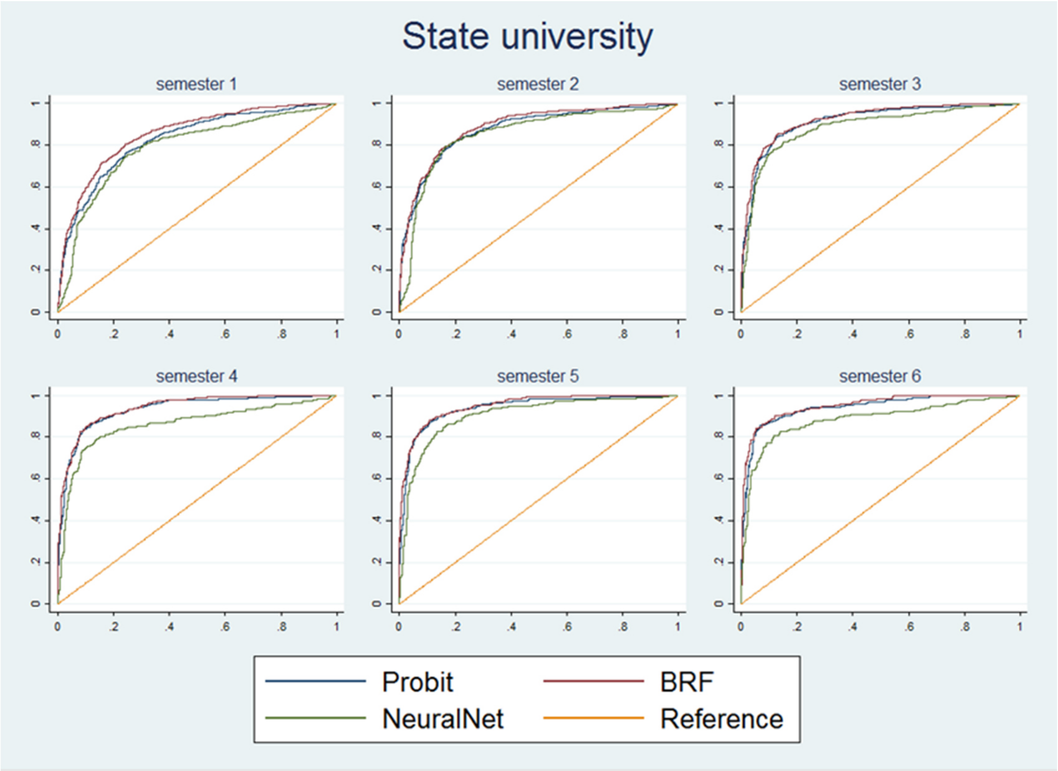


Figure 3a: ROC curve - state university

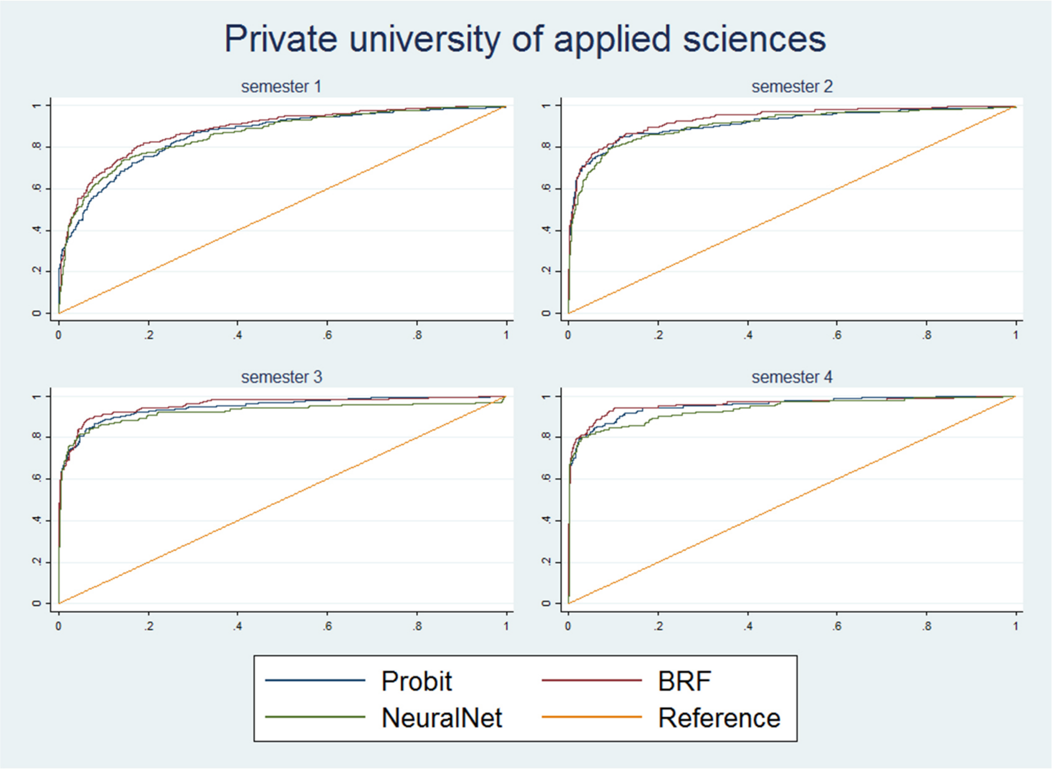


Figure 3b: ROC curve – private university of applied sciences

5.4. RESULTS FROM BOOSTING (ADABOOST)

Table 7 summarizes the forecast accuracy of the AdaBoost. 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: EDS performance accuracy 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	68.08%	78.87%	82.62%	87.62%	89.71%	67.17%	84.49%	89.70%	93.50%	95.51%
Recall	66.50%	74.24%	75.14%	80.58%	80.23%	49.78%	72.10%	76.69%	81.95%	83.22%
No. of graduates	1112	1039	1027	1015	992	940	940	938	933	925
No. of dropouts	1015	726	555	479	349	458	362	266	205	143
Correctly predicted-graduates	773	853	890	923	923	711	839	876	896	901
Incorrectly predicted-graduates	340	187	138	93	69	230	101	62	37	24
Correctly predicted-dropouts	675	539	417	386	280	228	261	204	168	119
Incorrectly predicted-dropouts	339	186	137	92	69	229	101	62	37	24
Correctly predicted-graduates	69.51%	82.10%	86.66%	90.94%	93.04%	75.64%	89.26%	93.39%	96.03%	97.41%
Incorrectly predicted-graduates	30.58%	18.00%	13.44%	9.16%	6.96%	24.47%	10.74%	6.61%	3.97%	2.59%
Correctly predicted-dropouts	66.50%	74.24%	75.14%	80.58%	80.23%	49.78%	72.10%	76.69%	81.95%	83.22%
Incorrectly predicted-dropouts	33.40%	25.62%	24.68%	19.21%	19.77%	50.00%	27.90%	23.31%	18.05%	16.78%

Next, we focus on information at time of enrollment, i.e., using only the demographic data. Of the final number of dropouts, about 21% left the PUAS and 28.5% left the SU before the end of the first semester. 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.

At both universities, the forecasting accuracy increases in semesters, as the probability of dropping out decreases with each additional semester. The forecast accuracy at the PUAS improves faster in the earlier semesters, whereas that of the SU increases at a steadier rate. The forecast accuracy at the SU was 90.99% (81.35% recall) and 91.85% (82.94% recall) for the fifth and sixth semesters, respectively.

At both universities, forecast results based on the information at the time of enrollment have a prognostic accuracy of about 68%. But they differ significantly with regard to recall and precision. While successful students are forecasted with a high degree of accuracy at the PUAS, dropouts are better forecasted at the SU. Using only performance data, predictions for the two universities are similar, as is summarized in Table 8.

Table 8: EDS performance accuracy 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%
No. of graduates	1039	1027	1015	992	940	938	933	925
No. of dropouts	726	555	479	349	362	266	205	143
Correctly predicted-graduates	833	880	915	915	836	881	890	897
Incorrectly predicted-graduates	207	147	100	77	109	58	43	28
Correctly predicted-dropouts	519	408	379	272	253	208	162	115
Incorrectly predicted-dropouts	206	147	100	77	104	57	43	28
Correctly predicted-graduates	80.17%	85.69%	90.15%	92.24%	88.94%	93.92%	95.39%	96.97%
Incorrectly predicted-graduates	19.92%	14.31%	9.85%	7.76%	11.60%	6.18%	4.61%	3.03%
Correctly predicted-dropouts	71.49%	73.51%	79.12%	77.94%	69.89%	78.20%	79.02%	80.42%
Incorrectly predicted-dropouts	28.37%	26.49%	20.88%	22.06%	28.73%	21.43%	20.98%	19.58%

At both universities, forecasts based on the performance data provide almost the same results as forecasts using both the demographic and 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 career before students take the first exams, demographic data is an important source of information. Once achievement data is available after the first semester, rich demographic data adds only 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.

5.5. VALUE ADDED FROM ADDITIONAL INFORMATION

The following variables were only available at one of the two universities: the grade on the university entrance qualification, student health insurance type, and the student's respective study program, the latter being a consequence of the SU facilitating enrollment in multiple study programs (as mentioned above, PUASs only allow enrollment in one study program). The predictive relevance of the above mentioned variables is unclear, but of political and theoretical importance. The regression results in Tables 4a and 4b show a significant effect from the variables in question. However, this does not imply that the information is important for the predictive power of the EDS using the AdaBoost algorithm. Table 9 summarizes the results of the AdaBoost with and without the type of health insurance and the grade of the university entrance qualification. As the table shows, the value added from the two variables is negligible after the first semester, implying that their explanatory power is captured by other variables, in particular by performance data.

Table 9: EDS performance accuracy for the state university (with and without health insurance and university entrance qualification grade)

AdaBoost	State university with university entrance grade and health insurance					State university without university entrance grade and health insurance				
	Enrollment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.	Enrollment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.
Accuracy	68.08%	78.87%	82.62%	87.62%	89.71%	64.65%	78.81%	82.62%	87.82%	89.71%
Recall	66.50%	74.24%	75.14%	80.58%	80.23%	62.86%	74.24%	75.14%	81.00%	80.23%
No. of graduates	1112	1039	1027	1015	992	1112	1039	1027	1015	992
No. of dropouts	1015	726	555	479	349	1015	726	555	479	349
Correctly predicted-graduates	773	853	890	923	923	737	852	890	924	923
Incorrectly predicted-graduates	340	187	138	93	69	377	187	138	91	69
Correctly predicted-dropouts	675	539	417	386	280	638	539	417	388	280
Incorrectly predicted-dropouts	339	186	137	92	69	375	187	137	91	69
Correctly predicted-graduates	69.51%	82.10%	86.66%	90.94%	93.04%	66.28%	82.00%	86.66%	91.03%	93.04%
Incorrectly predicted-graduates	30.58%	18.00%	13.44%	9.16%	6.96%	33.90%	18.00%	13.44%	8.97%	6.96%
Correctly predicted-dropouts	66.50%	74.24%	75.14%	80.58%	80.23%	62.86%	74.24%	75.14%	81.00%	80.23%
Incorrectly predicted-dropouts	33.40%	25.62%	24.68%	19.21%	19.77%	36.95%	25.76%	24.68%	19.00%	19.77%

We also use the choice of study program data to predict dropouts (Table 10). As argued above, this is only useful for the PUAS and has only limited value for the SU. As before, once the performance results are available at the end of the first semester, the predictive value from the study program diminishes; however, it remains relevant. For instance, even after the fourth semester, the percentage of true predicted dropouts increases by about 4 percentage points.

Table 10: EDS performance accuracy for the Private University of Applied Sciences

AdaBoost	PUAS with study program					PUAS without study program				
	Enrollment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.	Enrollment	1st Sem.	2nd Sem.	3rd Sem.	4th Sem.
Accuracy	67.17%	84.49%	89.70%	93.50%	95.51%	66.74%	84.87%	89.70%	92.97%	94.38%
Recall	49.78%	72.10%	76.69%	81.95%	83.22%	49.13%	72.65%	76.69%	80.49%	79.02%
No. of graduates	940	940	938	933	925	940	940	938	933	925
No. of dropouts	458	362	266	205	143	458	362	266	205	143
Correctly predicted-graduates	711	839	876	896	901	708	842	876	893	895
Incorrectly predicted-graduates	230	101	62	37	24	233	99	62	40	30
Correctly predicted-dropouts	228	261	204	168	119	225	263	204	165	113
Incorrectly predicted-dropouts	229	101	62	37	24	232	98	62	40	30
Correctly predicted-graduates	75.64%	89.26%	93.39%	96.03%	97.41%	75.32%	89.57%	93.39%	95.71%	96.76%
Incorrectly predicted-graduates	24.47%	10.74%	6.61%	3.97%	2.59%	24.79%	10.53%	6.61%	4.29%	3.24%
Correctly predicted-dropouts	49.78%	72.10%	76.69%	81.95%	83.22%	49.13%	72.65%	76.69%	80.49%	79.02%
Incorrectly predicted-dropouts	50.00%	27.90%	23.31%	18.05%	16.78%	50.66%	27.07%	23.31%	19.51%	20.98%

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 in support at the universities.

In addition to traditional regression analyses, we also employ machine learning algorithms, since the automatic EDS should perform without complex model building which has to be adjusted permanently. Instead of relying on a single method, we use the AdaBoost algorithm to combine the various methods employed, thus, reducing the disadvantages inherent in using any single method and also accounting for heterogeneity of study programs and the student body at 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 the AdaBoost is strong, and the accuracy of the results varies with increasing semesters and available data, i.e., performance and demographic data versus only demographic data. Depending on the semester and the corresponding information that is available, 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 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 presented system 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 the students' decision to drop out. That way, the public and private costs associated with attrition can be possibly reduced by building an EDS and use it as a starting point for allocating student support to the students in need and for testing the effectiveness of student support.

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? *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. & Burkhart, S. (2013) Bildungsinländer 2011 - Daten und Fakten zur Situain von von ausländischen Studierenden. Bonn.

Heublein, U. & Wolter, A. (2011) Studienabbruch in Deutschland. Definition, Häufigkeit, Ursachen, Maßnahmen. *Zeitschrift für Pädagogik*, 214-36.

Humpert, A. & Schneiderheinze, K. (2002) *Stichprobenziehung für telefonische Zuwandererumfragen. Praktische Erfahrungen und Erweiterung der Auswahlgrundlage.* Münster: Waxmann.

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.

Statistisches Bundesamt. (2015) Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund – Ergebnisse des Mikrozensus 2015.

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.*