European Journal of Higher Education

ISSN: 2156-8235 (Print) 2156-8243 (Online) Journal homepage: https://srhe.tandfonline.com/loi/rehe20

Predicting student dropout: A machine learning approach

Lorenz Kemper, Gerrit Vorhoff & Berthold U. Wigger

To cite this article: Lorenz Kemper, Gerrit Vorhoff & Berthold U. Wigger (2020): Predicting student dropout: A machine learning approach, European Journal of Higher Education, DOI: 10.1080/21568235.2020.1718520

To link to this article: https://doi.org/10.1080/21568235.2020.1718520

EUROPEAN JOURNAL OF HIGHER EDUCATION
https://doi.org/10.1080/21568235.2020.1718520

Predicting student dropout: A machine learning approach

Lorenz Kemper, Gerrit Vorhoff and Berthold U. Wigger

Department of Economics, Karlsruhe Institute of Technology, Karlsruhe, Germany

ABSTRACT

We perform two approaches of machine learning, logistic regressions and decision trees, to predict student dropout at the Karlsruhe Institute of Technology (KIT). The models are computed on the basis of examination data, i.e. data available at all universities without the need of specific collection. Therefore, we propose a methodical approach that may be put in practice with relative ease at other institutions. We find decision trees to produce slightly better results than logistic regressions. However, both methods yield high prediction accuracies of up to 95% after three semesters. A classification with more than 83% accuracy is already possible after the first semester.

ARTICLE HISTORY
Received 11 September 2019
Accepted 7 October 2019

KEYWORDS
Educational data mining; student dropout prediction; retention management; machine learning; student attrition

Introduction

The negative consequences of dropping out of the educational system are considerable, both for the individuals as well as the affected institutions. Indeed, dropout imposes costs on all parties involved, be it resources, time or money (Gansemer-Topf and Schuh 2006; Yu et al. 2010). Consequently, preventing educational dropout poses a major challenge to institutions of higher education (Zhang et al. 2010). The identification of potential dropout students constitutes a first step to improve retention policies such as learning assistance or mentorship programmes. Quantifying dropout risks may thus prove to be helpful in allocating pedagogical, psychological and administrative resources in an efficient way. The prediction of student dropout is generally inspired by churn analysis as employed in customer relationship management. Companies attempt to reduce customer attrition employing churn analysis, i.e. by identifying and maintaining at-risk customer relationships with the assistance of predictive data mining (Delen 2010).

So far, most of the research concerning dropout prediction has been conducted outside Germany. However, understanding and tackling dropout at German universities is particularly relevant for two reasons. Firstly, with about 30% of the students not completing their studies, dropout constitutes a widespread phenomenon in German higher education (Heublein 2014). Secondly, economic growth and demographic change have led to a shortage of qualified specialists (the so-called Fachkräftemangel) in the German labour market (Federal Labour Office, 2017). Therefore, the recruitment of university graduates poses a challenge to domestic firms. Since the Fachkräftemangel is particularly severe in technical domains, the current dropout rates in STEM fields[1] are especially critical (Hetze 2011).

CONTACT Lorenz Kemper kemper.lb@gmail.com

Consequently, at the Karlsruhe Institute of Technology (KIT), as one of Germany's leading technical universities, preventing dropout is high on the agenda. It, thus, provides a particularly suitable showcase for conducting analyses tackling student dropout.

As we aim to establish an applicable technique that can be readily implemented at other universities, the underlying data must come from sources that are generally accessible. More precisely, our approach relies on conventional study progression data in the form it appears on the typical student's transcript of records. This approach has the crucial advantage that it is easily replicable at other institutions in similar higher education systems and that it dominates approaches that rely on survey data (see, e.g. Caison 2007).

In order to analyse the data, we can choose from a wide set of machine learning techniques. For our purpose, the identification process should be easy to understand, so that it remains comprehensible to practitioners in the sense that course-specific insights on the internal drivers of dropout can be derived from the method. Therefore, we opt for the following two techniques: a logistic regression and a decision tree model. Even though more complex techniques like neural nets or random forests may outperform our approaches in terms of prediction accuracy (Yu et al. 2010), concessions should be made in favour of these practical considerations. We show that using the two methods and data, that is (1) collected at universities anyway and (2) directly related to individual study achievement, the machine is able to identify future dropouts with high accuracy. The advantage of our data selection is, that it is relatively harmless from a data protection perspective and, moreover, that it does not imply ex ante discrimination on the basis of criteria which are not related to study achievements.

The remainder of the paper is organized as follows. Section 2 discusses the related literature and further elaborates on the key idea of this paper. Section 3 introduces the dataset. Section 4 describes the methodology in detail. Section 5 presents the results and checks for their robustness. Section 6 concludes.

Related literature

Original approaches to model and analyse student dropout in higher education reach 40 years back to the famous survival model of Tinto (1975). In Germany, student dropout research is, above all, connected to Heublein et al. (e.g. 2003, 2010, 2014), which entail general analyses of the drivers of student success. Data mining techniques to predict individual student dropout have, to our knowledge, not yet been employed to German higher education institutions. In other countries, this topic has received more attention. Generally, this literature approaches the issue from a range of different angles. There are three main categories in which the approaches differ, namely setting, data and methodology.

With regard to the setting, it can be observed that all studies confine attention to single institutions. This reflects the fact that study programmes differ considerably between higher education institutions, so that machine learning is bound to institution-specific data. Therefore, the setup of the studies heavily depends on both, the availability and the access to the respective local information system. While some authors study the topic with respect to a single course or faculty (e.g. Dekker, Pechenizkiy, and Vleeshouwers 2009), others take an entire university into consideration (e.g. Djulovic and Li 2013), analyse a distance-learning environment (Kovacic 2010), utilize the availability of online-learning platforms (Yukselturk, Ozekes, and Türel 2014) or even analyse ID-card

transactions (Ram et al. 2018). Different setting circumstances yield different data bases. For example, with respect to timing, some studies are able to make early predictions as pre-enrolment data are available (Dekker, Pechenizkiy, and Vleeshouwers 2009; Djulovic and Li 2013; Yu et al. 2010). Other studies, such as ours, are only looking into data that is produced in the course of study.

Another aspect in which the studies differ is the structure of the data. While some information systems generate well-structured table data that is relatively easy to handle, other studies cope with unstructured data such as log data (Ram et al. 2018) or text data (Zhang et al. 2010). Finally, data are subject to different data privacy requirements and codes of ethics. For example, data features such as gender or ethnicity are especially sensitive when they are used for pre-enrolment models, as this will lead to discrimination within a university's selection process. Also, massive data records of daily activities such as payment transactions may be seen critically, since they could be perceived as surveillance. This is especially relevant for personalized data use such as assessing individual dropout risks.

Lastly, the studies differ with respect to methodology. This is especially true for pre-processing steps such as feature extraction, resampling and text tokenization. Also, studies choose from a large set of classification algorithms. The rise in popularity of neural networks, for example, is due to its performance on large, heterogeneous datasets. Therefore, neural networks are the method of choice in ID-card settings (Ram et al. 2018). In other data scenarios, classical classification methods are more typical. Classifiers in the literature include tree models such as decision trees or random forests (e.g. Aulck et al. 2016), linear models such as probit or logit regressions (e.g. Kovacic 2010), support vector machines (e.g. Zhang et al. 2010) or ensemble methods (e.g. Delen 2010). Since our dataset is both, comparatively homogeneous and comparatively small we employ decision trees and logistic regressions as classifiers. This has the advantage that the driving factors which lead to student dropout can be directly inferred from the methods. Other methods such as neural networks or random forests are less accessible in this respect. This is why our approach should be more revealing for practitioners in higher education.

Data

Empirical studies on dropout at German universities show that performance problems represent the single most important cause of leaving higher education among bachelor students (see figure 1).[2] If multi-selection was allowed, 70% of dropout students mention performance issues as a relevant factor for their decision (Heublein 2010). This finding underscores the use of study progression data for our investigation, as it displays the students' progress and performance in high detail. Since investigations of personal background information also show promise (Kovacic 2010), the analysis additionally includes few biographical parameters which are collected anyway by the university during enrolment for the course. The dataset covers full cohorts of Industrial Engineering students starting their studies at KIT between 2007 and 2012 (fall term). Until 2016, there have been 2556 cases of successful graduation, while 620 students have dropped out.[3]

The centrepiece of our dataset is made up of detailed observations of 487 different examinations (*taken examinations*, *grades*, *dates*, *results* and *number of attempts*). In addition, we received general student data including the *enrolment date*, the *final graduation result* (success, dropout, currently enrolled, paused) and, if available, the *date of*

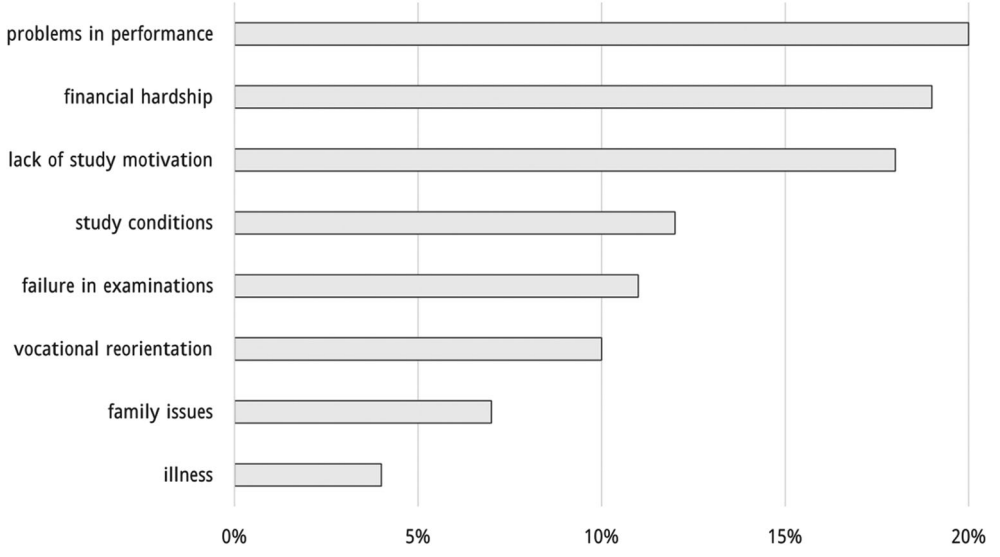

Figure 1. Crucial dropout motives of bachelor students: HIS dropout survey 2008 in (Heublein 2010), own representation.

graduation/dropout. This data can be assigned to the respective student by means of a generated anonymous *ID*.[4] In order to feed the models with second-level information, we add combined ('extracted') features, on top of which the models can be trained. These extracted features are *average grade performance*, *count of failed exams*, *count of participated examinations* and *count of active semesters*. Due to easy availability, the baseline data also comprises personal data including *age*[5], *gender* and *country of birth* (classified to German and non-German). Altogether, after pre-processing the data as well as selecting and extracting features our feature space can be summarized as portrayed in Table 1.

Table 1. Overview of the feature space.

Feature	Type	Description	Value
ID	nominal	Anonymous ID	
success	nominal	Indicator of study status	0 = dropout 1 = success 2 = enrolled
endat	date	Date of enrolment	10/1/07–10/1/12
sex_m	nominal	Gender	1 = male 0 = female
orig_d	nominal	Origin	1 = German 0 = non-German
age	numeric	Age at enrolment	ages 20–24 per year, >24 per five years
X*****_grade[12]	numeric	Exam grade	grade (1.0–5.0)
X*****_attempt	numeric	No. of exam attempts	attempts (1–3)
X*****_sem	numeric	Semester to take exam	≥ 0
X*****_status	nominal	Exam result	0 = failed 1 = passed 2 = applied
avg_grade	numeric	Average grade in all exams	grade (1.0–5.0)
avg_pass	numeric	Average grade: passed exams	≥0
avg_npass	numeric	Average grade: failed exams	≥0
*p*_count	integer	Count of exams	≥0
count_pass	integer	Count of passed exams	≥0
count_npass	integer	Count of failed exams	≥0

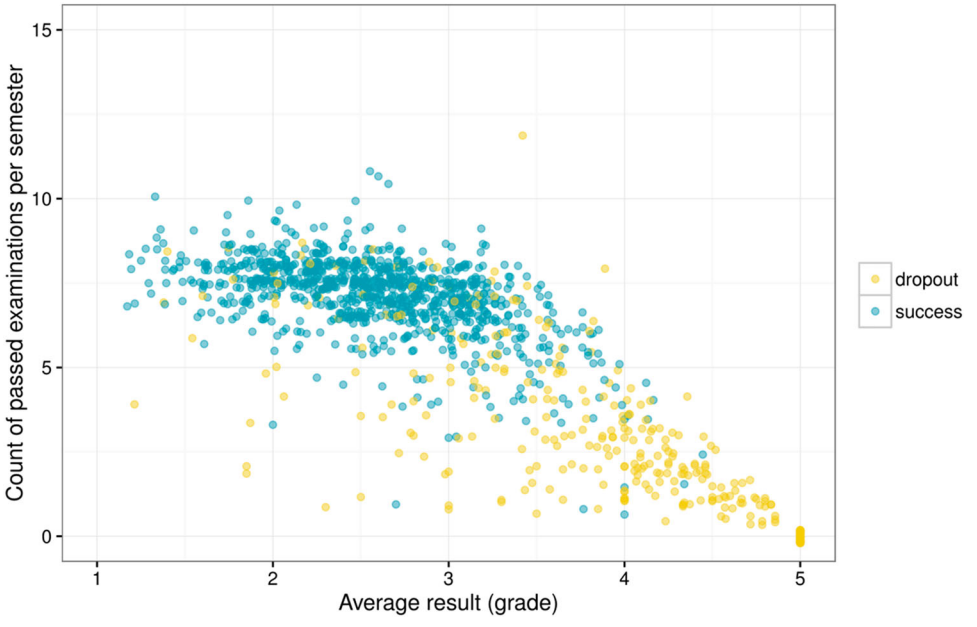

Figure 2. Successful and drop-out students compared by average result and average passed examinations per semester; Data of all semesters.

Data exploration

Plotting the average grade against the amount of passed examinations per semester indicates that a good split of successful and not successful students might already be possible using these features. A tendency for dropout students to cluster in the bottom right corner, representing fewer passed examinations at a lower average grade, can be observed. On the contrary, generally successful students pass more examinations and achieve better results (see Figure 2). This observation even holds if only data from the first two semesters are included. These students take more exams, while the overall grade performance remains the same. Apparently, student dropout may be predicted at an early stage. The importance of such an early forecast becomes apparent by looking at the timely dropout of our students. Indeed, 179 (331) of the 620 dropout cases can already be observed after the first (second) semester (see figure 3).

Even though early dropout occurs most frequently, an accurate prediction for later cases should not be neglected as the dropout distribution shows a long tail. Even more, both from an expenditure perspective as well as from a signalling perspective (cf. Spence 1973) the costs of dropout rise the later a student decides to end his or her studies. While an early prediction is desirable, late prediction comes with the advantage of a higher prediction quality, as accuracy generally rises with the amount of available data.

Selection and balancing of data

Exploring at which point of the studies reliable predictions can be made, we limit the exam data to the first three semesters of a student's studies. Thereafter, we create three different datasets representing the university's perspective at a certain point in time. Each dataset

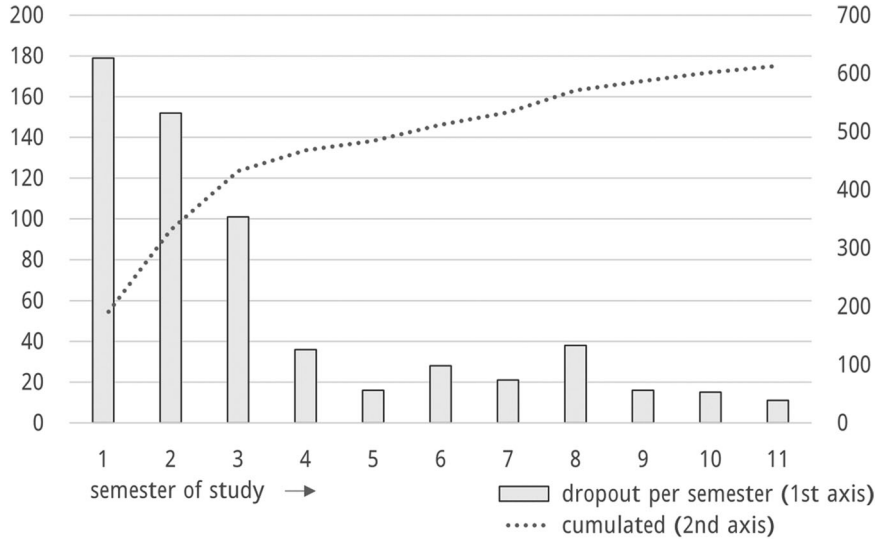

Figure 3. Aggregated and total number of student drop-out per semester.

contains only the data available *after* the respective semester. In the following, the models are built depending on the progress of the students. Models covering the same periods of semesters compete against each other.

In our scenario, we face the issue of an underrepresented minority class. Fortunately, only about a quarter of the students who begin their studies turn out not to finish it. In such an unbalanced data scenario, classification algorithms that are optimizing total prediction accuracy tend to calibrate towards detecting the specifics of the majority class (Chawla et al. 2002). In the case of student dropout this tendency is especially problematic as the minority class, the dropouts, are the class of particular interest. In order to tackle this problem, we create a second group of synthetically resampled datasets, on which the models are trained. For this purpose, we drop successful students while resampling new individuals utilizing the so-called synthetic minority over-sampling technique (*SMOTE*) (Chawla et al. 2002, 2008). This method uses bootstrapping and k-nearest neighbour analysis to synthetically create new examples closely positioned in between observations of the underrepresented category. Thereby, we strive to analyse whether over- and undersampling of the respective categories can produce superior predictions, especially with respect to the prediction of minority class cases.

Finally, we end up with six datasets which can be used to train our models (see Table 2).

Methodology

In our dataset a student's university career can result in *graduation* or *dropout*. Therefore, the problem can be stated as a supervised, binary classification model. During the training process, the algorithm optimizes a mapping function which is used to classify new unseen observations. The predicted results are then compared to real observations to compute the model's prediction performance. In this manner, the models can also be evaluated in their ability to generalize and predict the categories for new unseen data.

Table 2. Overview of different datasets (U: 'unbalanced', B: 'balanced').

Initial	Included semesters	Balanced
1U	1	no
1B	1	yes
2U	1–2	no
2B	1–2	yes
3U	1–3	no
3B	1–3	yes

The binary dependent variable y is therefore estimated by a function $f(\cdot)$. This function is calibrated using the examination data which represents the independent variables $x_m \in X^d$. The set of variable dimensions as well as the time horizons differ depending on the dataset used (see 3.2). Therefore, several models $f(\cdot)^d$ are trained. Each yields predictions for binarily-coded student outcomes y^d:

$$\hat{y}^d = f(X^d)^d = f(x) = \begin{cases} 0, & \textit{dropout} \\ 1, & \textit{graduation} \end{cases}$$

Logistic regression

Our first approach to model a function $f(\cdot)$ mapping the independent variables to a dropout prediction is logistic regression. In regression analyses this approach is frequently used to cope with the binary dependent variable violating the normality assumption in a typical linear regression (Aldrich and Nelson 1984). In this approach several parameters are calibrated in order to find an optimal fit of the feature variables to the dependent variable given a fixed functional form. In logistic regression this functional form is an S-curve in between the values of 0 and 1. How the S-curve is shaped is determined by estimating fitting parameter values, in our case using an iteratively re-weighted least squares method.

$$\hat{y} = f(x_1, \ \ldots, \ x_M) = I\left(\frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_M x_M)\right)}\right)$$

where $I(\cdot)$ is an indicator function that decides at which value $\hat{y}$ is set to 1 (or to 0). In order to perform a valid regression analysis, a few preliminary steps need to be taken. This includes handling of NA-data, excluding variables with few observations and the removal of correlation and multicollinearity among the predictors. Taking into account results of a recommended observation-variable rate (Peduzzi et al. 1996; Babyak 2004), unpopular examinations with less than 15 observations per outcome category were removed from the data set. This reduces the number of available examinations in the dataset from 487 to 29. This seemingly drastic reduction still leaves us effectively with more than 90% of all test results from the original dataset. This is due to the remarkably high freedom of choice at the advanced stages of the Industrial Engineering studies at KIT. The basic studies, however, are obligatory and comprise large numbers of graded tests.

Another essential precondition for the applicability of logistic regression is the absence of strong correlation and multicollinearity of the regressors (Aldrich and Nelson 1984). The existence of many correlated regressors, as in our case, poses the threat of overfitting the model (Babyak 2004). Therefore, we perform a pairwise-removal of highly correlated features (with a correlation greater than 0.9). Second, in order to remove multicollinearity we calculate the variance inflation factor (VIF) for each predictor (Kuhn et al. 2016).

Features are removed by backward selection of iteratively calculating VIFs (see Menard 1995; Kennedy 2003; Obrien (2007)). As demanded by Obrien 2007 the removal can also be justified theoretically, arguing that exams with correlated results are likely to measure roughly the same abilities. In consequence, exams may be dropped safely without the loss of crucial information.

Classification tree

Classification or decision trees are chosen as the second machine learning approach, since they are intuitive and easy to interpret. Decision trees owe their name to their tree-like structure. Represented as flowcharts, decision trees consist of three different types of nodes, the root node at the beginning/top, branches between the nodes and leaves at the end/bottom. At every internal node a specific feature of the input observation is tested to determine the further path (branch) down the tree structure until a final assignment is made at a terminal leaf-node. Because of this intuitive flow-chart structure, decision trees are easy to understand and easy to interpret. Also, decision trees work efficiently on large input data without the need to apply a complex parametric structure. Due to their usability and efficiency, decision trees have become one of the most effective and popular methods in machine learning since their introduction in the 1960s (Song and Ying 2015).

The starting point for the development of the decision tree model is, as with the regression model, the pivoted dataset of the respective semesters. In contrast to logistic regression, decision trees are less error-prone to NA-values (Song and Ying 2015). Consequently, our NA-handling turned out to be much easier. In addition, there is no need to analyse the dataset for correlation and multicollinearity. In fact, the removal of (multi-)correlated predictors may even reduce the prediction performance of decision tree models (Piramuthu 2008).

However, decision trees also come with disadvantages. Above all, they are prone to over- and underfitting, which can result in poor prediction performance. An underfitted model is typically represented by a small tree with very few decision levels that cannot express the data in sufficient detail. On the other hand, an overfitted model loses in its complexity the ability to make externally valid predictions. Both, over- and underfitting, lead to a poor generalizability making the model less robust.

Handling over- and underfitting

To produce models which do not overfit and therefore generalize to unseen data we employ two common approaches – *pruning* and *stopping*. Firstly, after a final, probably overfitted tree is calculated it may be shortened by 'cutting' the most detailed nodes. Cutting these nodes typically improves the generalizability of the model. Secondly, the learning process of the model may be stopped ahead of time preventing overfitting within the building process. To avoid a static threshold, such as a set number of tree levels, we chose a minimal leaf size as our stopping criterion. Here, leaves cannot contain less than a certain amount of observations at the end of the training process. This indirectly limits the possible levels of a tree, as every decision split increases the chance to reach the threshold (Song and Ying 2015).

Finally, in order to check the robustness of the approach, we monitor the model performance within a cross-validation framework. For this purpose, the dataset is split into 10 separate subsets each containing 10% of the original dataset.

Each fold is once used for testing, while the remaining 90% are used to train the respective tree. In the end, all 10 model accuracies are compared. If the results prove to be sufficiently stable, we shall be able to conclude that the tree building process can be generalized.

Model evaluation

Since in the application case Type I and Type II may be valued differently, we consult a variety of fit indicators for both the logistic regressions as well as for the classification trees. In order to find comparable evaluations, we choose indicators applicable to both model types.

Goodness of fit is evaluated by contrasting the predicted study outcomes to the actual occurrences of the labelled test-dataset. Predictions can take on four values:

- True positives (TP) count non-successful students which are correctly classified as dropouts.
- True negatives (TN) count successful students which are correctly classified as successful.
- False positives (FP) count successful students which are misclassified as dropouts (type II error).
- False negatives (FN) count non-successful students which are misclassified as successful (type I error).

In general, a functional prediction model should minimize the number of falsely classified individuals. The models can further be tuned towards the detection of a certain category. Focusing on one class is always combined with a change in the detection performance of another category. If the prediction of a special case should be emphasized, an improvement can be 'exchanged' by lowering the accuracy of another category. If a study considers the detection of dropout to be more important than the recognition of success, FN predictions are worse than FP predictions. Such a prioritization is useful when, for example, the costs of leaving the system exceed the costs of preventing the dropout of an at-risk student.

The list of priorities derives from the particular preferences of the practitioner. Conventional indicators that are typically derived from the four prediction outcomes are the following:

- The overall predictive precision is described by the accuracy (ACC) which measures the ratio of correct predictions to the total number of prediction cases.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- The sensitivity (SEN), or detection rate, is the probability that a dropout student is classified correctly. Since SEN measures the predictive performance for the dropout

category, it should be observed particularly in the resampled datasets.

$$SEN = \frac{TP}{TP + FN}$$

- The specificity (SPC) is the probability that a successful student is correctly defined as such. Therefore, SPC constitutes the opposite prioritization to SEN.

$$SPC = \frac{TN}{TN + FP}$$

- The precision (PRE) is defined as the conditional probability that an as dropout classified student is correctly classified. A larger number indicates that fewer students are suspected falsely and indicates a better prediction performance for the positive class. This measure may be especially interesting for the application case if incorrect predictions need to be avoided.

$$PRE = \frac{TP}{TP + FP}$$

The so-called ROC-curve includes both sensitivity and specificity as extreme cases of a range of possible priorities.[6] The curve portrays, how well the model performs depending on the valuation of a true-positives-rate in comparison to a true-negative-rate. In logistic regression this concept can be easily implemented by adjusting the thresholds of the indicator function as portrayed by Figure 4.[7]

For the decision tree model the certainty of a prediction for a student i, that is $P(\hat{y}_i = 1|X_i)$, can be estimated by finding the success-ratio of similar occurrences at i's respective terminal leaf node. A diagonal line connecting sensitivity and specificity marks the threshold of randomness regarding the ROC-curve. It is the most basic benchmark to evaluate the model evaluation versus guessing without information. The farther the ROC-curve is raised from the diagonal the better the particular model performs. Benchmarks are important in interpreting a model's goodness of fit. In our binary problem a random guess for example already yields a prediction accuracy of 50%. In an unbalanced data scenario this accuracy of a random guess is even significantly higher as the simple prediction rule favouring the more probable case 'all students succeed' correctly classifies about 76% = (1 - dropout rate) % > 50% of all instances.

Our last evaluation measure is Cohen's Kappa (KAP) (Cohen 1960). KAP incorporates a benchmark value of randomness comparing the models ACC with an expected accuracy ACC_r, which we would get predicting the majority category to all observations.

$$KAP = \frac{ACC - ACC_r}{1 - ACC_r}$$

A positive value indicates that the model improves the prediction relative to an assignment by chance. The literature is divided with respect to which KAP values accord to which level of agreement between the model and the real data Gwet (2014). A general indication as defined by Landis and Koch (1977) suggests a value of $0.4 - 0.6$ to be moderate, $0.6 - 0.8$ as substantial/good and $0.8 - 1$ to be (nearly) perfect agreement. Other researchers tend to demand even lower bounds to measure agreement (Gwet 2014).

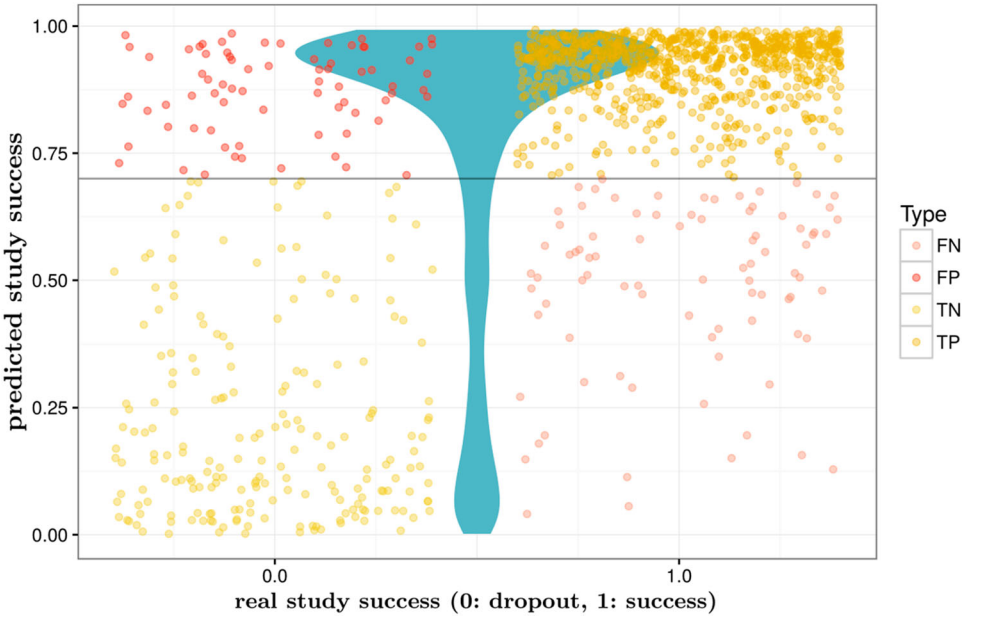

Figure 4. Classification of LR 1U predicted probabilities depending on indicator function (red line threshold). Dropout observations are scattered on the left (and vice versa).

Results

Overall, we find high prediction accuracies (ACC) reaching from a minimum of 85% on the smallest dataset (LR 1B) to a maximum of 95% after three semesters (DT 3U) Table 3.

Looking at Cohen's Kappa (KAP) all models make at least a 'substantial' improvement to the random model.[8] According to the classification of Landis and Koch (1977) the best models (LR 3U, DT 3U, DT) show a 'nearly perfect' agreement. Not just these extremes but all measures display increasing goodness of fit with increasing availability of data. One exception to this rule can be observed between the balanced data models DT 2B and DT 3B. The decline in model performance may, however, be explained by the randomness of sampling new data points to balance the dataset. Comparing the two models we find a slight edge in favour of the decision tree approach. Judging from the prediction accuracies only one decision tree is beaten by its counterpart (LR 1U). Yet, all measures are roughly equal, therefore we cannot speak of an evident winner. Especially, taking the trade-off between the true-positive and the true-negative rates into account, the performance between the model classes alternates too strongly to recognize a clear hierarchy. The comparability of the model performances is portrayed by the ROC curves of semester one models (see figure 5). Indeed, on first sight it is difficult to spot major performance differences between the model classes.

Regarding the differences in using balanced and unbalanced datasets, however, the evaluation measures show a clear pattern. All models trained on balanced data perform strictly worse in predicting successful students (SPC) while they perform strictly better in predicting student dropouts (SEN). Indeed, judging by the sensitivity, DT 3B is able to classify 89.6% of all dropouts correctly, which is about 9 percentage points better

Table 3. Evaluation measures of the logistic regression and the decision tree models for all datasets.

	Model	ACC	SEN	SPC	PRE	KAP	TP	TN	FP	FN
Logistic regression	LR 1U	0.881	0.629	0.960	0.828	0.642	173	853	36	102
	LR 2U	0.920	0.751	0.974	0.903	0.769	214	871	23	71
	LR 3U	0.945	0.785	0.984	0.924	0.816	256	1320	21	70
	LR 1B	0.845	0.775	0.866	0.642	0.598	213	770	119	62
	LR 2B	0.896	0.839	0.914	0.756	0.726	239	817	77	46
	LR 3B	0.918	0.862	0.932	0.755	0.754	281	1250	91	45
Decision tree	DT 1U	0.879	0.691	0.937	0.772	0.652	190	833	56	85
	DT 2U	0.930	0.765	0.982	0.932	0.796	218	878	16	67
	DT 3U	0.953	0.807	0.988	0.943	0.841	263	1325	16	63
	DT 1B	0.865	0.764	0.897	0.695	0.639	210	797	92	65
	DT 2B	0.933	0.828	0.966	0.887	0.813	236	864	30	49
	DT 3B	0.918	0.896	0.924	0.741	0.760	292	1239	102	34

than the decision tree model on unbalanced data (DT 3U). This difference in performance is especially prominent looking at the differences in SEN within the logistic regression models at semester 1. LR 1B outperforms LR 3U by about 15 percentage points. As expected, the improvement in detecting at-risk students with a balanced dataset is bought with a generally worse total prediction accuracy. Apart from DT 2B all balanced models are outperformed by their counterparts. This is due to the immensely high specificities reaching almost a value of 1. Also, as shown by the smaller PRE numbers balancing the datasets leads to a higher percentage in prediction errors given dropout is predicted. Observing the ROC curves these differences become apparent, when we focus on a single type of approach. Clearly, as portrayed by Figure 6, the balanced data models perform better in terms of sensitivity (left-bottom corner) while the unbalanced data

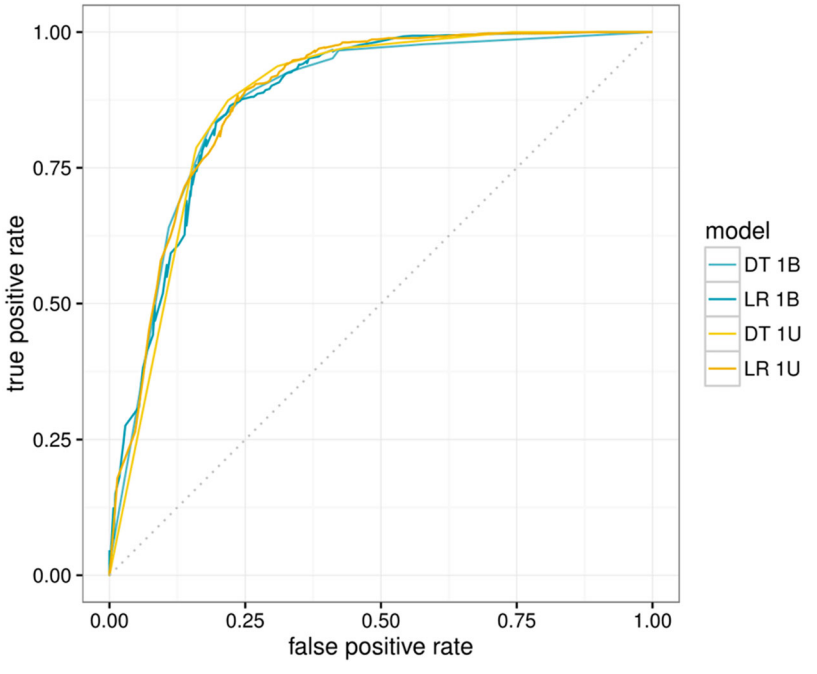

Figure 5. ROC curves of all models on first semester data.

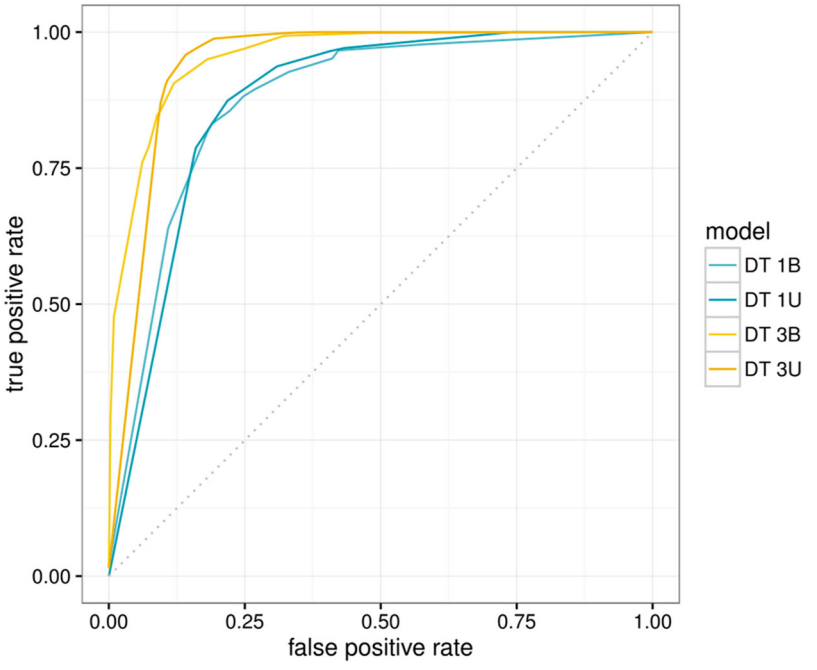

Figure 6. ROC curves of all decision tree models.

models take over in terms of precision (PRE), when a higher error rate in false success predictions is accepted (top-right corner). In addition, contrasting first and third semester models, the curve shows how the prediction accuracy increases as we utilize more data. This observation is not surprising as students that have already dropped out in earlier semesters can be spotted rather easily by the algorithm. As usual, in an actual application the designer has to decide if the gain in precision is worth the decrease in sensitivity. Balancing the dataset may be a helpful tool, if detecting dropouts is valued higher than misclassifying successful students.

Model interpretation

In the following, the influence of individual variables with regard to the interpretability of the models shall be examined in more detail. As the outcome structures differ by the general approach, the logistic regression models and the decision tree models are tackled separately.

Logistic regression

Exemplarily, Table 4 shows the regression parameters of the first semester model on unbalanced data.[9] It is easily apparent that the model comprises only a comparably low number of regressors, as at this moment only a few exams have been taken. On top of that, many variables are missing due to the removal of multicollinearity, which is especially prevalent in the first semester. Since so many variables have been left out, it is important to interpret these results with care. For example, the average grade appears only in the third-semester

Table 4. Regression parameters of the LR 1U model.

Variable	Parameter values[13]	Variable	Parameter values
(Intercept)	−4.149***	X57521_sem	0.166
Age	0.102**	X57522_grade	0.177*
Origin (Ger = 1)	0.742*	X65018_grade	−0.403**
Sex (m = 1)	−0.271	X65018_sem	2.569***
count_npass	−0.892***	X67503_grade	0.202*
X30009_attempt	−0.399	X70041_attempt	1.223***
X32723_attempt	−0.770**	X77001_grade	−0.212*
X32725_attempt	−0.293	X77001_attempt	1.217***
X57512_grade	0.105	X77508_grade	−0.03
X57521_grade	0.041	X77508_attempt	0.628

data. The low number of exams leads to a high correlation between the performance in certain single exam grades and the total average grade. Therefore, the variable is excluded from the dataset beforehand. Nevertheless, most probably it is a fallacy not to rate the average grade as a powerful indicator for study success. Also, included features that significantly influence the model prediction may, in fact, measure the real effect of correlated but excluded features. Even personal features such as *age* or *nation* vary when exam variables are in – or excluded. They even lose significance in some models. Consequently, we will voluntarily forgo to interpret the influence of individual features in detail. In general, however, we can observe that the combined variables, especially the count of failed exams, prove to have the highest impact on all models.

Decision trees

The decision tree trained on the 2U dataset is exemplarily displayed in figure 7. An internal node splits the data by testing a parameter on a threshold displayed below the node. If the test result is positive, the observation is moved further left in the graph, if not, the data point follows the branch to the right. Every node displays four internal values. The first is the assigned category at the current state, which indicates the final assignment if the node were a leaf-node. Second and third, the *purity* of the current data subset is displayed. The purity measures the percentages of '0' and, respectively, '1' classifications among all observations that pass through the particular node. The last value is the node *size*. It measures the percentage of observations that pass through the particular node relative to all observations of the training dataset (Milborrow 2017). In tendency, an impression of a variable's relative importance can be gained by combining these measures. The so-called score of a node is calculated by multiplying its purity and its size (Song and Ying 2015).

As an example, 7 shows the decision tree built on the 2U data. The feature 'count of passed exams' (*count_pass*) constitutes the tree's root node. If a student's parameter for this feature is above 10, the observation is immediately classified as 'successful', as the right branch leads directly into a leaf-node with leaf category '1'. This classification is performed with a high confidence since this leaf has a purity of 94% for successful students.[10] Looking at this far-right leaf node, its undermost value shows that 76% of all observations passed-on directly from the root node. The remaining 26% move into the root node's left branch. This path is taken, if the *count_pass*-variable takes on a value below 10. From there on, the next features are repeatedly tested until the observation reaches a leaf-node. Finally, after having passed the tree completely from top to bottom, a final classification is made.

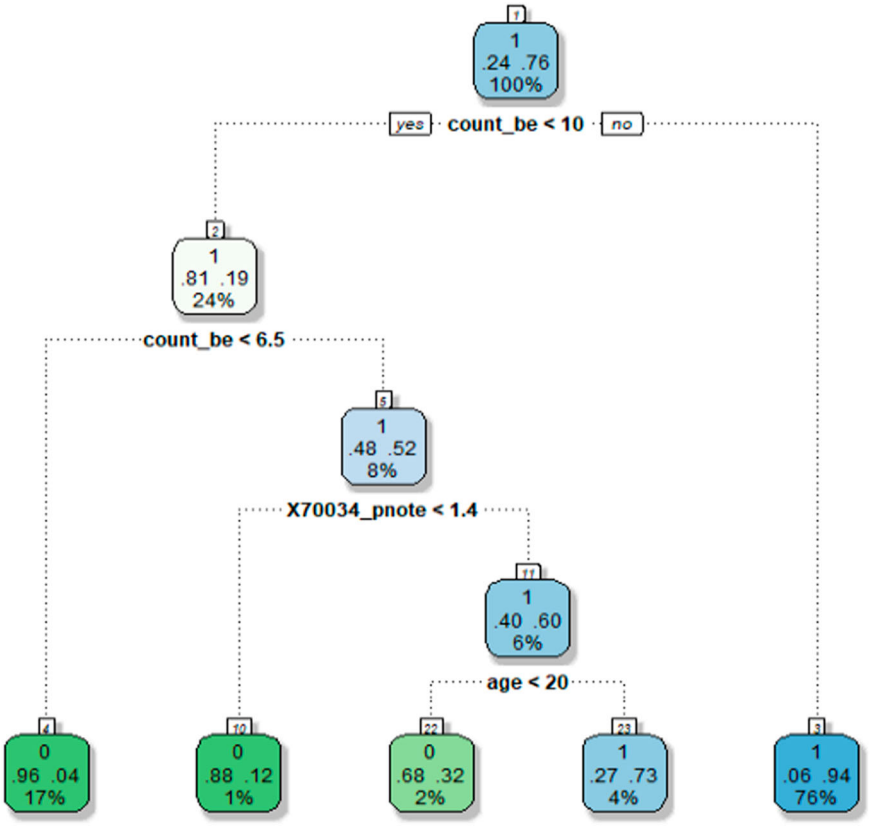

Figure 7. DT: Model Plot, data: 2U (count_be = count_pass).

Table 5 shows the aggregated variable scores over all models. Thereby, we can observe the most important variables that make up the decision tree structures.

In general, the decision trees are consistent with the regression models regarding the discriminatory power of extracted variables (e.g. average grade *avg_grade*, the count of passed examinations *count_pass* and the passed examinations per semester *avg_pass*). However, also individual examinations find their way into the tree structures – among them *Material Science* (*X77508, semester 1*), *Operations Research* (*X40009, semester 3*), *Business Administration I* (*X30009, semester 1*) and *Informatics I* (*X45006, semester 2*).

Overall, the decision trees prove to perform well, focusing their split on extracted variables. Regarding the anticipated challenge of under- and overfitting they remain robust, yet not unaffected. As portrayed in Table 6, a 10-fold cross-validation shows moderate variation in prediction accuracies depending on the subset of the data on which the model is trained (90%) and tested (the remaining 10%).

Practical validation

In order to test the *external validity* of the prediction performance, additional models are built. The models are trained on top of data up to a certain semester. Afterwards, they are tested with data recorded during the *next* suitable semester. The model evaluation on top

Table 5. Score (size × purity) over all models.

Variable[14]	Age. score	Occurrences	Variable	Agg. score	Occurrences
avg_grade	3123	7	X70034	138	3
count_pass	2261	4	X57523	116	1
avg_pass	1684	4	age	115	3
X77508	1609	4	avg_npass	81	1
X40009	1341	1	X30010	60	1
X30009	1020	4	X70035	47	1
X45006	961	3	X77002	39	1
count_sem	818	2	X67503	38	1
X32725	325	3	X45007	37	1
X70041	226	1	X57522	35	1
X57512	223	1	X57521	32	1
X77001	203	4	X72504	25	1
X65018	198	3	X32723	12	1

Table 6. Accuracy (ACC) values of a 10-fold cross-validation on unbalanced datasets.

Model	fold-1	fold-2	fold-3	fold-4	fold-5	fold-6	fold-7	fold-8	fold-9	fold-10
DT 1U	0.888	0.897	0.846	0.905	0.846	0.862	0.862	0.812	0.845	0.855
DT 2U	0.940	0.932	0.873	0.890	0.932	0.924	0.958	0.898	0.924	0.932
DT 3U	0.940	0.928	0.934	0.928	0.898	0.940	0.964	0.922	0.892	0.910

Table 7. Prediction evaluation of decision tree models tested on the subsequent semester data. Data: 1U.

Semester	ACC	SEN	SPC	PRE	KAP
10.01.2009	0.893	0.588	0.954	0.714	0.583
10.01.2010	0.908	0.433	0.980	0.763	0.505
10.01.2011	0.897	0.257	0.990	0.792	0.347
10.01.2012	0.862	0.378	0.994	0.944	0.476

of data of future cohorts can be used to test how well the models would have performed in reality. Therefore, this perspective serves as a practical setting to evaluate how well the approach generalizes; that is, how the results vary for different training and test sets that reflect altering circumstances within the university. For example, the first row of Table 7 is computed by training a model only on data up to 2008. This model is then tested by predicting the outcomes for data gained solely during the semester beginning in October 2009. Thereafter, the test and training data windows change accordingly.[11]

Even though the accuracies vary, the results appear to be sufficiently stable to conclude the method's external validity (see Table 7). As the general DT 1U model reaches an accuracy of 0.879 the yearly prediction models appear to even outperform the general case. This perception, however, proves to be overly optimistic when considering the lower values of Cohen's Kappa. Indeed, the KAP values reflect the fact that dropout rates in the KIT's industrial engineering programme have gradually fallen over the years. Therefore, the predictions have become easier as compared to the random guess. Nevertheless, the prediction quality remains meaningful. The shrinking minority class also leads to a model orientation in favour of the majority class. This is reflected by comparably high specificities and low sensitivities. This observation underscores the importance of balancing the data in such use cases.

Conclusion

The first and general finding is that the prediction of study success and dropout is possible on the basis of individual student transcript records. Since nowadays these data are digitally recorded at virtually all institutions of higher education, approaches similar to ours can be implemented at other institutions without substantial upfront costs. In addition, our approach has two beneficial properties from a broader socioeconomic point of view. Firstly, the data used is directly related to individual student performance and, therefore, does not require any further collection of possibly privacy-related information. Secondly, because our approach uses only study records, it does not pre-discriminate between students based on non-study-related information. Both properties should make the present approach socially more acceptable. The fact that the future performance of students can best be predicted by their previous performance goes in line with the findings of Heublein (2010) that student dropout is essentially a performance issue. An approach confining attention to student record data is sufficient to identify possible dropouts.

Since it was possible to identify individual dropout probabilities, the results can be used to establish specific retention policies, such as assistance programmes for at-risk students. By identifying those students, resources can be assigned efficiently to students who would otherwise drop out with high probability.

The methods employed in this paper, logistic regressions and decision trees, produce promising results with respect to their prediction accuracy. Further, we have shown that resampling the dataset can be used to improve the identification of dropout cases, as they are underrepresented relative to successful students in the initial dataset. Downsides of logistic regression models are more complex dataset preparation and model interpretation which both leave room for potential errors. Regarding decision trees, dataset preparation is less complicated, and the models proved to be easier to compute and to interpret.

The most relevant single factors to predict study success and dropout are combined features such as the count and the average of passed and failed examinations or average grades. This implies that, in our case, we can reduce the models' complexity considerably, without losing too much predictive power when focusing on few but powerful extracted features. It should be noted, however, that the analysis was conducted on data of one specific study programme. Therefore, the results cannot be directly transferred to other institutions or courses. Rather than the results as such, it is the technique that led to the results, which should be transferable to other institutions of higher education.

Notes

1. In German the technical subjects are referred to as 'MINT' fields.
2. This does not include forced dropout by failing an examination.
3. The students that had still been enrolled in 2016 or were pausing their studies have not been considered in this analysis.
4. The data were made available and released under the approval of the data protection office of the KIT and in compliance with all data protection guidelines.
5. To ensure data privacy age data were split up in classes. The number of years per class was increased for older cohorts with fewer observations, so one could not infer from the year to an individual student. For the same reason the origin data were classified in German and non-German,

6. Statistically speaking, the curve portrays the prioritization of type I versus type II errors.
7. The blue shape displays a mirrored density distribution of the predicted probabilities for study success.
8. With LR 1B scratching at the 0.6 threshold.
9. The X … variables refer to the semester, attempt or grade of a certain examination.
10. The threshold for the split-test itself is relatively high, which is explained by additional pre-examinations during the first semester.
11. Using classification trees based on unbalanced data from only the freshmen years (1U). Data from the summer-term semesters are excluded because they contain only very few individuals. Also, data from 2013 onwards is not used for testing since most of those students are still enrolled and the final status is yet unknown. Predictions of these semesters are not compared or evaluated.
12. X***** is an encoding for a certain examination (for example, X45006 refers to *Informatics I*),
13. Significance levels: *0.05, **0.01, ***0.001,
14. All single observations (attempt, grade, semester) are aggregated per individual examination.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the German Federal Ministry of Education and Research under [Grant 01PB14008]. The responsibility for the content of this publication lies with the author.

Notes on contributors

Lorenz Kemper has been research associate at the Chair of Public Finance and Public Management. His research lies in the areas of econometrics, data science and institutional economics. In 2019, Lorenz Kemper defended his dissertational thesis about econometric works in higher education economics.

Gerrit Vorhoff has been a research fellow at the Chair of Public Finance and Public Management at the Karlsruhe Institute of Technology. He has a background in industrial engineering and works as a consultant for business processes and digitization.

Berthold U. Wigger heads the Chair of Public Finance and Public Management at the Karlsruhe Institute of Technology. He is an expert in public finance and focuses his research on taxation and public expenditure. Berthold Wigger is a member of the scientific advisory council at the Federal Ministry of Finance and the Kronberger Kreis of the Stiftung Marktwirtschaft.

ORCID

Lorenz Kemper http://orcid.org/0000-0002-0479-7155

References

Aldrich, J. H., and F. D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Vol. 45. Sage.
Aulck, L., N. Velagapudi, B. Joshua, and W. Jevin. 2016. "Predicting Student Dropout in Higher Education." ICML Workshop on Data4Good: Machine Learning in Social Good Applications, 411–421.

Babyak, M. A. 2004. "What You See May Not Be What You Get: A Brief, non-Technical Introduction to Overfitting in Regression-Type Models." *Psychosomatic Medicine* 66 (3): 411–421.
Caison, A. L. 2007. "Analysis of Institutionally Specific Retention Research: A Comparison Between Survey and Institutional Database Methods." *Research in Higher Education* 48 (4): 435–451.
Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–357.
Chawla, N. V., D. A. Cieslak, L. O. Hall, and A. Joshi. 2008. "Automatically Countering Imbalance and Its Empirical Relationship to Cost." *Data Mining and Knowledge Discovery* 17 (2): 225–252.
Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 10 (1): 37–46.
Dekker, G. W., M. Pechenizkiy, and J. M. Vleeshouwers. 2009. "Predicting Students Drop out: A Case Study." Proceedings of the 2nd International Conference on educational data mining, 41–50.
Delen, D. 2010. "A Comparative Analysis of Machine Learning Techniques for Student Retention Management." *Decision Support Systems* 49 (4): 498–506.
Djulovic, A., and D. Li. 2013. "Towards Freshman Retention Prediction: A Comparative Study." *International Journal of Information and Education Technology* 3 (5): 494–500.
Gansemer-Topf, A. M., and J. H. Schuh. 2006. "Institutional Selectivity and Institutional Expenditures: Examining Organizational Factors That Contribute to Retention and Graduation." *Research in Higher Education* 47 (6): 613–642.
Gwet, K. L. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*, 4th ed. Advanced Analytics, LLC.
Hetze, P. 2011. "Nachhaltige Hochschulstrategien für mehr MINT"-Absolventen. Technical report, Stifterverband für die Deutsche Wissenschaft; HeinzNixdorf-Stiftung.
Heublein, U. 2010. "Ursachen des Studienabbruchs in Bachelor- und in herkömmlichen Studiengängen." Deutsches Zentrum für Hochschul- und Wissenschaftsforschung.
Heublein, U. 2014. "Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen: statistische Berechnungen auf der Basis des Absolventenjahrgangs 2012." Deutsches Zentrum für Hochschul- und Wissenschaftsforschung.
Heublein, U., H. Spangenberg, and D. Sommer. 2003. "Ursachen des Studienabbruchs." Analyse 2002. Technical Report, Hannover: HIS.
Kennedy, P. 2003. *A Guide to Econometrics*. MIT press.
Kovacic, Z. 2010. "*Early Prediction of Student Success: Mining Students' Enrolment Data.*" Proceedings of Informing Science and IT Education Conference, 647–665.
Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, et al. 2016. "Caret: Classification and Regression Training." R package version 6.0-73.
Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics*, 159–174.
Menard, S. 1995. *Applied Logistic Regression Analysis: Sage University Series on Quantitative Applications in the Social Sciences*.
Milborrow, S. 2017. "rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'." R package version 2.1.2.
Obrien, R. M. 2007. "A Caution Regarding Rules of Thumb for Variance Inflation Factors." *Quality & Quantity* 41 (5): 673–690.
Office, F. L. 2017. "Fachkräfteengpassanalyse." Technical report, Arbeitsmarktberichterstattung.
Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. 1996. "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis." *Journal of Clinical Epidemiology* 49 (12): 1373–1379.
Piramuthu, S. 2008. "Input Data for Decision Trees." *Expert Systems with Applications* 34 (2): 1220–1226.
Ram, S., Y. Wang, S. A. Currim, and F. A. Currim. 2018. "Predicting student retention using smart-card transactions." United States Patent Application Publication. US Patent 2018/0144352 A1.
Song, Y.-Y., and L. Ying. 2015. Decision Tree Methods: Applications for Classification and Prediction." *Shanghai Archives of Psychiatry* 27 (2): 130.

Spence, M. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87 (3): 355–374.
Tinto, V. 1975. "Dropout From Higher Education: A Theoretical Synthesis of Recent Research." *Review of Educational Research* 45 (1): 89–125.
Yu, C. H., S. DiGangi, A. Jannasch-Pennell, and C. Kaprolet. 2010. "A Data Mining Approach for Identifying Predictors of Student Retention From Sophomore to Junior Year." *Journal of Data Science* 8 (2): 307–325.
Yukselturk, E., S. Ozekes, and Y. K. Türel. 2014. "Predicting Dropout Student: an Application of Data Mining Methods in an Online Education Program. *European Journal of Open Distance and E-Learning* 17 (1): 118–133.
Zhang, Y., S. Oussena, T. Clark, and H. Kim. 2010. "*Using Data Mining to Improve Student Retention in Higher Education: A Case Study*." International Conference on Enterprise information systems.