I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

RESEARCH REPORT SERIES

IZA Research Report No. 89

Early Identification of College Dropouts Using Machine-Learning

Conceptual Considerations and an Empirical Example

Ingo E. Isphording (IZA)
Tobias Raabe (IZA)

JUNE 2019

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Disclaimer and acknowledgements: This report is part of the research project PIRanHa, funded by grant 01PX16013 from the German Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF). The sole responsibility for the report's contents lies with the authors. This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort First-Year Students, doi:10.5157/NEPS:SC5:11.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. We would like to thank Christoph Schwerdtfeger, Arnim Seidlitz, Pamela Qendrai and Florian Wozny for support in the data preparation and analysis.

1. Introduction

Can and should we use modern big data methods to predict which students will drop out of their studies? Such an early detection of at-risk groups of students would allow practitioners and policy-makers to timely intervene through targeted counter-measures as for instance mentoring, counseling, or institutional support. This report discusses the feasibility, benefits and potential hidden costs of such an approach which would allow to mitigate substantial costs of educational career frictions.

Resources of educational production – monetary as well as non-monetary, public as well as private – are scarce and should be efficiently invested into human capital acquisition. College dropout is often viewed as a misallocation of those resources. While from an individual perspective college dropout might be a rational correction of the educational career, it is associated with societal costs that are difficult to quantify. Costs will be specifically high in subjects where demand for tertiary education exceeds the number of college places and future college dropouts prevent more suitable candidates from beginning their studies. In addition, college dropout is still linked to a societal stigma of failure – despite the fact that newspaper headlines rather focus on success stories of college dropouts.

Underlying economic changes increase the demand for academically qualified workers. Ongoing digitalization of workplaces and demographical trends of an aging society lead to ever-growing demand for academically trained workers. Accordingly, debates about skill shortage are closely related to discussions about how to increase the share of academically qualified workers. Increasing the number of academically trained workers cannot entirely be achieved through pushing more students into college: the stabilization of individual educational careers has to receive similar attention.

To support this goal, this project aims at evaluating the feasibility of an early warning system of student dropout. Using data from a national representative student survey, we conduct a "horserace" of different machine-learning-based prediction models to understand (1) to what extent more complex models yield a higher prediction quality, and (2) how far acquiring more in-depth data about the student body produces more precise predictions. We compare prediction results from different models according to their ability to precisely identify at-risk students while keeping the rate of students falsely classified as potential dropouts as low as possible.

Our results show that machine-learning-based prediction allows for a significant increase in prediction quality compared against naïve classification approaches (e.g. using population shares of dropout as dropout probability for any student), while the precise choice of the prediction model is not overly affecting prediction quality. Rather straightforward to implement machine-learning methods which maintain a reasonable level of interpretability should be preferred as they allow practitioners to infer the sources of dropout and to tailor intervening measures. Prediction quality does, however, vary substantially with the amount of available data. While even focusing on the most basic student characteristics that are part of university registers can predict student dropout to some degree, additionally collecting subjective information at the point of enrolment, and more so in later semesters, can lead to distinct increases in prediction quality. A prediction model based on a so-called random forest approach and relying on an extensive set of basic student characteristics and subjective performance indicators yields a considerable increase in prediction quality. This increase covers a substantial part of the difference between *naïve classification* (basic guessing based on population shares) and *perfect knowledge*.

We discuss the usefulness of prediction results for cost-efficient roll-out of intervention measures, such as counselling and self-commitment devices. The design of a prediction model necessarily includes a trade-off between the number of at-risk students correctly identified and the number of students falsely identified as at risk of dropout. Which model to choose will accordingly be influenced by the cost structure of dropout – how costly student dropout is from a welfare perspective and how cost-efficient potential counter-measures are expected to be.

We further discuss our results in relation to additional (hidden) costs of data acquisition. High levels of mistrust against data collection and strong privacy concerns among the student body and administration might render additional data acquisition unfeasible, specifically for items that are perceived as sensitive information. Here, arising implementation costs have to be evaluated against the increase in expected prediction quality.

2. Background and Literature

With this study, we contribute to a long-standing literature in social and pedagogical sciences on understanding the heterogeneity in students' study persistence and predicting the occurrence of student dropout. The former part of this literature asks questions about the underlying factors and determinants that increase or mitigate the risk of dropping out of college. In general, the literature aims at understanding the role of different individual and institutional factors in student dropout decisions. Studies in this field often rely on established theoretical frameworks. For example, Tinto (1975) explicitly describes causal chains between a student's own characteristics (the 'study capital'), contextual factors stemming from peers and social environment and institutional background factors and the decision to drop out. Despite the causal claims made by these models, and explicitly causally framed questions ("What is causing student dropout?''), the literature is in general lacking quasi-empirical strategies to separate causal effects from student self-selection and unobservable confounding factors. Notable recent exceptions applying quasi-experimental methods to establish causality between single determinants and dropout are Stinebrickner and Stinebrickner (2014a,2014b), Adamopoulou and Tanzi (2017) and Horstschräer and Sprietsma (2015).

A second strand of literature focuses on *predicting* student dropout, instead of understanding its sources. From a practitioner's perspective, accurately assessing the risk of a student to drop out enables university administrations to address at-risk groups with meaningful interventions. In recent years, this literature has strongly benefited from the emergence of an evermore powerful machine learning methodology. The main comparative advantage of this methodology is to provide accurate out-of-sample predictions using information from earlier stages of a certain educational path. Additionally, improved access to administrative data for forecasting analyses has led to significant advances in this field.

Sara et al. (2015) predict high-school dropout using large scale Danish data. Based on basic demographics, school properties, household income and students' performance collected during the first six months of high school, they predict whether the students will drop out in the three subsequent months. Among the applied prediction methods (Support Vector Machine, Random Forest, CART and naïve Bayes classifier techniques), random forest provided the most accurate prediction. Sansone (2018) develops an algorithm that identifies students that are at risk of dropping out using high-school longitudinal data from the US. The prediction is based on students' information from their first year of high school to predict the dropout later on in secondary education. Compared to traditional methods, techniques such as Support Vector Machine, Boosted Regression and Post-LASSO are found to be more accurate in predicting dropouts. Berens et al. (2018) use administrative data on undergraduate students students from a private and a public university in Germany to predict student dropouts. Similar to our own approach, they employ machine learning techniques such as decision trees, neural networks and random forest as well as ensemble methods combining the predictive power of these three techniques. The most accurate prediction is obtained utilising an ensemble algorithm that summarizes results from several single prediction models. Using only demographic determinants at the time of enrolment, the algorithm provides substantially better prediction results than naïve classification. Surprisingly, adding information on students' performance does not further improve the prediction accuracy in their case.

These exemplary studies show that machine learning techniques seem to be complementary to traditional methods and can facilitate the detection and prediction of students that are at the risk of dropout. We add to these studies a comprehensive perspective using representative data, and explicitly taking a practitioner's perspective on the prediction exercise. To do so, we distinguish data by its (potential) availability and costs of acquisition. We further distinguish between different definitions of student dropout that are relevant for different kinds of stakeholders. To address an audience of practitioners, we maintain a non-technical language throughout most of the paper.

3. Data and Method

3.1 Data

We base our analysis on the starting cohort 5 of the National Educational Panel Survey (SC5/NEPS) provided by the Leibniz-Institut für Bildungsverläufe e.V. (LIfBi). This data set targets the population of students who started their studies in the winter semester 2010/2011 at a German university. Initial sampling was based on a mixture of roster-based sampling and snowball procedures. Students are bi-annually surveyed by computer- and telephone-assisted interviews.

We construct a student × term level data set that tracks a student's progress by subject and institution. In each term, we therefore know whether a student is studying the subject she has initially chosen and whether she is still studying at the original institution. If not, we track whether she has received a degree or has dropped out of her studies without degree. Further, the data includes information on both time-invariant characteristics fixed at enrolment and time-variant characteristics monitored throughout the educational career.

To track the progress of students, we rely on detailed spell information where each spell covers an episode at a tertiary education institution. For each change in episodes, we determine whether a student has obtained a degree in the previous spell and therefore successfully completed her studies. If not, we document whether the transition leads to a change in subject and institution and record the time between the spells. If a spell is incomplete and there is no later spell, we document whether it would have been possible for the individual to report a new spell in a following interview. If it was not possible for the individual to report a new spell, we cannot determine the outcome of the student. If a new spell was possible, we assume that individuals always report new spells and document the time since the last spell.

To determine when a dropout has occurred, we apply three different definitions of dropout that differ in their scope and relevance for different stakeholders:

Definition 1: Any student who ends a study program non-successfully and does not begin new studies of any subject at the same or another institution. Definition 1 displays the narrowest definition of dropout. This definition takes the policy-makers' macro perspective on in- and out-flows into the education system. This definition is relevant for general targets of academization, i.e. increasing the number of college-educated individuals.

Definition 2: Every student who ends a study program non-successfully and does not start a study program of the same subject at the same or another institution. This definition again displays a macro perspective on student dropout and disregards the role of a single institution. This definition is specifically relevant whenever shortages of labor supply can be traced to specific fields, e.g. against the background of a lack of STEM (science, technology, engineering and math) students.

Definition 3: Every student who ends a study program and does not start a new study program at the same institution in the same subject[1]. This definition takes a more micro perspective of a single department of an institution. This definition is relevant from the perspective of a department head who aims at minimizing dropout from a programme which can be seen as a quality indicator. Similar definitions have been applied in previous case studies relying on data from single programs (e.g. Berens 2018).

We link these outcomes to a broad range of student level characteristics. We categorize predictors into four different variable subsets. These subsets describe information that is 1) available to each German university out of legal reasons at college entry, 2) could potentially be acquired through in-depth student self-assessments at college entry, 3) could potentially be assessed through student progress assessment

[1] Due to limitations of the data, we do not have the exact subject of a student and therefore must rely on a subject group variable. Furthermore, educational episodes in the spell data are continued for institution changes if the main subject stays the same. Thus, there are potential student dropouts which are misclassified as graduates.

or 4) is acquired for the purpose of legally mandated student monitoring. In the following, we describe these sets of variables in more detail.

Basic student characteristics are collected by each university to fulfil the requirements of the *Hochschulstatistikgesetz* (Tertiary Education Statistics Act). Hence, these variables are directly available for any student starting her studies at a German university. This first information set includes basic socio-economic background information: gender, age, month of birth, nationality, chosen secondary subjects, intended degree, high school GPA and year of high school graduation.

An **Initial student assessment** can be used to allow universities to gain further insights into their student body. Such questionnaires might use well-known survey inventories to learn about a student's personality, motivation and beliefs, information that would not be readily available through register sources.[2] Additionally, student entry tests might be used to test students' cognitive skills. These data are substantially more expensive to acquire and might lead to privacy and data security concerns. This information set contains a large number of variables with respect to the student's socio-economic background (parental education and occupation, household structure, language of origin), her personality, educational career, own and parental subjective educational attitudes and cognitive test scores.

Student progress assessment, e.g. through mandatory meetings with a student counsellor or through online questionnaires, could be used to elicit subjective indicators of student progress. This information set contains subjective assessments about the satisfaction with one's studies, expected returns of studying after graduation, the probability of graduation and dropout. Student progress assessments will display a significant investment into data acquisition for most universities. Additionally, like in the case of self-assessments, students might raise privacy concerns about revealed information and constraints to academic self-determination.

Student monitoring is again standard in all universities which are required for legal reasons to collect a students' progress (grades and credit points) and course choices. This data should be readily available with a student progressing through her studies. To approximate this information set, we rely on a student's current GPA.

Basic descriptive statistics of a subset of core variables are summarized in Table 1. Comparing these numbers with representative numbers from official student statistics reveals that the data is comparable in most characteristics, although we observe a larger share of female students (60 vs 50 percent), a smaller share in law, economics and social sciences (25 vs 32 percent), and a smaller share of students at universities of applied sciences (colleges) of 22 vs 33 percent in the representative student statistics.

[2] As an example of such student self-assessments, the TH Nürnberg gathers data for each applicant through an in-depth online assessment. The data has been used in Himmler/Jäckle/Weinschenk (2019).

Table 1: Descriptive Statistics

	Mean	SD	Min	Max	N
A. Individual Data					
Personal characteristics					
Female	0.60	0.49	0.00	1.00	9809
Age	21.40	3.53	13.00	65.00	9809
Migrant	0.20	0.40	0.00	1.00	9809
High school GPA	2.21	0.62	1.00	4.00	7786
No Abitur	0.05	0.21	0.00	1.00	9809
School in Germany	0.00	0.00	0.00	0.00	9809
Highest parental education					
Middle school	0.02	0.13	0.00	1.00	9809
Secondary	0.54	0.50	0.00	1.00	9809
Tertiary	0.44	0.50	0.00	1.00	9809
B. Subjects					
Agriculture, forest and food sciences	0.02	0.15	0.00	1.00	9809
Human and health sciences	0.02	0.14	0.00	1.00	9809
Engineering	0.16	0.37	0.00	1.00	9809
Art and sciences of art	0.03	0.16	0.00	1.00	9809
Mathematics and natural sciences	0.24	0.43	0.00	1.00	9809
Law, economics and social sciences	0.26	0.44	0.00	1.00	9809
Sport sciences	0.01	0.12	0.00	1.00	9809
Linguistics and cultural sciences	0.26	0.44	0.00	1.00	9809
C. Institutional characteristics					
University	0.74	0.44	0.00	1.00	9809
College	0.23	0.42	0.00	1.00	9809
Private	0.03	0.16	0.00	1.00	9809
Admission restriction: grades	0.91	0.29	0.00	1.00	6029
Admission restriction: test	0.19	0.39	0.00	1.00	5981
D. Degree					
Bachelor	0.65	0.48	0.00	1.00	9809
Bachelor (Lectureship)	0.14	0.35	0.00	1.00	9809
State examination	0.04	0.19	0.00	1.00	9809
State examination (Lectureship)	0.16	0.37	0.00	1.00	9809

***Notes**:* This table displays means and standard deviations of core variables.

3.2 General approach: the prediction horserace

The aim of this project is to analyze the feasibility and expected performance of machine-learning-based early detection algorithms aiming at forecasting at-risk groups of student dropout. Specifically, we want to answer the question to what extent already available data provides sufficient levels of prediction quality, and which data would have to be acquired to further increase the quality of prediction. We address these questions by applying a range of different prediction models and specifications in what we label as a *prediction horserace*. We test a large number of prediction models that differ in the information that is used and also with respect to the applied method in order to elicit structural differences in prediction quality and variable importance across models. These differences allow us to elicit generalizable regularities. We use these regularities to provide advice for the implementation of prediction models in practice.

3.3 Models

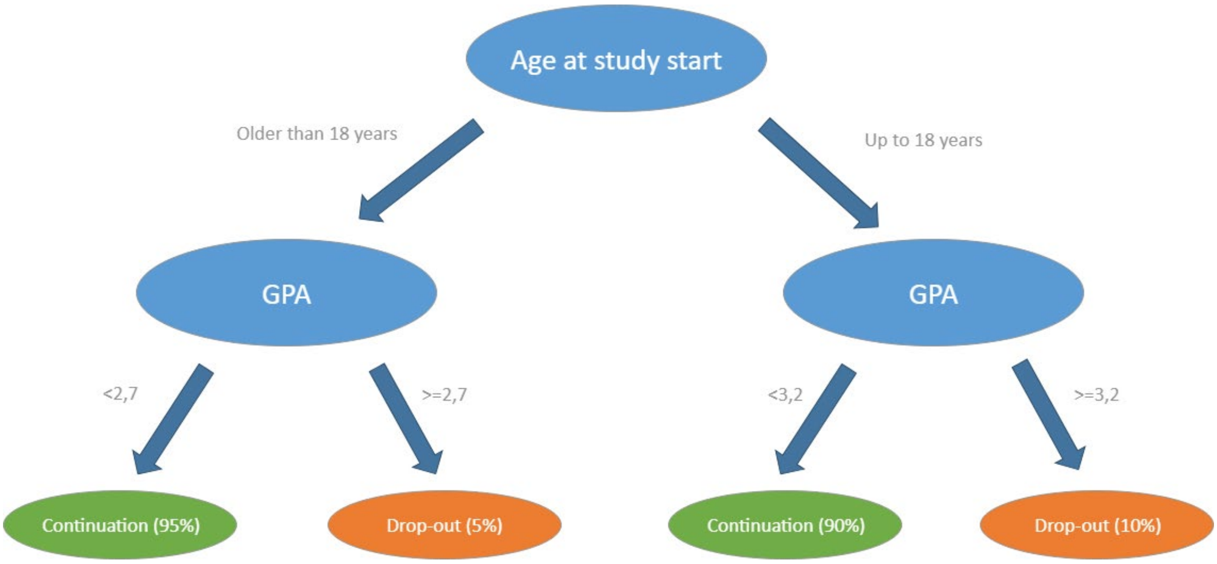
For the *prediction horserace*, five different models of increasing complexity compete with each other in terms of prediction quality. The models range from a baseline OLS regression to an ensemble method incorporating results from several complex prediction methods. With increasing complexity, we face a trade-off between more precise prediction and loss of interpretability of the results. In the following, we describe the different applied methods as intuitively as possible.

OLS regressions are the workhorse of empirical social sciences and are characterized by high levels of interpretability and robustness. Dropout as dependent variable is modeled as a simple linear combination of a group of independent variables. Coefficients can directly be interpreted as partial correlations between individual variables and the outcome. The estimation of the OLS model is based on the idea of minimizing the deviations between predicted and observed values in the present data. By doing so, OLS optimally adjusts to the training data. In such a case, the estimated model would very accurately describe the data it was fitted ("trained") on, but would perform badly in predicting dropout in new samples. This property might not be desirable for a prediction exercise as in our case, because it leads to "over-fitting" of a model. Over-fitted model parameters might represent particularities of the training sample instead of actual underlying structural relationships. In turn, this leads to low out-of-sample prediction quality.

Logistic LASSO (Ridge) regressions differ from simple OLS in two ways. First, a logistic regression does a better job in capturing the non-linear relationship between a binary variable (such as dropout yes/no) and a set of independent variables. Second, the regularization of LASSO (or Ridge) directly addresses the over-fitting problem of OLS regressions. It does so by not only minimizing deviations between predicted and observed dropout, but also by holding the number of variables used for the prediction as small as possible. Thus, the estimated model is forced to maintain a higher level of generalizability by adding a penalty for coefficient size to the minimization problem.

Tree classification methods are the most common and successful machine-learning methods used to classify observations according to a categorical outcome. A tree classifier estimates a decision tree that splits the sample at each node by a binary decision rule and by doing so narrows down the likeliest outcome for each subpopulation. Figure 1 displays an example of a simplified decision tree. Each node of a tree represents a decision rule. End points of the tree are classified as dropout / success by the share of dropouts in the respective sub-population. Single predictors can have different roles / classification thresholds in different branches.

Figure 1: Example of a decision tree

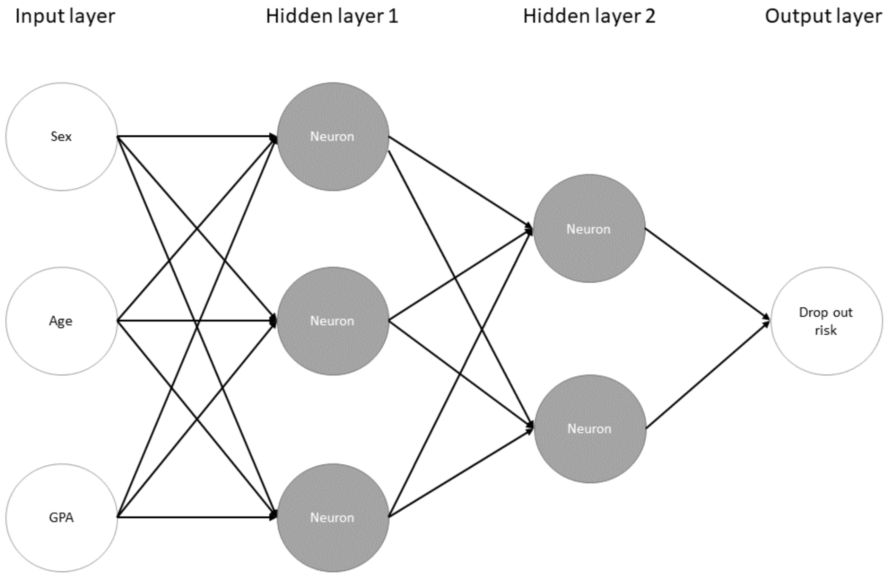


Tree classifiers can be prone to over-fitting, too. If the number of branches becomes too large, again they might represent particularities of a sample instead of underlying structures of dropout. To avoid such over-fitting of single trees, more complex methods such as the LightGBM[3] rely on a multitude of different trees, which are then called *forests* instead of on a single tree.

The intuitive decision trees maintain a large degree of interpretability. In forest methods, the number of nodes classified by a specific variable can be taken as a measure of these variables' importance for the decision making process.

Neural networks are meant to imitate human decision making processes. Intuitively, a neural network consists of several layers of *neurons*, which are best described as decision nodes consisting of highly interactive combinations of student characteristics. Neural networks differ, though, substantially from simple decision trees. Instead of branching out, each neuron of a layer feeds into each neuron of a subsequent layer (see Figure 2). The first layer consists of the observable inputs of the model. The arrows display the weights that are estimated in the learning process for each neuron. The last layer of neurons feeds into the prediction outcome, in our case a risk score for student dropout. The resulting complexity makes neural networks a powerful prediction tool, but it comes at the expense of a lack of interpretability about what variable is important in determining the outcomes.

Figure 2: Example of a neural network



Ensemble methods finally combine the strength of multiple prediction models to overcome individual weaknesses. In our case, we use LightGBM and provide it with both the predictions of the previous models and the underlying student level variables, so that the resulting model can weight the different predictions based on student characteristics.

[3] LightGBM is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms developed by Microsoft under free usage MIT license under https://github.com/Microsoft/LightGBM.

3.4 Data pre-processing

To implement the models laid out above, the raw data of the NEPS undergoes a necessary pre-processing. This `data pipeline' includes a number of necessary steps that differ between models.

For the OLS model first, missing values are replaced with variable means of all observations with available information. Categorical variables with nominal or ordinal scale are transformed into binary indicators for each category. Numerical variables on a cardinal scale are standardized to mean zero and standard deviation of 1. For the logit classifier and the neural network, we additionally apply a principal component analysis to reduce the number of variables used. The LightGBM, that turns out to be our preferred model in the later `horse race', does not rely on any pre-processing.

3.5 Metrics

To assess the performance of the prediction models in our *prediction horserace*, it is important to choose appropriate metrics that measure the distance between prediction and truth. Such metrics have to fulfil three criteria. First, appropriate metrics have to yield information on the objective of the project: the identification of student dropout. Second, appropriate metrics should be easy to interpret for a non-technical audience. Third, appropriate metrics have to be robust to the imbalanced distribution of graduates and students who drop out: approximately three out of four students successfully complete their degree. Metrics not taking into account this imbalance yield an over-optimistic view on prediction performance and might lead to false conclusions.

We illustrate the importance of choosing the appropriate metric with the matrix describing distributions of predicted vs actual dropouts in a so called *confusion matrix* in Table 2. The described situation involves 100 students out of which 20 students are *actual* dropouts while the remaining 80 students graduate. A fictive prediction model is able to correctly classify half of the dropouts. It performs better, though, for the graduates and correctly classifies three quarters of those. Thus, after running the prediction model, 10 of the actual dropouts are identified (*true positives*) while 10 are missed (*false negatives*). Instead, the model correctly classifies 60 out of 80 *actual* graduates (*true negatives)* but mistakes 20 actual graduates as dropouts (*false positives*).

Table 2: Exemplary confusion matrix

		Prediction	
		Dropout	*Graduate*
Truth	*Dropout*	10 *(true positives)*	10 *(false negatives)*
	Graduate	20 *(false positives)*	60 *(true negatives)*

Notes: This table displays an exemplary confusion matrix. The underlying population consists of 20 dropouts and 80 graduates. The examplary algorithm correctly identifies 10 dropouts (true positives), but misclassifies 10 dropouts as graduates (false negatives). Of the graduates, the algorithm correctly identifies 60 (true negatives), but mistakes 20 students as dropouts (false positives).

Accuracy. The most straightforward and often applied metric to assess the performance of the prediction model is its *accuracy*. The accuracy of a model is defined as the number of correctly classified students divided by the total number of students:

$$Accuracy = \frac{True\ negatives + True\ positives}{100}$$

In the example of Table 2, the accuracy would be $\frac{10+60}{100} = 0.7$. Thus, 70% of the students are correctly classified as either dropout or graduate. Although easy to interpret, the simple accuracy can be misleading in cases of imbalanced data, i.e. where the share of dropouts is relatively small. In the example of Table 2, even a naïve approach randomly assigning dropout status by its population share would yield in expectation a high accuracy of 0.68 (by correctly identifying 64% of all students as true negatives and 4% of all students as true positives).

Cohen's Kappa. In such cases of imbalanced data, an alternative and more appropriate metric is provided by *Cohen's Kappa* (Cohen, 1960). Cohen's Kappa corrects for the level of *expected accuracy* that would arise from pure chance. Thus, it describes the increase of accuracy compared to a naïve classifier which would randomly assign graduates or dropouts according to their known population shares. Cohen's Kappa is defined as

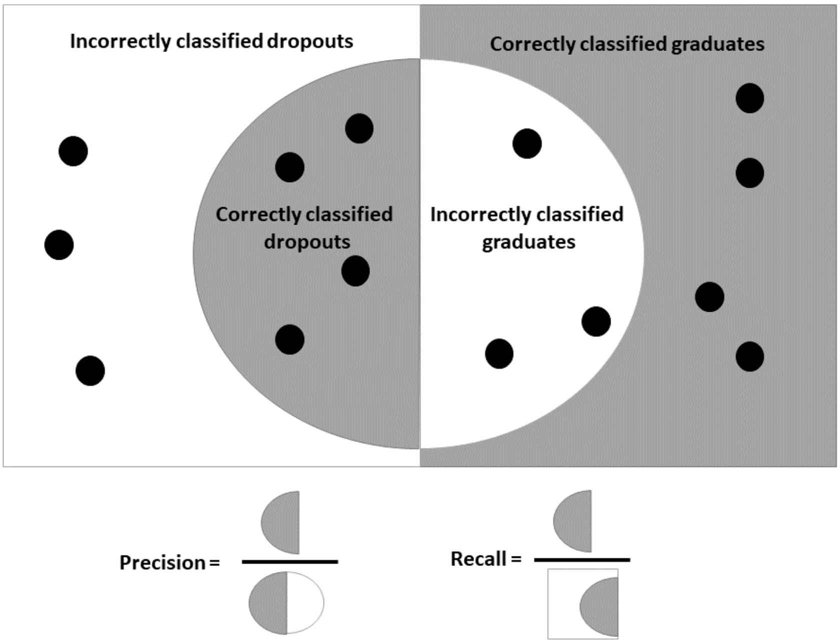
$$Cohen's\ \kappa\ = \frac{Accuracy\ \ - p_c}{1 - p_c}$$

, where p_c is the expected accuracy based on population shares. Accordingly, the setting described in Table 2 would lead to a Cohen's Kappa of $\frac{0.7-0.68}{1-0.68} =\ 0.0625$. The algorithm would therefore capture about six percent of the gap in prediction quality between the naïve approach and perfect knowledge.

Bakeman et al. (1997) propose the following intervals to classify the performance of prediction models by its Cohen's Kappa: 0-0.2 means a slight, 0.21-0.4 a fair, 0.41 – 0.6 moderate, 0.61-0.8 a substantial and 0.81-1 a perfect *agreement* (i.e. classifying each observation according to its true status).

Precision, Recall and classification thresholds. As it becomes clear from the example above, prediction models might differ, though, in their ability to correctly classify the different classes. In the example of Table 2, the prediction model can correctly classify only half of the dropouts, but three quarters of graduates. A disadvantage of Cohen's Kappa is that it is a global metric: it aggregates the performance of the classification model for both classes. If we are interested in separating these error rates, we can focus on *precision* and *recall* of a model (Perry et al. 1955). Figure 3 describes the two concepts. Here, the black dots display data points that are either correctly classified as dropouts (*true positives*, grey circled area), incorrectly classified as dropouts (*false positives*, squared white area), correctly classified as graduates (*true negatives,* grey squared area), or as incorrectly classified graduates (*false negatives*, white circled area).

Figure 3: Precision and recall


Incorrectly classified dropouts
Correctly classified graduates
Correctly classified dropouts
Incorrectly classified graduates
Precision =
Recall =

Precision is defined as the ratio of *correctly classified* dropouts over all students classified as dropouts (both true and false positives). It displays the probability of being a true positive conditional on positive classification and can be expressed as the ratio of true positives over the sum of true and false positives:

$$\boldsymbol{Precision} = \frac{true\ positives}{true\ positives + false\ positives}$$

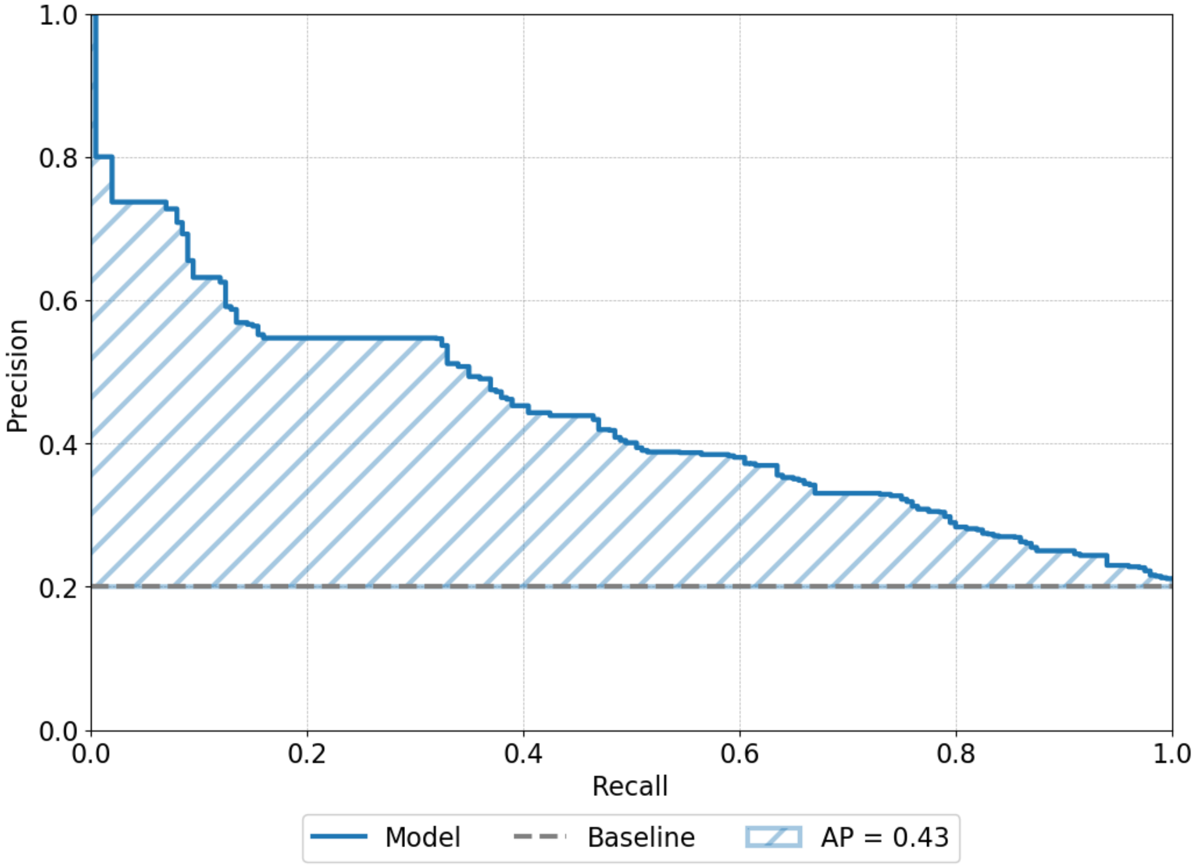
Recall (or *sensitivity*) instead displays the share of correctly classified dropouts among all *actual* dropouts (true positives and false negatives). It can be expressed as the ratio of true positives over the sum of true positives and false negatives:

$$\boldsymbol{Recall} = \frac{true\ positives}{true\ positives + false\ negatives}$$

In practice, prediction models do not provide binary outcomes (graduate/dropout) per se, but implicitly evaluate a predicted "risk score" against a classification threshold. This classification threshold can be set by the researcher to influence the different errors made by the prediction model. Precision and recall can be used to describe the performance of a prediction model in response to changes in the classification threshold. Therefore, by choosing the classification threshold, the researcher has to decide on the trade-off between precision and recall. This trade-off can be illustrated in a precision-recall curve.

Figure 4 displays an example of such a precision-recall graph. The horizontal dashed line indicates the ratio of dropouts in the population which also corresponds to the precision of a naïve prediction model which always predicts dropout. The solid line displays potential combinations of precision and recall for different classification thresholds which are not shown in the graph. Higher threshold levels are in principle related to higher precision but lower recall. The *area* below the precision/recall curve can be used as a measure of the overall performance of a prediction model, i.e. the average precision.

Figure 4: Example of a Precision-Recall Curve



We will later discuss consequences of the trade-off between precision and recall for cost/benefit considerations. Should the prediction model be used to assign high-cost interventions, practitioners want to avoid false positives (e.g. well-performing students being assigned to mentoring) and will opt for high precision. If costs of student dropout are substantial, practitioners would want to increase the number of identified dropouts at the cost of higher numbers of false positives and will opt for high values of recall.

4. Results

In the following sections, we describe the results of our *prediction horserace* of prediction models which differ by prediction method and set of available information. We further discuss the sensitivity of our results to changes in the definition of dropout.

To provide credible estimates of the performance of the prediction models, all results are based on out-of-sample predictions. To do so, we separated the main sample into a *training data set* for model selection as well as model fitting and a *test data set* that has been stored separately for the project period to be reserved for the final predictions. The ratio between test and training data is 3:1. In that sense, the results are based on the actual performance of a prediction model with respect to data "the model has not seen before". Thus, we emulate a practitioner's perspective whose objective it is to predict student dropout in newly arriving cohorts. Table 3 provides a comprehensive overview on all tested prediction models of our "horserace" and lists comparable core metrics for each of the models. Listed models differ by prediction model, applied definition of dropout and the used information sets.

Table 3: Metrics

Information set	**Model**	**Accuracy**	**Kappa**	**Precision**	**Recall**	**Average Precision**	**Num. Obs.**
By prediction model							
Data at enrolment	OLS	0.70	0.25	0.42	0.51	0.48	2449
	Logit	0.72	0.28	0.44	0.49	0.46	2449
	LightGBM	0.73	0.29	0.46	0.48	0.49	2449
	Neural Network	0.67	0.21	0.38	0.50	0.37	2449
	Ensemble	0.73	0.29	0.46	0.50	0.48	2449
By dropout definition							
Broad definition	LightGBM	0.77	0.28	0.37	0.50	0.39	2316
Intermediate definition		0.72	0.25	0.38	0.50	0.43	2316
Narrow definition		0.73	0.28	0.42	0.50	0.44	2316
By available information							
Base characteristics	OLS	0.65	0.18	0.36	0.52	0.42	2449
	LightGBM	0.67	0.21	0.38	0.51	0.44	2449
Bas. char. + Progress assessment	OLS	0.79	0.41	0.59	0.50	0.59	2449
	LightGBM	0.80	0.42	0.62	0.50	0.60	2449
Basic characteristics	OLS	0.66	0.18	0.35	0.50	0.41	2139
	LightGBM	0.67	0.20	0.36	0.51	0.43	2139
Bas. char. + Initial assessment	OLS	0.71	0.25	0.40	0.50	0.45	2139
	LightGBM	0.74	0.31	0.45	0.51	0.49	2139
Basic characteristics	OLS	0.63	0.07	0.14	0.51	0.17	1337
	LightGBM	0.61	0.06	0.13	0.51	0.15	1337
Bas. char. + Monitoring	OLS	0.70	0.13	0.17	0.51	0.23	1337
	LightGBM	0.75	0.17	0.20	0.50	0.25	1337

Notes: This table summarizes metrics of prediction performance for a range of models differing by applied method, range of information and applied definition of dropout.

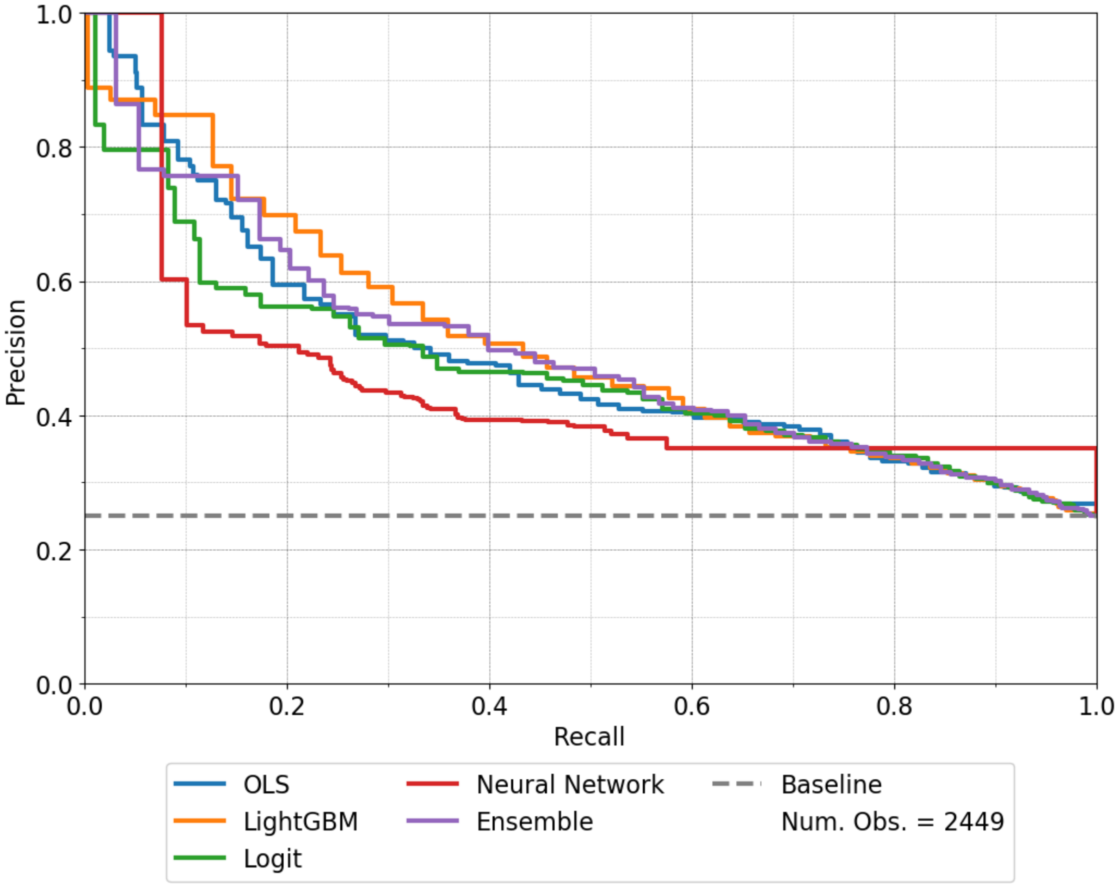
At first glance, models differ substantially in prediction quality. Most of this variation in prediction quality stems from the amount of information used. Instead, the actual choice of the prediction model does not seem to affect prediction quality substantially. In the following, we will systematically trace these sources of differences in prediction quality. To do so, we compare marginal increases in prediction quality associated with specific information sets and methods used to provide guidance about which information set is expected to identify the "biggest bang for a buck" in terms of prediction quality.

4.1 Classification performance by prediction model

We first examine the sensitivity of prediction quality with respect to the choice of the prediction model. We do so by using an information set including basic student characteristics and additional data that could potentially be acquired at the point of enrolment. We use this data to describe performance changes by different prediction models based on this early information. We rest this first exercise upon a sample of 2,449 students from the test data, these observations therefore represent about a quarter of the full sample.

Figure 5 compares precision-recall curves for each of the applied five estimation methods – simple OLS regression (OLS), tree-based LightGBM, Logit regression, Neural Network and the ensemble method. The area below a precision-recall curve describes the overall performance of a prediction model – the more area below the curve, the better is the expected performance in new data. The different points on the precision-recall curve display trade-offs between falsely classifying non-dropouts as dropouts (*precision)* and missing out actual dropouts (*recall*) that are associated to a specific decision threshold. We discuss the information content of precision-recall curves in more depth in Section 3.4.

Figure 5: Prediction performance by model



Notes: This figure compares precision-recall curves for different models based on information potentially available at the time of enrolment.

The comparison of the different precision-recall curves indicates a clear ranking of methods according to their prediction quality. On average and over all possible thresholds, LightGBM and the flexible ensemble method incorporating all tested methods jointly outperform their competitors. Especially for small and intermediate thresholds, i.e. for threshold values valuing precision over recall and therefore accepting larger numbers of missed dropouts, LightGBM is the model of choice, even outperforming the ensemble method.

The quantitative metrics associated to this comparison are summarized in the upper panel of Table 3. Applying LightGBM to the maximum amount of information leads to a prediction quality of $\kappa = 0.30$ (Cohens' Kappa, see discussion in Section 3.1). This means that the model gains about one third of the way between randomly assigning a dropout according to population shares (*random classifier)* and having *perfect knowledge* about who is a dropout. Under common comparison thresholds used in the machine learning literature, the model accordingly provides *substantially better* predictions than a naïve classification.

Looking at precision and recall separately and in more detail further reveals that models differ mainly in their precision, i.e. their capability not to classify non-dropouts as dropouts. We will later discuss the trade-off between precision and recall from a cost-benefit perspective in Section 5.1.

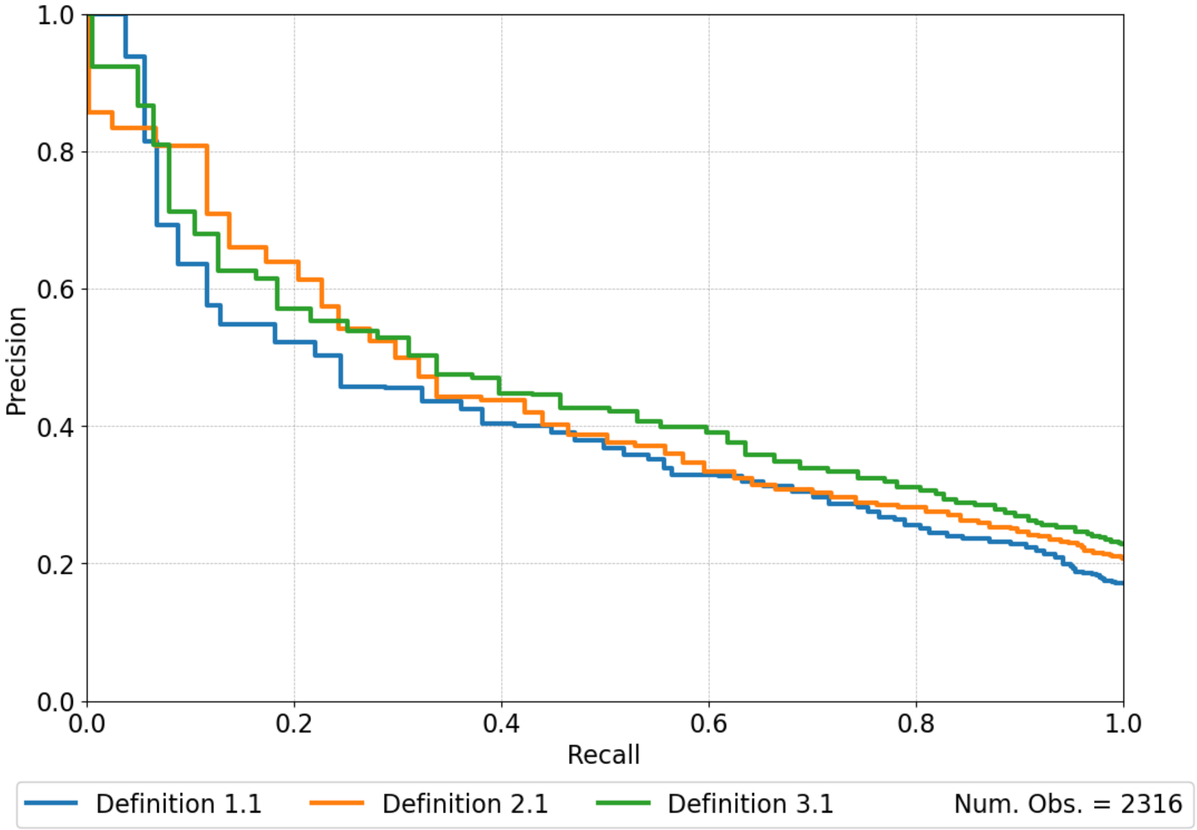
Taken together, we conclude from this comparison of models under full available information that LightGBM and ensemble methods are the models of choice. As LightGBM maintains a much higher level of interpretability and for example allows to better describe the importance of single predictors, we specify LightGBM as our model of choice from here on.

4.2 Classification performance by dropout definition

We next test the sensitivity of our results with regard to changes in the applied definition of dropout. Several different disruptions to a student's career can be defined as "dropout" – students leave an institution, but commence their studies at another institution, students decide to change to a different major, or decide to leave the tertiary education system entirely. Different decision makers have different stakes in different definitions of dropout. Government officials care about inefficiency in college placements and wasted resources through late college dropout. College administrations and department heads care about dropouts of their own student body which might affect resource allocation between colleges or departments. Changes in the definition of dropout do not only matter from a conceptual point, but might also affect the performance of prediction algorithms. We therefore test the robustness of prediction with respect to the three different definitions described in Section 3.1.

Figure 6 displays precision-recall curves of the same prediction model (LightGBM, variables as in Section 4.1) being applied with the three different definitions of dropout. The prediction performance is fairly stable across all three variants. For the remaining analysis, we therefore focus on the intermediate definition 2 which best captures a balanced view relevant for decision makers at various levels.

Figure 6: Prediction performance by definition



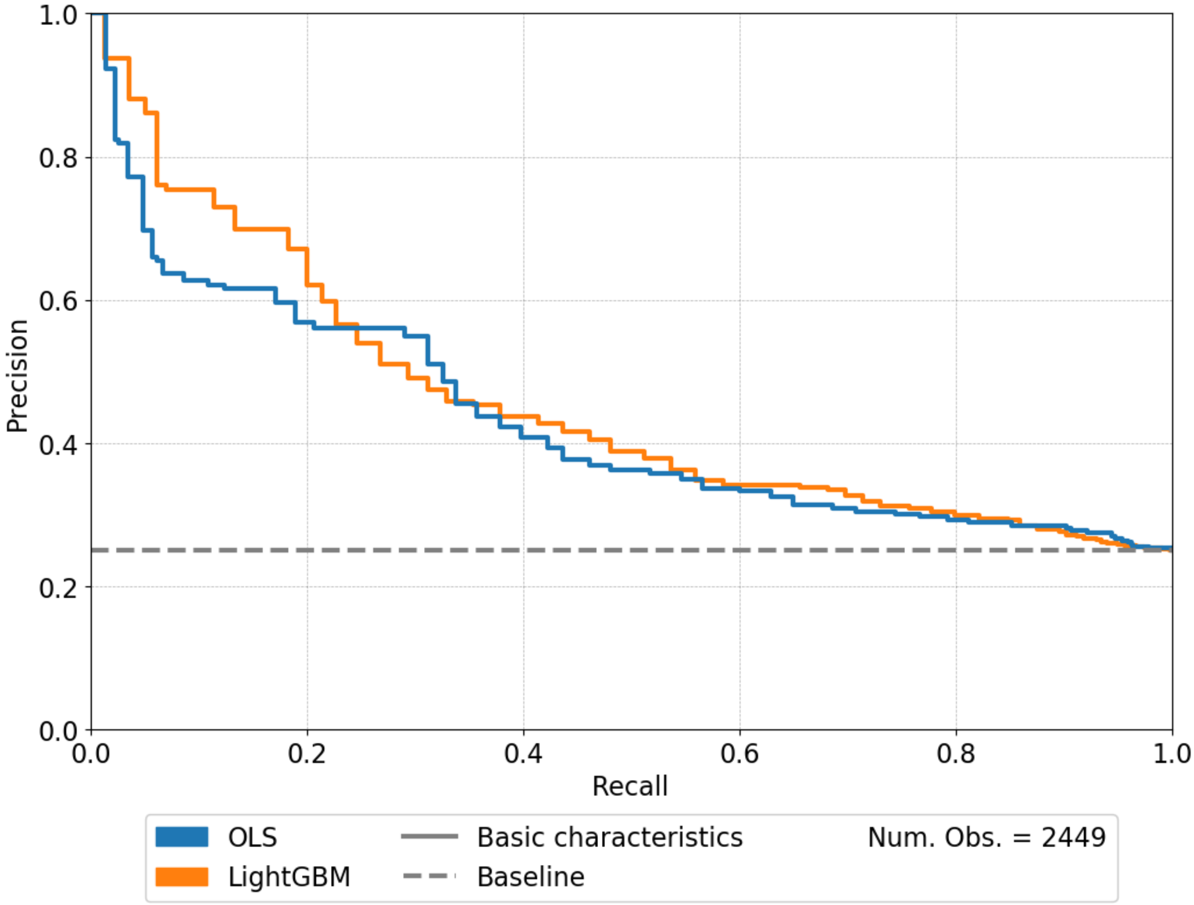
Notes: This figure compares precision-recall curves for different definitions of dropout, based on a LightGBM model with information potentially available at enrolment.

4.3 Prediction with basic characteristics only

Having established in the earlier section that in our setting LightGBM in general outperforms further prediction models for most decision thresholds with a fairly extensive set of predictors (although differences by choice of model are small), we now proceed to analyze how far results are sensitive with regard to changes in the amount of data used. We divide the available variables into the four subsets defined in Section 3.1: *basic student characteristics, initial student assessment, student progress assessment* and *student monitoring.* We test the different information sets with a basic OLS regression and the LightGBM model.

The most parsimonious set of information, *basic characteristics,* is restricted to variables that universities are legally obliged to record: basic socio-economic background information: gender, age, month of birth, nationality, chosen secondary subjects, intended degree, high school GPA and year of high school graduation. Using only these basic characteristics as input for the prediction model leads to a significantly worse prediction performance than providing the full information. Figure 7 compares the prediction performance of simple OLS and LightGBM using these basic characteristics. Again, LightGBM improves over simple OLS especially for low thresholds preferring precision over recall. Cohens' κ reaches $\kappa = 0.18$ for OLS and $\kappa = 0.21$ for LightGBM. Compared to the specifications using the full information, the accuracy decreases by about a third.

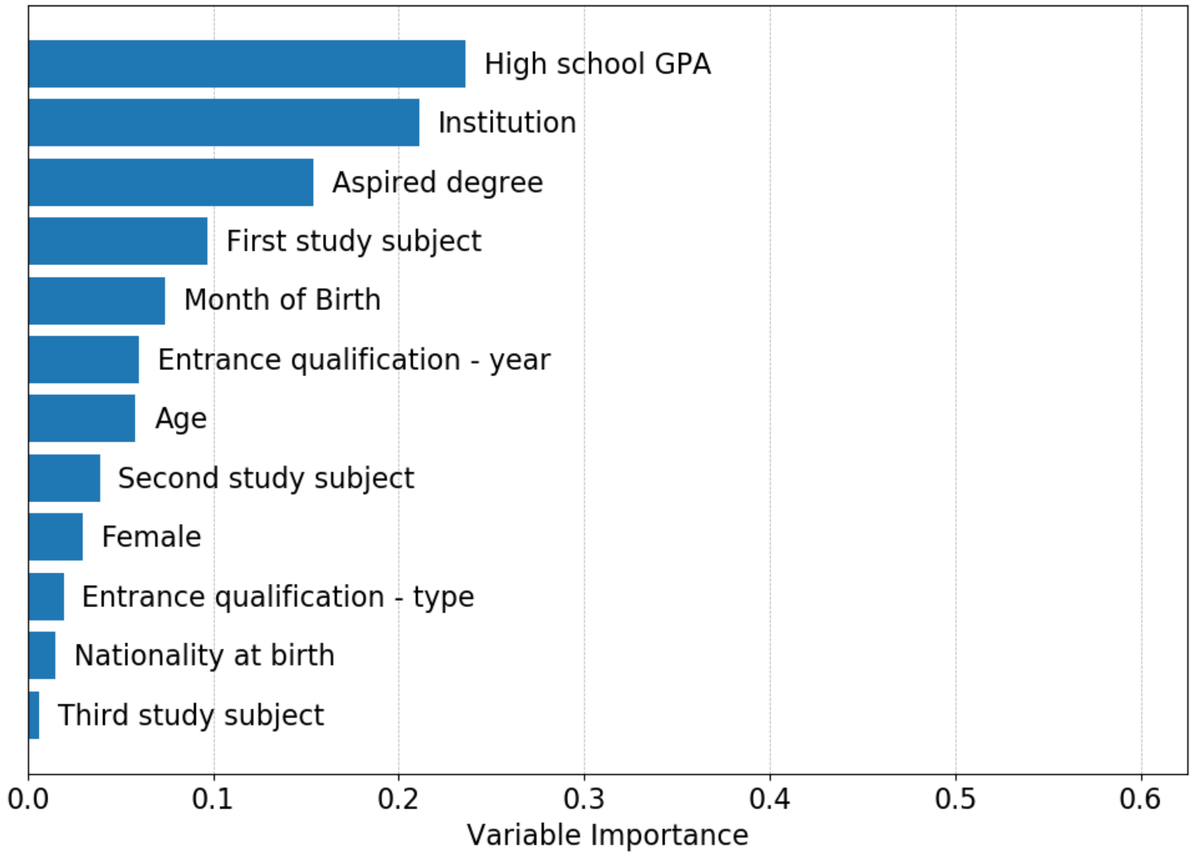
Figure 7: Prediction with basic characteristics only



Notes: This figure compares precision-recall curves for LightGBM and OLS based on basic student characteristics only.

Figure 8 displays the variable importance, i.e. the importance of a single predictor relative to the others, among the variables of the basic information set. The high school grade point average (*Abiturnote)* turns out to be by far the most important predictor. This is in line with the existing empirical literature describing high school performance as the most important predictor for student dropout (see for an overview Chingos 2018). This importance of high school performance as predictor for student dropout is unsurprising. In absence of better information, variation in high school GPA picks up both cognitive and non-cognitive skills, thus a range of individual traits that are expected to be highly relevant for study persistence (Borghans et al. 2016).

Figure 8: Feature importance: Basic characteristics



Notes: This figure displays the relative importance of the top predictors based on basic student characteristics only.

Although prediction quality is reduced significantly when restricting the available information to these basic characteristics, even with the small information set machine learning-based prediction can provide substantially better classifications than naïve or OLS-based prediction. As the basic information set is defined as characteristics which universities are already legally obliged to obtain, the usage of this information to design an early warning system is straightforward and cost-efficient.

4.4 Prediction with enrolment assessment

The amount of information that universities are legally obliged to collect and which was used in the previous section is restricted to the most basic demographic variables. For most institutions, this information could be easily augmented by additional information about a student's *study capital* (Tinto 1975), i.e. traits and characteristics, that a student brings with her and that shape her study persistence. Additionally, students are likely to differ in their beliefs and expectations they bring to university.

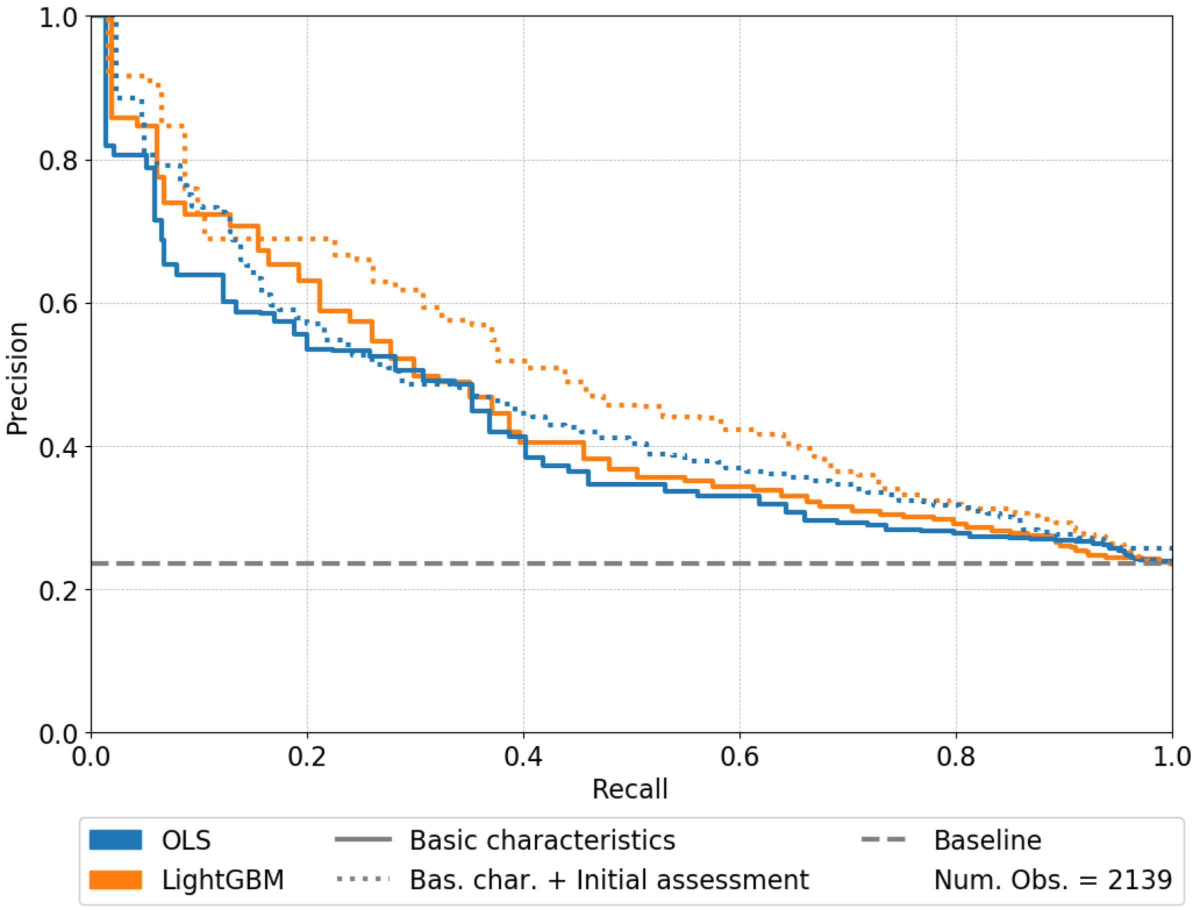
There exist several examples where universities induce students to partake in (mandatory) *enrolment assessments* at the beginning of their studies. We mimic such an enrolment assessment by drawing characteristics from the NEPS data that potentially could and that have been asked in student self-assessments in practice: the socio-economic background of a student, a student's personality[4], opinions

[4] Student personality was assessed by a Big 5 short form questionnaire. The earliest assessment of Big Five happens in the third wave of the survey. As commonly assumed in the psychological literature, we take the assessed personality traits as constant over the critical period.

and beliefs of the student, information on her friends and family and (for students for which it was available) test scores on cognitive ability.

Figure 9 compares precision-recall curves for a LightGBM and OLS model including this rich additional information to a variant using only basic information, restricting the sample to those students for which the relevant information is available. We observe that the prediction quality of a standard OLS regression does not substantially increase when adding the new information. Only when using the assessment information set in combination with the LightGBM model, prediction quality strongly increases for the entire range of classification thresholds. Adding the set of student assessment variables to the basic characteristics increases the predictive power of the model by a 50 percent higher prediction quality as measured by Cohens' Kappa. This substantial increase mostly stems from improved precision, i.e. the ability of the model to identify dropouts accurately.

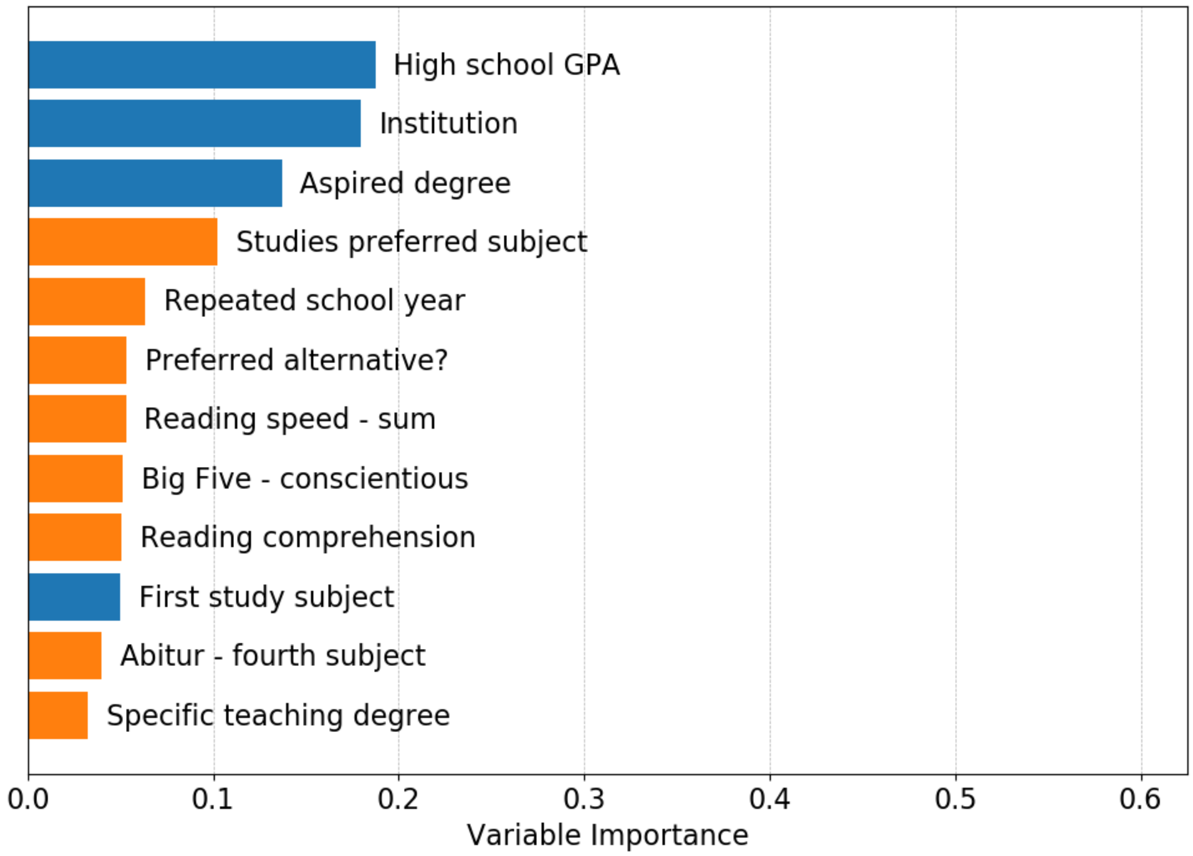
Figure 9: Prediction with student enrolment assessment variables



Notes: This figure displays precision-recall curves based on basic student characteristics and variables potentially to be acquired through student assessments at enrolment.

Figure 10 compares the variable importance of the assessment indicators. While the high school GPA and factors fixed to a single institution remain the most important explanatory factors, some new variables show up among the top predictors, including for example an indicator of whether a student studies her preferred subject. This information stands exemplary for a piece of information missed out by administrative data collections, but which is easy to assess and which has high predictive power. Also a student's personality, represented by the conscientiousness (a personality trait describing the desire to do tasks well and to take obligations seriously) becomes an important predictor, as well as different cognitive test scores.

Figure 10: Feature importance: Student enrolment assessment variables



Notes: This figure displays the relative importance of the top predictors based on basic student characteristics and variables potentially to be acquired through student assessments at enrolment.

The increase in prediction quality by including early assessment variables is the largest increase to prediction quality we observe in our "horserace". Thus, subjective information assessed at enrolment appears to be a very promising candidate to achieve a high prediction quality of later dropouts. The additional information can be assessed efficiently if included into already existing questionnaires that elicit basic information, as commonly used at most institutions. Nonetheless, some of this information, such as personality and beliefs, might be seen as sensitive. Asking for this kind of information might trigger resistance and indignation among the students. We discuss this kind of hidden cost in Section 5.2.

4.5 Prediction with student progress assessment

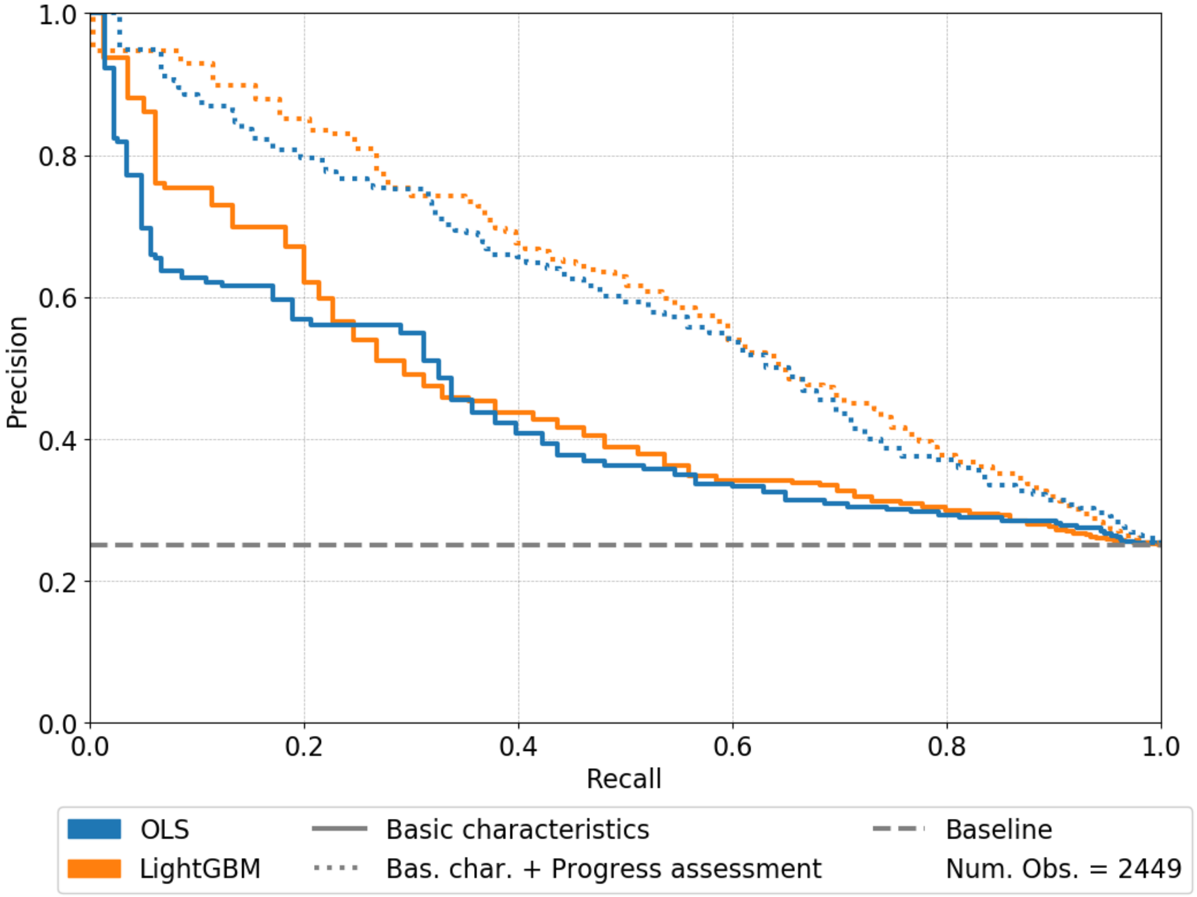
Many predictors of student dropout are likely to vary over time but are not directly observed in administrative records. The early years of tertiary education are a formative period for a student's beliefs and attitudes. Old social networks evolve and change, new networks at university emerge. Students deal with changing financial situations as well as personal shocks, e.g. emerging health issues, partner or family problems. Only relying on information assessed at enrolment might overlook these important predictors for student dropout.

We therefore now turn to determinants which could potentially be assessed through what we label *student progress assessment*. Student progress assessment can be implemented using a range of tools through which ongoing information, either objective or self-reported, could be assessed throughout a student's study progress. Students can fill in update questionnaires which ask about their well-being, integration, and ask directly about dropout intentions. Other institutions provide voluntary or mandatory mentoring or counseling where students are asked about their subjective progress, beliefs about success and intentions to drop out.

We mimic the information that could be assessed through counseling or surveying the student body by generating indicators for stated intention to drop out, enjoyment of one's studies, the subjective probability of graduation and a student's self-reported knowledge of the studied subject.

Adding these variables to a set of basic student controls substantially increases the prediction quality. Figure 11 compares precision-recall curves for LightGBM and OLS. For the LightGBM, prediction quality measured by Cohens' Kappa almost doubles. Given the comparably low number of additional included features, this substantial increase in prediction quality is specifically notable when compared to the marginal increase associated to the much larger set of assessment variables at enrolment described in the previous section.

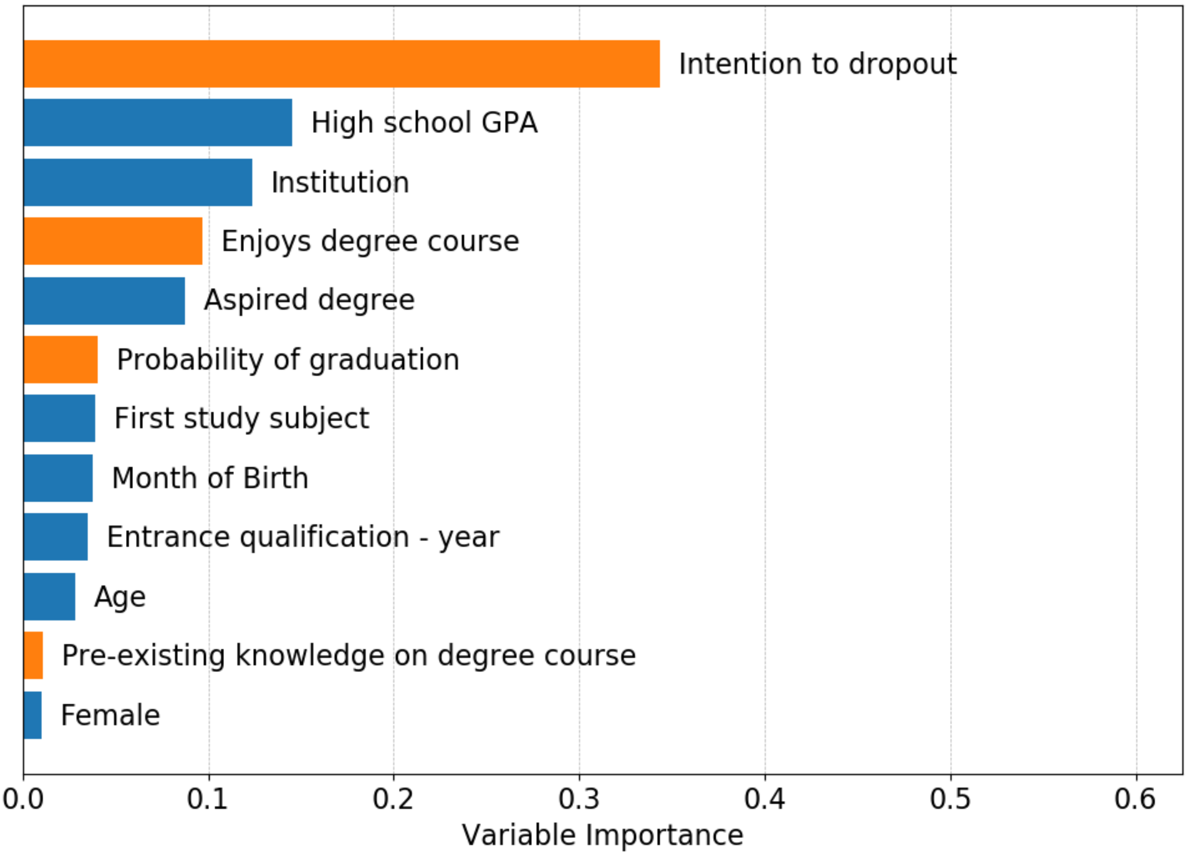
Figure 11: Prediction with student progress assessment



Notes: This figure displays precision-recall curves based on basic student characteristics and variables potentially to be acquired through student progress assessments.

Figure 12 summarizes the relative variable importance of the added subjective features. Especially the self-reported *intention* to dropout is unsurprisingly strongly predictive for actual realized dropout. While one might argue that this is rather a preceding measure of the outcome itself, in practice it would nonetheless be feasible to survey students about their intentions, and to intervene if the decision process is sufficiently long.

Figure 12: Feature importance: Student progress assessment



Notes: This figure displays the relative importance of the top predictors based on basic student characteristics and variables potentially to be acquired through student progress assessments.

The assessment of these variables is expected to be more difficult and cost-intensive than the previous information sets. Variables cannot be taken from existing administrative records. Making questionnaires mandatory interferes with a student's academic freedom and might reduce the quality of information. Keeping questionnaires voluntary strongly reduces response rates and leads to selective response. Further, study intentions and well-being are often perceived as sensitive private information. Student representations are likely to oppose the acquisition and use of this kind of information for predicting student outcomes.

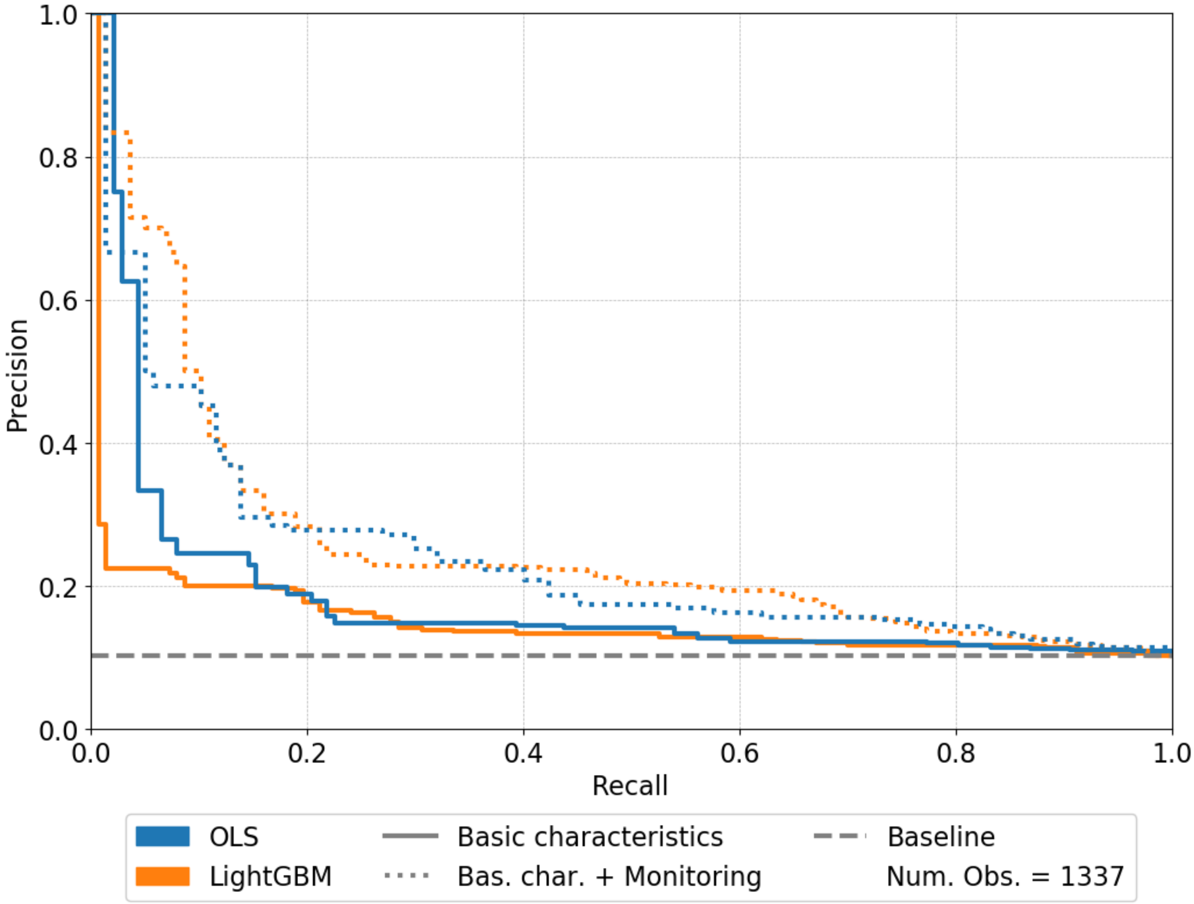
4.6 Prediction with student monitoring

In many cases, study dropout will simply be a response to slow progressions through one's studies. Comparing oneself with fellow peers, students might infer from inferior credit point accumulation that they will not reach their degree. Such belief updates have been analysed recently as a major determinant of student dropout (Zafar 2011, Arcidiacono et al. 2016). Including a student's ongoing performance into the prediction of dropout therefore appears to be the most straightforward way to increase the quality of prediction. Again legally obliged to monitor a student's progress in ECTS, universities can readily exploit this information.

Accordingly, the running GPA of a student, or the amount of achieved credit points, are of course readily available and potentially important predictors of student dropout. Unfortunately, the NEPS does only inadequately describe a student's progress in such objective performance measures, further the information is missing for large parts of the sample. Not all students report the GPA. The earliest information on the GPA is collected in the second wave of interviews from July 2011 in the students' mid second to mid third semester. A significant part of dropout has already happened at this time and the smaller sample with valid information on running GPA differs significantly from the larger sample in the previous sections.

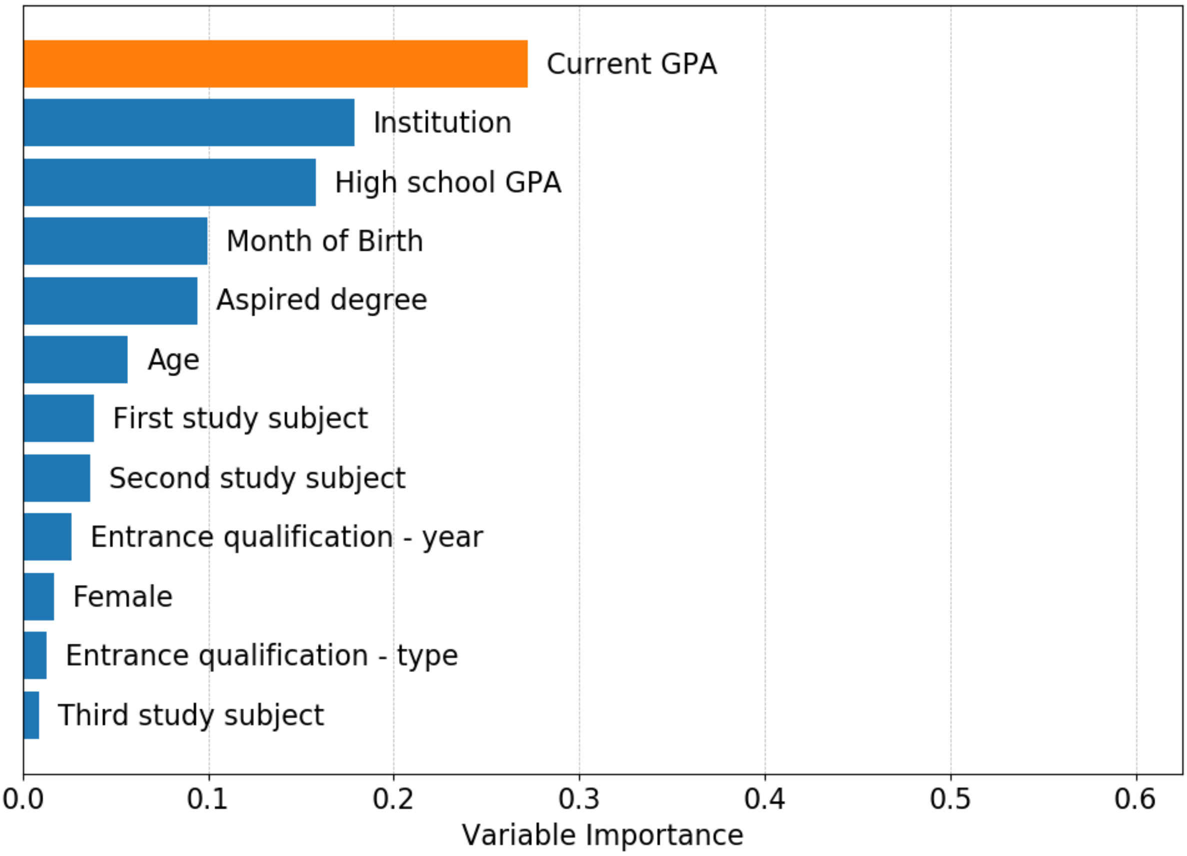
Despite this shortcoming, we report the results of a comparison of a model including running GPA to a model with basic characteristics in Figure 13. The overall performance of the model underperforms compared to any previously described model, due to the aforementioned data constraints. The relative variable importance of the running GPA, displayed in Figure 14, still is as expected very high. Study progress turns indeed out to be the most important predictor in this model. Therefore, as grades are readily assessed and recorded as inherent feature of a university's administrative processes, grades should be an integral part of any attempt to predict study dropout.

Figure 13: Prediction with student performance



Notes: This figure displays precision-recall curves based on basic student characteristics and variables potentially to be acquired through student progress assessments.

Figure 14: Feature importance: student performance



Notes: This figure displays the relative importance of the top predictors based on basic student characteristics and GPA as measure of student performance.

5. Discussion

5.1 Cost-benefit considerations

In the previous section, we compared different prediction models according to two core metrics: a model's *precision* and its *recall*. The possible combinations of these metrics corresponding to a specific prediction model are displayed by a precision-recall curve. Each point on this curve implies a single classification threshold associated to a certain precision-recall combination. This threshold is used to dichotomize the predicted risk score into a binary dropout/success classification. Which point on the curve, i.e. which threshold, to choose is up to the practitioner and has consequences for the expected number of false positives and false negatives of the classification. In the following, we discuss how cost-benefit analyses of dropout and interventions should guide this choice.

Precision and recall describe two different error dimensions of a model. Precision measures the ability of a model to identify true positives, i.e. actual dropouts, among all observations that are classified as dropouts, regardless whether this classification is actually correct or incorrect. Recall instead describes the ability of a model to identify as many true-positives, i.e. predicted dropouts, out of the population of all true dropouts.

To choose the optimal trade-off between the two error dimensions, it is important to understand the cost-benefit structure behind available intervention measures. This cost-benefit structure includes both the expected costs of a student dropping out of her studies and the direct costs and benefits of the intervention that is meant to be rolled out based on predicted dropout risk.

The costs of student dropout again consist of both private and social costs. From a student's perspective, costs include forgone earnings during the time spent at university, scarring effects from a later entry into the labor market, as well stigma and psychological costs of failure. From a societal perspective, costs are driven by withholding a college place from another student und the actual invested resources for training of the student subsequently dropping out. Further, lower earnings and higher unemployment risk of dropped out students lead to lower tax returns and higher welfare payments.

Similarly to the costs of student dropout, also intervention measures can differ strongly in their costs and efficiency. A simple information campaign raising the salience of dropout risks, associated costs and potentially informing about counseling services is relatively cheap to roll out, but might be of low efficiency. Providing a private 1:1 mentoring to at-risks students is expected to be of high efficiency in reducing dropouts, but constitutes a very costly intervention.

Taken together, the benefits of treating an at-risks student increase with the efficiency of an intervention and the costs that do not occur if a dropout is prevented. Costs of treating an at-risk student instead increase with the costs of the intervention. How is this cost structure of interventions linked to precision and recall?

Choosing a higher precision reduces the number of false positives, i.e. students falsely classified as dropouts who are indeed not at risk of dropping out. But this comes at the expense of a lower recall. The model will thus identify fewer actual dropouts in total. A practitioner is well advised to aim for a high precision/low recall of the prediction model in case that costs of interventions are relatively high compared to expected costs of dropout. In this case, falsely identifying and treating students who are not at risk of dropout is costly and the according risk should be minimized.

If however the costs of an intervention are low or do not play a role (e.g. the department is well funded), a practitioner might rather be advised to aim for a high recall. In this case, the prediction model might wrongfully classify a higher number of graduates as dropouts, but at the same time most dropouts are affected by the intervention.

This ensures that all potential dropouts receive the treatment. It should be clear from the above cost-benefit considerations that simply choosing the best performing model is not necessarily the best choice.

Practitioners have to take into account the comprehensive cost-benefit structure of both student dropout and intervention measures when choosing the appropriate prediction model. The cost structure and manifold counterfactual scenarios of dropout render the estimation of the costs of dropout difficult. Considering all personal or societal costs of dropout in the decision to implement an intervention, the resulting higher costs tilt the precision-recall trade-off in favour of higher recall.

5.2 Hidden costs through privacy concerns

In the previous sections, we have demonstrated that machine-learning-based prediction of dropout is feasible and can lead to significant increases in risk assessment compared to naïve or regression-based classification. Gains in prediction quality increase the more information is provided to the prediction model. Some of this information is already available to universities, other information sets can be acquired at different costs.

Some features that stand out in their power to predict student dropout might be perceived as sensitive by parts of the student body. Eliciting this information by surveys and using it to predict student careers might lead to strong resistance among students and administration staff on the basis of privacy and data security concerns. Germany has a long-running history of public distrust against centralized data collection, with the massive protests against and ultimately the cancellation of 1987's Census as a prominent example. In university settings, such upheavals have been common in the past whenever universities decided to implement or reform centralized student data collection and storage.

Practitioners have to weigh these potentially arising costs against the benefits from having access to additional information. Early release of information towards student body and administration about the scope and reason behind additional data collection might mitigate such arising hidden costs.

6. Conclusion

In this paper, we assessed the feasibility and discussed concerns and cost-efficiency of a machine-learning-based prediction model of university dropout. We tested several prediction models based on data from the National Education Panel Study.

We used several sets of information – basic student information, subjective information potentially to be assessed at enrolment and information that could be acquired through ongoing student progress assessment, to let different prediction models compete in a "prediction horserace".

Our results indicate that prediction quality is rather insensitive to the choice of the prediction method. Models turn out to be very sensitive in their precision – the ability to correctly classify actual dropouts while keeping the number of non-dropouts falsely classified as dropouts (false negatives).. Subjective information, such as a student's personality and cognitive skills that could potentially be assessed through mandatory surveys at enrolment substantially increase prediction quality. More so, continuously surveying students about their subjective satisfaction with their studies could generate information that closely predicts later dropout.

We discussed implications of different performance metrics (precision and recall) for cost-efficiency considerations of the implementation of early dropout detection systems. Practitioners are advised not only to choose the best performing prediction model, but to be guided in their choice by the cost-efficiency structure of dropout and intervening measures. Finally, we discussed the role of potential hidden costs that might arise from students' indignation triggered by being asked for seemingly sensitive information.

References

Adamopoulou, E., Tanzi, G. M. (2017) "Academic Dropout and the Great Recession," *Journal of Human Capital* 11, no. 1: 35-71.

Arcidiacono P., Aucejo E., Maurel A., Ransom, T. (2016) "College Attrition and the Dynamics of Information Revelation," *NBER Working Papers* 22325

Bakeman, R., McArthur, D., Quera, V., Robinson, B. F. (1997) “Detecting sequential patterns and determining their reliability with fallible observers” *Psychological Methods*, American Psychological Association, 2, 357

Berens, J., Schneider, K., Goertz, S., Oster, S., and Burghoff, J. (2018). Early Detection of Students at Risk-Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods. *CESifo Working Paper*, No. 7256.

Borghans, L., Golsteyn, Bart H. H., Heckman, J. J., Humphries, J. E. (2016), “What Do Grades and Achievement Tests Measure”, *Proceedings of the National Academy of Science*, Volume 113, Issue 47, November 2016, Pages 13354-13359.

Chingos, M. M. (2018). What matters most for college completion? Academic preparation is a key predictor of success. *AEI Paper & Studies*, 3A.

Himmler, O., Jäckle, R., & Weinschenk, P. (2019). Soft Commitments, Reminders, and Academic Performance. American Economic Journal: Applied *Economics*, 11(2), 114–142

Horstschräer, J., Sprietsma, M. (2015), The Effects of the Introduction of the Bachelor Degree on College Enrollment and Dropout Rates, *Education Economics* 23(3).

Perry, J. W., Kent, A., Berry, M. M. (1955), “Machine literature searching x. machine language; factors underlying its design and development”, *American documentation, 6*, 242-254

Sansone, D. (2018). Beyond Early Warning Indicators: High School Dropout and Machine Learning. *Oxford Bulletin of Economics and Statistics*.

Sara, N. B., Halland, R., Igel, C., Alstrup, S. (2015). High-school dropout prediction using machine learning: A danish large-scale study. In *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence* (pp. 319-24).

Stinebrickner, Todd R. Stinebrickner, R. (2014) “A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout,” *The Review of Economic Studies*, 81(1), 426-472

Stinebrickner, Todd R. Stinebrickner, R. (2014) “Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model,”*Journal of Labor Economics*, 32(3)

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of educational research 45(1), 89-125.

Zafar, B. (2011) “How Do College Students Form Expectations?”, *Journal of Labor Economics*, 29, issue 2, p. 301 - 348,