



# Temporal and Between-Group Variability in College Dropout Prediction

Dominik Glandorf  
dominik.glandorf@student.uni-tuebingen.de  
University of Tübingen  
Germany

Hye Rin Lee  
hyerin@udel.edu  
University of Delaware  
USA

Gabe Avakian Orona  
gabriel.orona@uni-tuebingen.de  
University of Tübingen  
Germany

Marina Pumptow  
marina.pumptow@med.uni-tuebingen.de  
University Hospital Tübingen  
Germany

Renzhe Yu\*  
renzheyu@tc.columbia.edu  
Teachers College, Columbia  
University  
USA

Christian Fischer\*  
christian.fischer@uni-tuebingen.de  
University of Tübingen  
Germany

## ABSTRACT

Large-scale administrative data is a common input in early warning systems for college dropout in higher education. Still, the terminology and methodology vary significantly across existing studies, and the implications of different modeling decisions are not fully understood. This study provides a systematic evaluation of contributing factors and predictive performance of machine learning models over time and across different student groups. Drawing on twelve years of administrative data at a large public university in the US, we find that dropout prediction at the end of the second year has a 20% higher AUC than at the time of enrollment in a Random Forest model. Also, most predictive factors at the time of enrollment, including demographics and high school performance, are quickly superseded in predictive importance by college performance and in later stages by enrollment behavior. Regarding variability across student groups, college GPA has more predictive value for students from traditionally disadvantaged backgrounds than their peers. These results can help researchers and administrators understand the comparative value of different data sources when building early warning systems and optimizing decisions under specific policy goals.

## CCS CONCEPTS

• **Applied computing** → Education; • **Information systems** → Decision support systems.

## KEYWORDS

college dropout prediction, administrative data, machine learning, temporal dynamics, student heterogeneity

\*Equal contribution



This work is licensed under a Creative Commons Attribution International 4.0 License.

LAK '24, March 18–22, 2024, Kyoto, Japan  
© 2024 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-1618-8/24/03  
<https://doi.org/10.1145/3636555.3636906>

## ACM Reference Format:

Dominik Glandorf, Hye Rin Lee, Gabe Avakian Orona, Marina Pumptow, Renzhe Yu, and Christian Fischer. 2024. Temporal and Between-Group Variability in College Dropout Prediction. In *The 14th Learning Analytics and Knowledge Conference (LAK '24)*, March 18–22, 2024, Kyoto, Japan. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3636555.3636906>

## 1 INTRODUCTION

Preventing college dropout is a long-lasting goal of modern post-secondary education institutions [22]. Succeeding Tinto's groundbreaking theory of academic integration [55], large-scale administrative data and machine learning (ML) algorithms have been leveraged to build early warning systems (EWS) for student attrition in the recent decade [24]. However, the definition and nature of college dropout can vary significantly across institutional contexts, student populations, and application scenarios. Therefore, the full potential of early warning algorithms has yet to be systematically evaluated. Modeling decisions regarding the time of prediction and potentially different dropout mechanisms across student subgroups have to be better understood to build robust and reliable prediction systems.

This study aims to bridge the understanding of the complex dynamics of student dropout factors and real-world applications of college dropout models. Our analyses focus on the relevance of individual predictors and potential group differences, such as gender or other traditionally underrepresented groups, for dropout risks. We integrate the temporal dimension of the dropout prediction problem by comparing different time points at which dropout can be predicted. We target both further hypothesis-driven research and the construction of early warning systems (EWS) by answering the following research questions (RQs):

- RQ1: How well do college dropout prediction models perform when utilizing only administrative data, and which predictors within these models are the most important?
- RQ2: How do the predictability and relevance of predictors of college dropout temporally change throughout enrollment?
- RQ3: How do the predictability and relevance of predictors of college dropout vary between different student populations (i.e., underrepresented minorities, low-income-family, female, STEM, first-generation college students)?

## 2 RELATED WORK

### 2.1 Definition of College Dropout

The opposite of academic success, i.e., no graduation or the absence of course grades in a given period, generally defines dropout [7, 30]. Academic success has been measured by time to degree, time of absence, or graduation [9, 13]. Therefore, the following dimensions appear most helpful in defining dropout: retention, non-completion of the degree program, and elapsed time. Retention refers to whether a student reports back at the beginning of the term [7]. Non-completion refers to students not returning and at the same time not having completed the program [11]. Elapsed time refers to the number of consecutive terms students were not enrolled in coursework [52]. Using these principles, we consider “a student that has not taken any courses at the university for at least four consecutive terms and has not completed the degree program” as dropout.

### 2.2 Predictors of College Dropout

**2.2.1 Administrative Data.** This data source refers to the information collected by the institution at the onset of and throughout a student’s college trajectory [24]. It includes variables such as demographics (e.g., ethnicity, income status, first-generation status), academic performance measures (e.g., high school grade point average, scholastic aptitude test scores), and course-level outcome data (e.g., final grades). High school grade point average (GPA) and entrance test scores are mainstay predictors in most educational studies investigating college dropout [7–9, 16, 30, 39, 43, 46]. Similarly, demographic characteristics such as age, ethnicity, gender, and socioeconomic status predict student dropout [7, 9, 16, 28, 30, 39, 40, 43]. However, some studies indicate significant contributions to dropout predictions [9, 30, 39, 41, 48], whereas other studies fail to replicate such findings [31, 43, 59]. More work is needed to understand how various demographic characteristics predict students’ academic success.

**2.2.2 Alternative predictors.** Surveys add various motivational belief constructs, such as academic self-efficacy, values, and motivation, to the traditional predictors [18, 19, 28, 30, 38]. However, the degree to which these attributes improve dropout predictions is debated, with modest improvements reported in prior work [22]. The cost of this type of data and low response rates to non-obligatory student surveys (e.g., 9% in [50]) led us to omit survey data in our prediction models. A growing set of predictors is derived from learning management systems to predict students’ engagement [10, 32]. Clickstream data refers to time-stamped records of student interactions triggered by the use of digital course material [47, 49]. Clickstream-based measures, such as idle time, number of keystrokes, and frequency of clicks within a particular page of the online learning environment indicate students’ engagement [23]. Unfortunately, the heterogeneity of these environments and course-dependent use cases make this type of data hardly available for administrators as an off-the-shelf predictor in ready-to-scale prediction models.

### 2.3 Prior early warning systems

In recent years, the described data sources have been used for EWSs to improve educational practice. Early examples of such uses include Arnold and Pistilli’s study [5], which leveraged demographic, learning-management-system, and previous academic records to identify students at risk of not being retained in courses. Similarly, Brown et al. [12] utilized standardized test scores, course information, and demographics to implement early warnings for the performance of students enrolled in general education programs. More recent studies have branched towards additional data sources. A model that added formative assessments and online activity to predict final grade outcomes supplied approximately a 94% accuracy rate by week 6 [33]. More recently, used E-book log data and Wifi connections served to predict the risk of course failure [1, 60]. Our approach has the potential to achieve similar predictiveness by incorporating only a cheap subset of these rich data sources. It can create a non-perfect but easy-to-implement first-level detection system that can indicate which at-risk students need more careful examination.

### 2.4 Temporal dimension and subpopulations

Dropout is often predicted only at one single point, such as the time of initial enrollment [7], end of the first term [15, 44], or end of the first year [7, 8]. Different prediction time points were typically only spread across one term to make course-related predictions [1, 46, 60]. Only a few studies predict at times ranging from initial enrollment until the end of the second year [29], and a single one reported the change of predictor importance for one of their models [9]. Apart from that, changes in college dropout factors have been only tracked over cohorts [54]. This change of predictor importance between two or more time points within a student’s trajectory was only systematically tracked for high school dropout [25] or within a survival analysis of college dropout focused on the time point of dropout [36]. Therefore, we emphasize a systematic comparison between time points in this study concerning the prediction quality and data sources.

Group-specific dropout factors were previously only modeled in classical statistical models (regressions and structural equation modeling) [4, 53]. To the best of our knowledge, no prior work using ML methods has focused on group-specific college dropout predictors. We could identify almost no studies that reported the change in importance of these predictors on the between-group dimension. Only one study reported how factors may vary between public and private institutions [9]. We see this as a chance to combine the strength of ML methods with traditional interaction analyses (see Section 3.2).

## 3 METHOD

### 3.1 Study Setting and Sample

This study is conducted at the University of California, Irvine, a large public research university in Southern California, that enrolls more than 25,000 undergraduate students. This university features a diverse undergraduate student body and received federal designations as a Hispanic-Serving Institution (HSI) and an Asian American and Native American Pacific Islander-Serving Institution

(AANAPISI). Data for this study were provided from a multitude of services that collect and curate institutional data, including Admissions, the Registrar's Office, the Office of Institutional Research, the Office of Information Technology, and the Office of Financial Aid and Scholarships. Notably, this study is tied to a large institution-wide measurement project to understand the value of undergraduate educational experiences and promote evidence-based models of undergraduate student success trajectories [6]. The investigation uses data from six cohorts of degree-seeking non-transfer students (2011-2016 entrance dates) to capture both four- and six-year graduation rates. This led to a total sample size of 33,133 students, with records of 367,761 terms and 1,466,260 course enrollments, spanning twelve years of data.

### 3.2 Research Design

To answer RQ1 and identify the best model, we evaluate all models' ability to predict dropout after one and two years of the initial enrollment. Based on typical dropout, we select two observation spans to account for dynamic relations between predictors and dropouts over time. RQ2 and RQ3 are computed based on the best-performing model from RQ1.

RQ2 analyzes the temporal dynamics of the dropout process. Ten subsets of the data span periods of up to the first three years of study (three terms per year). Starting from information available at the moment of the first enrollment, data obtained later than  $n$  terms after enrollment is discarded for  $n \in \{0, 1, \dots, 9\}$ . Students already known to have dropped out are excluded from later subsets. For each observation span, the relative importance of predictors is identified to track their changes over time. For time points later than this, models were not analyzed anymore due to the very low base rate of dropout by then.

RQ3 is structurally analogous to RQ2 but aims to identify differences between subgroups of students. We choose the attributes *female*, *first-generation student*, *low-income family*, *underrepresented minority*, and *STEM major* (students in science, technology, engineering, and mathematics degrees) for the comparison of predictability and predictor relevance as they are often of particular interest for college administrators. The main analysis is conducted for a respective observation span of three terms, but a robustness analysis has been carried out for observation spans of two and four terms (see supplementary material<sup>1</sup>).

### 3.3 Data

**3.3.1 Outcome.** Dropouts are identified according to our dropout definition (see Section 2.1) and marked as such if students show at least four consecutive terms of no re-enrollment and were never reported to graduate. This leads to a prospective dropout rate of 13.2% after the first year and 11.4% after the second year. The descriptive statistics of all predictors and the conditional dropout rates can be found in the supplementary material.

**3.3.2 Pre-entry predictors.** Demographic data is usually available at the time of admission and remains invariant. Gender is simplified to a binary variable of whether a student is *female*. The *age*

*at enrollment* is derived from the date of birth. *International students* are annotated with the additional information if they took the TOEFL. The *ethnicity* is captured as "Asian / Asian American", "Black", "Hispanic", "Indigenous", or "White non-Hispanic". For RQ3, "Black", "Hispanic", and "Indigenous" are summarized to the binary label *underrepresented minority*, following a standard definition [21]. The *citizenship* is indicated as "US Citizen", "Permanent Resident", and "Not US Citizen". On the geographic scale, the *residency within the state* at the time of application is known as "In-State", "Bona Fide", and "Out-of-State". The *geographical category* contains more specific categorical information about residency before enrollment: "Foreign Country", "Out-of-State", "Northern California", "Southern California", and "University County". The university's *distance from home* is enquired at the same point. Students within the *first generation* of their family to study and students from a *low income* family are flagged. The *parents' education* is indicated per parent with the categories "No high school", "Some high school", "High school graduate", "Some college", "2 year college grad", "4 year college grad", and "Postgraduate study". The *household size* at the time of admission is registered as the number of members, capped at a maximum of six. A binary variable indicates whether a student is a *single parent*. Lastly, it is stated if a student is an *English language learner* (i.e., non-native speaker).

Performance data collected prior to the studies contain the *high school GPA* as well as the *math*, *writing*, and *reading entry test score*. The *best score in Advanced Placement (AP) exams* is used when available; otherwise, it is set to 0. The resulting year of study in the first term ("Freshman", "Sophomore", "Junior/Senior") is also recorded.

**3.3.3 Post-entry predictors.** Stemming from term-level information, the data contains the current *number of declared majors* with the corresponding *number of school affiliations*. The *primarily affiliated school* and *major* were indicated as a categorical variable. The number of *changes of major*, *school*, and *total enrolled terms* are respectively derived from multiple term records. Note that our definition of the outcome only considers students as dropouts as soon as they have not re-enrolled for four consecutive terms. A binary flag indicates whether a student was declared as *honors* for at least three terms and if at least one of their majors is a *STEM major*. Predictors also include the average *number of courses* taken per term and the *current year of study*.

On the course level, both demographic and performance data are captured. The *number of credits*, if the course has been *passed*, and the numeric *final grade* indicate the performance. The *number of total students* and the relative amount of students of the *same gender*, *first-generation status*, and *ethnicity* are captured as demographic indicators. All numeric information is first aggregated per term and then incorporated into cumulative averages up to the time point of prediction. The *linear change in the number of credits* from the first to the current term is also calculated. Another statistic is the *number of credits relative to the major's average*, and whether taken courses were *offered by a school of one of the majors*.

### 3.4 Modeling

A range of binary classification models is used in dropout prediction tasks [1, 7, 9, 11, 17, 28, 43, 51, 58]. We trained them on all

<sup>1</sup><https://github.com/dominikglendorf/college-dropout-prediction>

predictors except the declared major as it shows redundancy with the school variable and would introduce too many variables. The *logistic regression*, which assumes a linear relation between predictors and logarithmic odds of the outcome, is widely used. Besides being relatively simple to apply, it usually provides accurate predictions [23]. The used ML methods include *random forests* (RF), which are based on decision trees. Each tree recursively splits the data into two subsets based on the feature out of a random feature selection of a specified size that yields the best class impurity until the tree is grown to a specified size. An RF is assembled of a specified number of decision trees trained on different subsets of the training set. The predictions are averaged to make a robust prediction that maintains the quality of the individual trees. We use the R implementation in the package *randomForest* [42].

The *support vector machine* (SVM) chooses the position of a hyperplane in a multidimensional space, optimizing the class separation and the margin to their data points, incurring a certain cost on violations. Non-linear kernels that transform the input data into high-dimensional spaces often perform best. Performance also depends on regularization for the decision boundary (cost) and class weights. In the case of radial basis functions as a kernel, one must additionally specify the radius of influence (gamma). We use the implementation in the R package *e1071* [45]. The *naive Bayes* classifier assumes an independent effect of categorical predictors on the outcome. Hence, it predicts its joint probability based on the observed class frequencies. Therefore, continuous predictors are discretized. The classifier can integrate a regularization value to generalize better to joint probabilities unobserved in the training data (Laplace parameter). We use the implementation in the R package *e1071* [45].

*k-nearest neighbors* identifies the  $k$  closest instances in terms of the Euclidean distance and predicts class membership based on the majority class within the neighborhood. Categorical predictors have to be dummy-coded for this purpose. We use the implementation in the R package *class* [57]. Feed-forward *neural networks* model non-linear functions by hierarchically applying linear transformations and non-linear activation functions to the input predictors. The two output neurons representing the two classes are normalized to probabilities using the softmax function. The error function is the binary cross-entropy loss, which is used to train the model weights and biases via backpropagation. We use the R interfaces to Keras [2] and TensorFlow [3] using Python 3.10.

**3.4.1 Missing data imputation and hyperparameter tuning.** Many predictive models require complete data. Due to the amount of missing data in some predictors, we prefer data imputation over keeping only complete data points. Creating multiple imputations to reflect the uncertainty in the missing data prediction and calculating results for all of them is a common approach. The R package *mice* [56] starts from simple baseline imputations and recursively repeats more sophisticated model-based predictions to improve them. We choose RF as the underlying single imputation method because of its suitability for categorical and continuous predictors. We generate ten imputed datasets, on which we run the entire training routine to ensure the robustness of our results against the randomness of the imputation. To ensure that classifiers perform best, their non-trainable parameters are optimized heuristically. This study

uses grid search over all combinations of a careful selection of hyperparameter values (see Table 1). Our performance evaluation is based on the hyperparameter-tuned models and averaged metrics across imputations.

## 3.5 Evaluation

**3.5.1 Performance.** The performance of dropout prediction is estimated via 3-fold cross-validation, with a held-out test dataset to evaluate future performance. In dropout prediction, the imbalanced classes are often not addressed [51], e.g., when *accuracy* (ratio of correct predictions) is the only measure of performance [8, 46]. Overly predicting no dropout inflates accuracy without addressing the problem. Reporting per-class accuracy fixes the problem but hampers comparing datasets with different base rates. However, we include accuracy in the results to show how unsuitable the metric is in practice. It is calculated for 200 thresholds, ranging from 0 to 1, to choose the threshold that yields the maximum possible score optimally.

The more comprehensive receiver-operator characteristic (ROC) summarizes all possible thresholds for the *sensitivity* (ratio of actual dropouts detected, also called *recall*) and *specificity* ( $1 - \text{false positive rate}$ ) and can be summarized in a scalar, called area under ROC curve (AUROC), which is often used in dropout prediction. The precision-recall curve (PRC) is more informative for imbalanced classes because it incorporates the *precision* (correct ratio of predicted dropouts) instead of the specificity. It also can be summarized in the area under PRC (AUPRC) [26]. For all metrics, the best possible classifier with no predictors implies the baseline. Given the base rate for dropout is  $r_d$ , the baseline accuracy in binary classification is  $1 - r_d$ . In the case of AUROC, it is 0.5; for AUPRC it is  $r_d$ .

**3.5.2 Predictor Importance.** To ensure the meaningfulness of our results to administrators, the global scores of predictor importance are of most interest. We choose a model-agnostic approach to calculate the significance of single predictors, such that scores are available independent of the model performance ranking and can theoretically be compared across models. The *Permutation Feature Importance* (PFI) measures how much the test set performance of a model decreases when one variable is randomly permuted [8]. For RQ2, the PFI is based on the more sensitive AUPRC. In RQ3, the different base rates led us to use the AUROC with its constant baseline. By excluding predictors with a variance inflation factor (VIF) larger than 5, we ensure that predictors are independent enough for meaningful PFI scores. The PFI is averaged over test sets and imputations in the same way as the performance metrics.

## 4 RESULTS

### 4.1 RQ1: Dropout predictability and predictor importance of different models

All models' performance in dropout prediction on the general population is summarized in Table 2. The metrics are almost invariant across the ten missing data imputations and hence averaged, enabling us to compare differences between the models independent of the data imputation (see standard deviations in supplementary material). The relatively small differences in the accuracy metric

**Table 1: Sets of hyperparameter values used in grid search tuning procedure by model. |predictors|: total number of predictors**

| Model                  | Hyperparameter                           | Values                             |
|------------------------|--|------------------------------------|
| Logistic regression    | -  | -                                  |
| k-Nearest Neighbors    | Number of nearest neighbors              | 9, 19, 39, 59, 99, 199, 299        |
| Naive Bayes            | Laplace regularization                   | 0, 0.1, 0.5, 1.0                   |
| Neural Network         | Number of neurons in first layer         | 256, 512, 1024                     |
|                        | Neurons in second layer (ratio of first) | 0%, 25%, 50%                       |
|                        | Dropout rate after each layer            | 0%, 50%                            |
|                        | Training epochs                          | 5, 10                              |
| Random Forest          | Number of grown trees                    | 500, 1000, 1500                    |
|                        | Number of candidates for split           | 3, 5, 6, 7                         |
| Support Vector Machine | Kernel                                   | RBF, linear, polynomial            |
|                        | Cost                                     | 0.1, 0.5, 1.0                      |
|                        | Gamma (for RBF kernel)                   | {0.01, 0.1, 1.0, 10.0}/ predictors |
|                        | Class weight for dropouts                | 1, 3, 5                            |

**Table 2: Estimated performance metrics of fine-tuned models on the entire data respectively up to one year and two years after first enrollment.**

| Time of prediction     | After first year |              |              | After second year |              |              |
|------------------------|------------------|--------------|--------------|-------------------|--------------|--------------|
| Model                  | AUPRC            | AUROC        | Accuracy     | AUPRC             | AUROC        | Accuracy     |
| Baseline               | 0.131            | 0.500        | 0.869        | 0.112             | 0.500        | 0.888        |
| Logistic Regression    | 0.553            | 0.816        | 0.898        | 0.712             | 0.899        | 0.938        |
| k-Nearest Neighbors    | 0.493            | 0.785        | 0.890        | 0.620             | 0.863        | 0.922        |
| Naive Bayes            | 0.501            | 0.789        | 0.890        | 0.680             | 0.885        | 0.928        |
| Neural Network         | 0.561            | 0.823        | 0.899        | 0.722             | 0.903        | 0.939        |
| Random Forest          | <b>0.577</b>     | <b>0.827</b> | <b>0.902</b> | <b>0.735</b>      | <b>0.908</b> | <b>0.940</b> |
| Support Vector Machine | 0.556            | 0.821        | 0.898        | 0.718             | 0.903        | 0.938        |

result from the high baseline accuracy. As predicted, the AUPRC shows a larger variation than the AUROC, which empirically justifies its use. Especially the area-based metrics show that all models perform above the baseline. The RF model emerges as the best by dominating all metrics by a notable margin at both prediction times. The neural network does perform second-best, whereas the traditional logistic regression shows performance almost on par with the SVM. These four models will be referred to as the top 4.

We define the most important predictors of student dropout in terms of their PFI as the most insightful information for administrators. These predictor importances are listed in Table 3. The number of essential predictors is relatively sparse compared to the overall number of predictors (39). Depending on the model and time of prediction, only seven to ten predictors impact the AUPRC by more than 1%.

When predicting dropout after a student's first year, performance indicators related to grades and passing are most important, along with continuous enrollment (i.e., a higher number of enrolled terms) and being on track in the current year of study in all portrayed models. The acquisition of English as a second language and other demographic factors such as individual and peer ethnicity also impact the prediction. Overall, recorded behavior within the first year

at college seems to have more impact than demographic variables. This pattern does not only occur for random forests but across all well-performing models. Depending on the specific model, some rankings might change by one or two positions, but the pattern is comparable.

Remarkable differences emerge if we compare the importance between the two prediction time points. A fine-grained analysis of importance change over time is the subject of RQ2. Two years after initial enrollment, the number of enrolled terms is by far the best predictor and is preferred by every top-4 model. In contrast, the college GPA has lost its initial importance. By the ensemble of the top 4 models, passed courses are preferred over the college GPA as a predictor. The number of passed courses and year of study become much more important. The fact that English learners are less likely to drop out is reflected by all the models and maintains a PFI of more than 1% among the pre-entry predictors. Also, at this point, the ranking changes slightly when considering different models but the magnitudes and relative importance of predictor pairs are model-independent.

**Table 3: The 15 most important predictors after one and two years measured by Permutation Feature Importance. RF: Random Forest. Top 4: Mean of Random Forest, Linear Regression, Support Vector Machines, and Neural Network. Left: after one year, Right: after two years.**

|    | Predictor                     | RF    | Top 4 |
|----|-------------------------------|-------|-------|
| 1  | College GPA                   | 23.0% | 16.1% |
| 2  | Number of enrolled terms      | 15.0% | 16.2% |
| 3  | Passed courses                | 7.4%  | 11.6% |
| 4  | Number of school affiliations | 4.9%  | 3.6%  |
| 5  | English language learner      | 3.5%  | 3.8%  |
| 6  | Current year of study         | 2.8%  | 1.6%  |
| 7  | Number of declared majors     | 2.6%  | 1.9%  |
| 8  | Primarily affiliated school   | 2.0%  | 2.6%  |
| 9  | Ethnicity                     | 1.3%  | 2.6%  |
| 10 | High school GPA               | 1.1%  | 0.5%  |
| 11 | Same ethnicity in courses     | 0.9%  | 0.6%  |
| 12 | Residency at application      | 0.7%  | 0.8%  |
| 13 | Distance to university        | 0.7%  | 0.6%  |
| 14 | Pre-entry reading score       | 0.5%  | 0.5%  |
| 15 | Courses per term              | 0.5%  | 0.4%  |

|    | Predictor                     | RF    | Top 4 |
|----|-------------------------------|-------|-------|
| 1  | Number of enrolled terms      | 48.3% | 54.1% |
| 2  | College GPA                   | 8.5%  | 3.5%  |
| 3  | Passed courses                | 4.8%  | 7.5%  |
| 4  | Current year of study         | 4.1%  | 2.8%  |
| 5  | Number of school affiliations | 2.5%  | 1.3%  |
| 6  | English language learner      | 2.1%  | 2.3%  |
| 7  | Primarily affiliated school   | 1.0%  | 1.4%  |
| 8  | Number of declared majors     | 0.9%  | 0.4%  |
| 9  | Distance to university        | 0.5%  | 0.4%  |
| 10 | Same ethnicity in courses     | 0.5%  | 0.3%  |
| 11 | Ethnicity                     | 0.4%  | 1.1%  |
| 12 | Courses per term              | 0.4%  | 0.2%  |
| 13 | Pre-entry reading score       | 0.3%  | 0.3%  |
| 14 | Female                        | 0.3%  | 0.5%  |
| 15 | High school GPA               | 0.3%  | 0.1%  |

## 4.2 RQ2: Dropout predictability and predictor importance over time

The results for RQ2 and RQ3 are based on the RF because it performed best and due to comparable predictor rankings (and performance differences) between models. Figure 1 depicts the performance of this model trained with data up to a certain time point relative to the initial enrollment. The general trend is an increased model performance with a growing observation span. The fact that AUPRC is not monotonically increasing over time is due to the change in the base rate of dropout, which starts to shrink after five terms when the first students are known to have dropped out. The accuracy shows no improvement over the baseline at the pre-entry time point. Nevertheless, the two area-based metrics demonstrate that the model already outperforms the baseline at this early time point. Over time, and especially during the beginning of the second year, accuracy stands slightly out from the baseline. The AUROC shows a steady increase, most strongly in the first terms. The most variation and increase can be observed in the AUPRC, which was expected to be the most suitable evaluation metric in this scenario.

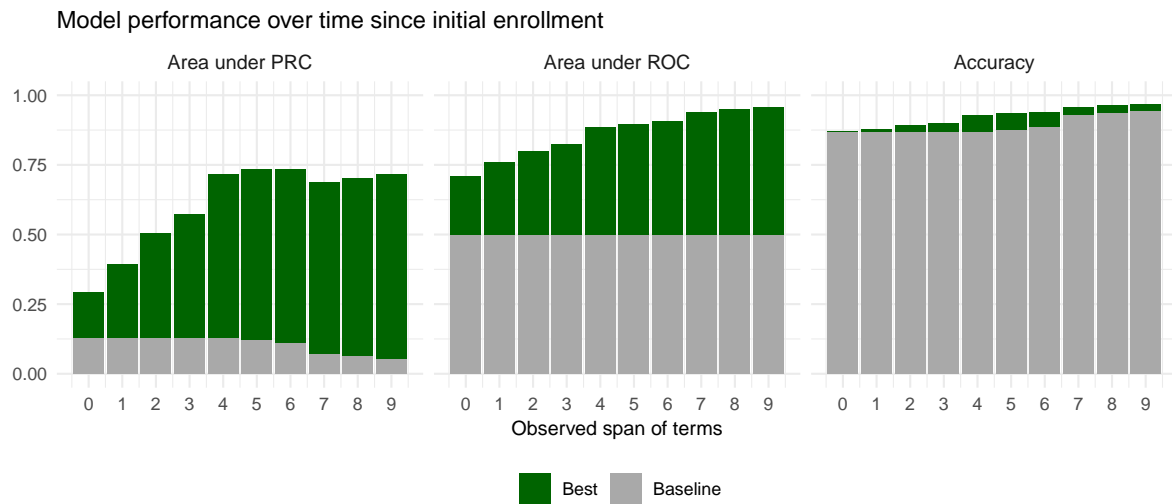
In Figure 2, the change of relative predictor importance is plotted. At the time of initial enrollment, a mixture of demographic and performance indicators weighs the most in the prediction. The status as an English language learner contributes the most information, closely followed by high school GPA, the best AP exam score, and the student's ethnicity. The geographic information (mostly distance from home), the pre-entry reading score, and gender play minor but significant roles. All other predictors influence the performance by less than 4%. The ranking is strongly perturbed as soon as behavioral data from college studies is available. By the end of first term, the GPA leads the ranking by a large margin (22.2%). The English language learner status retains a part of the initial information, ranking second. The five most important factors after one term are completed by performance-related indicators: passed courses, number of school affiliations, and number of declared majors. Ethnicity follows with less than 4% impact, which also applies to the high school GPA and best AP score by then.

The relative importance ranking does not consolidate after the first term but continues to change. Most strikingly, the number of enrolled terms drastically increases its predictive importance over time. Variance in this predictor is possible from the second term on and directly reflects in an importance score of 13.7% by then. At that time point, the average grade is still the most informative predictor (27.6%). The relative ranking does not change much within the remainder of the first year of enrollment. After four terms, the number of enrolled terms eventually ranks first by a significant margin. The college GPA gets degraded to less than 10%. English learner status still plays a role, although it is very minor. During the second year, the ranking stays rather stable. During the third year, the ratio of passed courses and the current year of study overtake the importance of GPA. The current year of study slightly gains importance, starting from the third term. Interestingly, the English language learner status gained some meaning during that period.

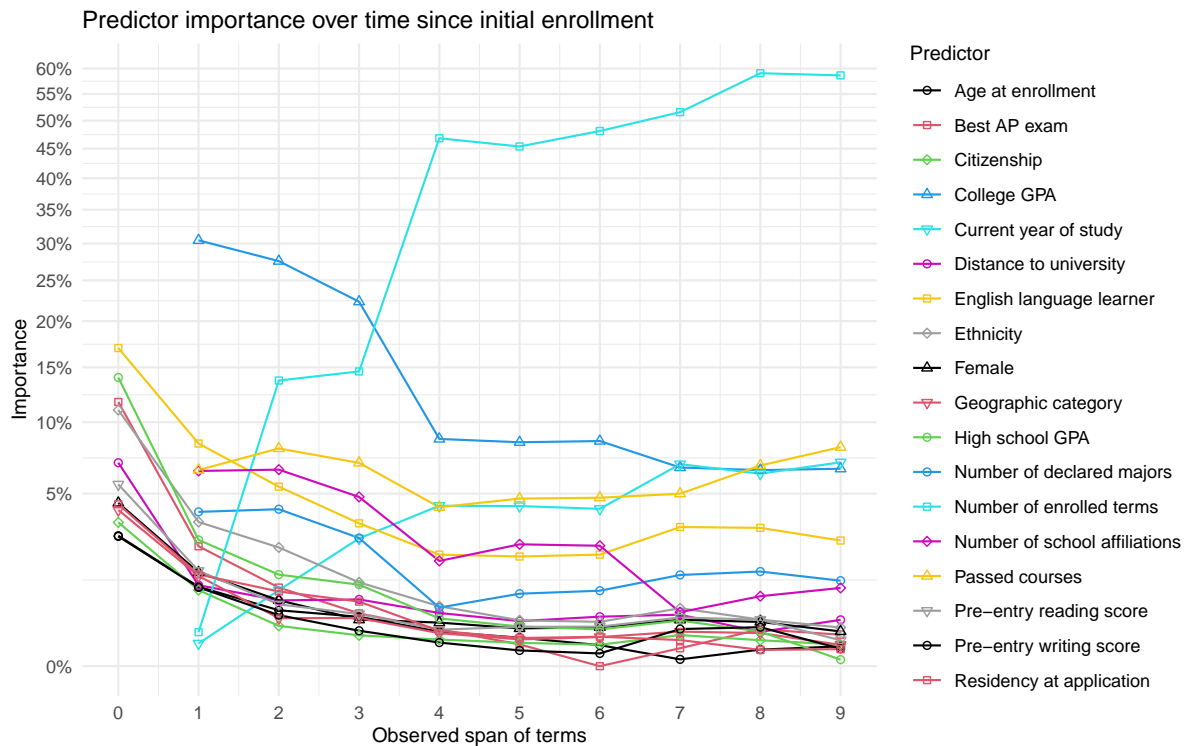
## 4.3 RQ3: Dropout predictability and predictor importance for subpopulations

To compare the dropout predictability and predictor importance between different groups of students, we resort to the AUROC due to its fixed baseline of 0.5 and its sufficient sensitivity to performance differences in RQ2, especially between the second and the fourth term. Figure 3 shows the predictive performance as a function of groups, along with the group sizes and dropout rates. Although there are some unbalanced grouping factors in terms of population size (e.g., only 33.1% of students stem from low-income families) and dropout rates (18.9% of students from an underrepresented minority drop out), the predictability of dropout does not vary enormously between the respective groups. Prediction seems slightly easier for first-generation students, non-STEM majors, and students from low-income families and underrepresented minorities (URM).

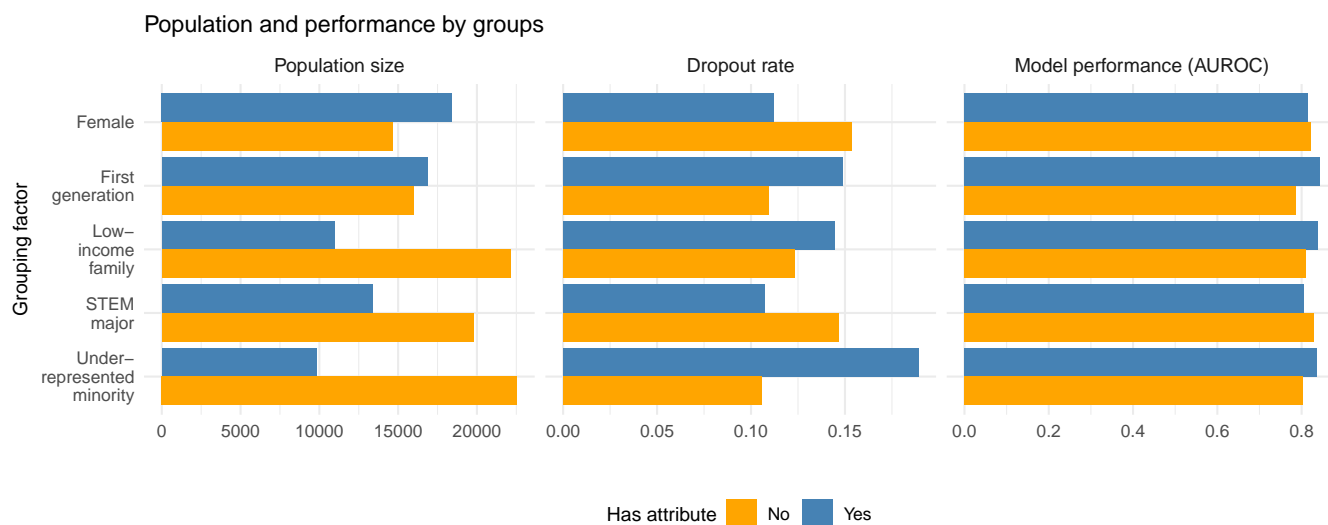
Figure 4 visualizes the importance of predictors between the groups. The generally most important predictor after one year, college GPA, is the top predictor for all subpopulations. Despite its group-independent overall relevance, we can highlight some



**Figure 1: Performance metrics for different time points of predictions against their respective baselines. PRC: precision-recall curve, ROC: receiver-operator curve. The baseline is a random prediction for curve-based metrics, while based on the best possible threshold for accuracy.**



**Figure 2: Predictor importance for different time points of predictions. Predictors always below 2.5% are omitted. Due to the root transformation of the importance for better readability, differences in the area below 5% may seem aggravated.**



**Figure 3: Model performance and population sizes by grouping factors. Both are based on the data available one year after initial enrollment.**

notable differences between groups. Within the PFI of the number of enrolled terms and passed courses ratio, we observe similar between-group differences that also rarely change the overall ranking of importance within one group. For generally less important factors, such as major and school information, current year of study, and English learner status, one can observe minor but potentially significant differences in absolute value.

For female students, the GPA is notably less relevant with respect to dropout than for non-female students which also applies to course passing. The number of enrolled terms contributes almost equal information to both group's predictions. Prediction for female students may benefit more from major and school information and the year of study. Ethnicity is also more relevant for female dropouts whereas English learner status does not vary in its importance between the two groups. First-generation student dropout stronger depends on GPA than its counterpart, for which course passing is more informative. The first-generation dropout prediction benefits more from the number of school affiliations and ethnicity, whereas English learner status and high school GPA show the opposite interaction effect. This trend is mirrored for students from low-income families. Only the number of school affiliations has a less pronounced effect. Moreover, the importance of ethnicity differs much less between the induced groups by this criterion.

Being part of a URM mostly means an increased importance of the same factors than for low-income families. However, English learner status is more important for non-URM students. We also observe the clear pattern that for URM students the high school GPA and residency location at the time of application are irrelevant while having predictive power for the remainder of students. Comparing STEM majors and non-STEM majors also reveals a prominent difference in GPA importance, being more predictive in the case of STEM majors. Course passing is also more helpful in predicting this group's dropout, whereas the number of enrolled

terms is much more interesting for non-STEM majors. The same applies to the number of school affiliations, which are irrelevant for STEM dropout in this model. The pace of study is only relevant for non-STEM dropout.

## 5 DISCUSSION AND CONCLUSION

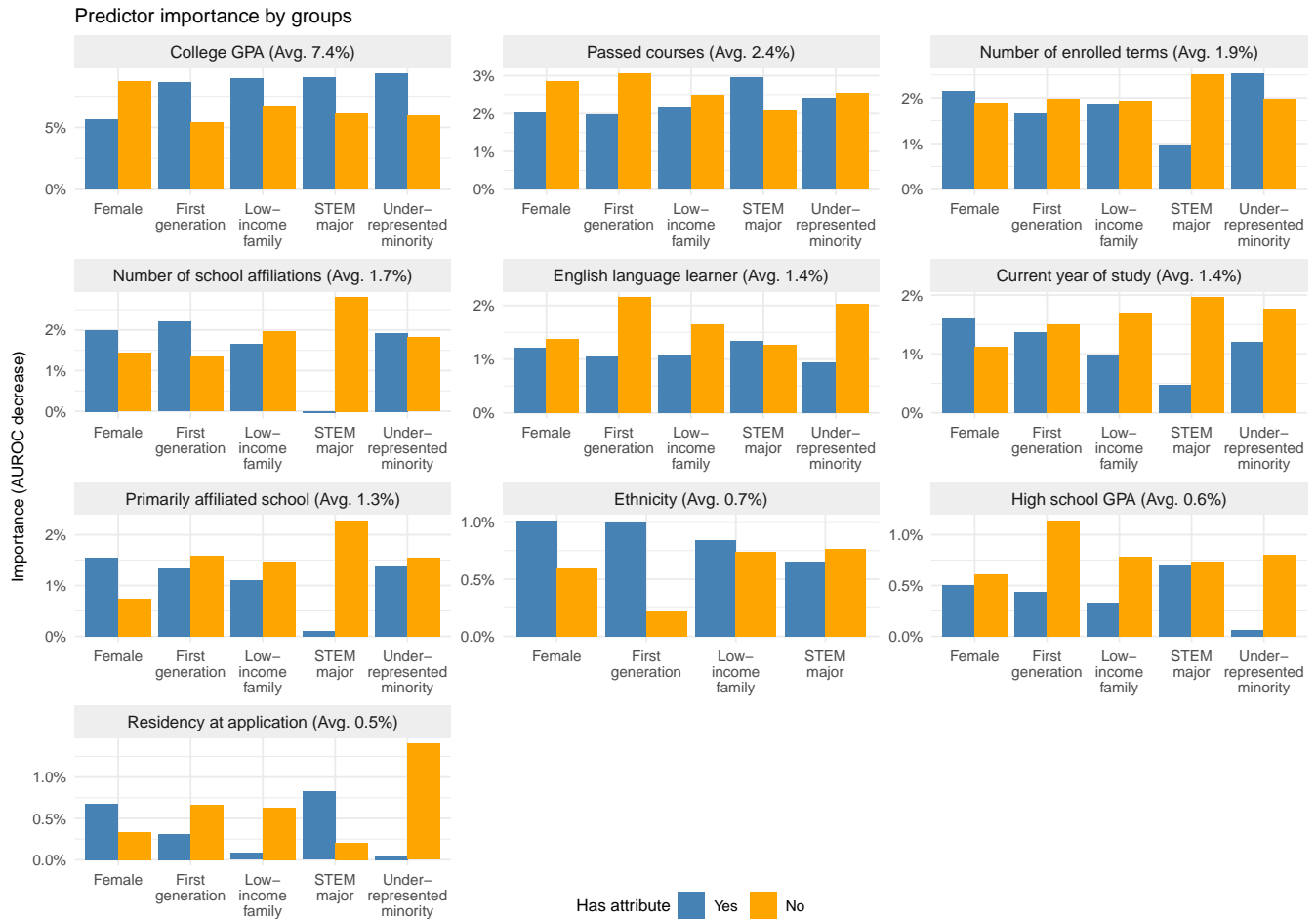
This study has successfully employed modern classification models to predict college dropout, relying on cost-effective large-scale administrative data that can be adopted for an early warning system. The model-agnostic PFI yielded valuable insights into the dropout factors. As novel contributions, we traced the predictability of dropout over a large span of the student lifecycle and the importance of single predictors. Moreover, the analyses address student heterogeneity by distinguishing the predictive importance of factors between important grouping factors of college success.

### 5.1 Scholarly Significance

Our model's performance is mostly comparable with previous studies using the same data type. Studies that made predictions after one term based on institutional data reported an AUROC between 0.69 and 0.88 (depending on the school) [44] compared to our score of 0.76. In the literature, scores rise to the range between 0.81 and 0.93 after one year of enrollment [7, 9], where we achieve 0.83. Results for datasets from different institutions and regions remain hard to compare. However, considering the low overall impact of sociodemographic factors in our analyses and the standardized federal reporting of many predictors, the findings may be generalizable to other 4-year colleges in the US.

Due to the rich temporal structure of our data, we traced how dropout prediction factors evolve. Although Tinto's theory of dropout [55] emphasizes the longitudinal character of the dropout process, even recent studies did not include this critical dimension of dropout prediction (see Section 2.4). Therefore, we mapped





**Figure 4: Differences in predictor importance between groups when predicting dropout one year after initial enrollment. Twenty-nine predictors with a maximal score of 1% or below for every group are omitted in this plot.**

how pre-entry information decreases in value over time and that college performance data is generally most valuable in the first year, whereas continued enrollment is most important after the first year. Academic integration (as measured by GPA) may become substituted by social integration (indicated by the number of enrolled terms) to predict dropout over time. Interestingly, prior work suggested the reverse [35]. However, course peer composition measures were not highly predictive as potential indicators of social integration. An outstanding overall predictor was students' status as English language learners, which may be related to social integration and is worth investing more in-depth for future research. Overall, administrative data focuses on academic integration and contains limited information regarding social integration.

The large-scale character of our data yielded large enough subpopulations to examine differences in predictive factors. We believe that this interaction analysis, based on ML models, can open up new directions in dropout prediction research. We found significant differences between the dropout factors for certain subgroups. Considering that college GPA, the ratio of passed courses, and the

number of enrolled terms reflect a spectrum from performance to non-performance-related behavior, students from traditionally disadvantaged groups may be more reliant on grades regarding dropout. This could be an exemplified interaction of academic and social integration. It may be worth investigating if these groups are more likely to stay enrolled and perform worse instead of taking a semester off before dropping out. The most striking differences were found for students in STEM majors, whose dropout is better predicted by their grades, and course passing. Non-STEM dropouts are easier identified by the number of enrolled terms, number of majors, and school affiliations. Again, academic integration may be more crucial in STEM fields compared to other subjects for the student's commitment that eventually leads to dropout, resonating with the literature [20].

## 5.2 Implications for Educational Stakeholders

Our results further support the value of administrative data that every institution already has in standard formats [24], compared to costly interviews or cleaning-intensive process data. We provide

our analysis scripts to encourage other universities to replicate the analyses with their data. Generally, our methodology allows administrators directly or by re-analysis to create a precise and smaller hence cheaper set of dropout predictors. First, the temporal dimension of the prediction can be considered when implementing an early warning system. At the moment of enrollment, predictions are much more error-prone than after the first year. Based on the change over time in dropout predictability, educational stakeholders may choose their individual preference along the trade-off between early interventions against dropout and accumulated evidence for actual dropout risk. Second, the selection of predictors to obtain can now be adapted to the actual time point where a prediction should be made. Although administrative data itself comes at a lower cost compared to other data sources, the integration of different information systems across a higher education institution always implies effort or may be subject to data privacy regulation [9]. As we have shown, pre-entry data rapidly loses value as soon as behavioral data at the college becomes available. Depending on the chosen time point for dropout prediction, administrators can estimate the value of predictors more precisely. Third, an EWS can be tailored to specific subpopulations of the student body using group-specific factors. We found that dropout prediction fortunately works almost equally on groups induced by various grouping factors. However, in the case of STEM majors, the predictor collection can vary as a function of the targeted group.

Overall, an effectively designed EWS may help reduce overall dropout via resource-efficient intervention by administrators. It allows for reducing false negatives by tolerating more false positives using the classification threshold. Nevertheless, the following risks should be mitigated: Biases in helpfulness associated with protected attributes [59], which our results, fortunately, did not reveal; biasing faculty in their behavior towards at-risk students [27], which may be prevented sending warnings directly to students or to central consulting offices; and misuse of the EWS for admission decisions, which requires appropriate legislation.

### 5.3 Limitations and Future Work

When students at risk are identified, possible interventions may want to address underlying reasons. Most of the identified predictors in this study may serve as an explanation for the risk of dropout but are usually hard or even impossible to change directly. The prediction of dropout still requires careful human examination of individual cases.

Although we use the VIF to make sure that our univariate importance metric is meaningful, there is still an inherent correlation between most predictors. Differences in predictor importance are in this approach naturally dependent on the existence of the other predictors in the training. For example, while high school GPA loses its predictive value relative to the college GPA after the first term it does not mean that it becomes useless. Instead, other predictors may just be more suitable for subsequent performance. This possibility should be considered when deciding in favor or against the survey of certain predictors as an educational stakeholder.

The approach of this study hopes to identify some potential "blind spots" of traditional hypothesis-driven analyses. It provides

more insights for theory development and further hypothesis testing compared to other more data- and algorithm-driven approaches that often underlie EWS used for predicting and reporting risks [37]. Similarly, our approach to using existing large-scale administrative data provides more cost-efficient alternatives to expensive questionnaire-based surveys or complex clickstream data to support educational administrations in their decision-making. Ultimately, we believe that the automatization of EWS may allow for adaptive student support systems to foster more learner-centered environments, enhance learning benefits, and reduce dropout risks [14, 34].

## REFERENCES

- [1] Gökhan Akçapınar, Arif Altun, and Petek Aşkar. 2019. Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education* 16, 1 (2019), 1–20. <https://doi.org/10.1186/s41239-019-0172-z>
- [2] Joseph J. Allaire and François Chollet. 2023. *keras: R Interface to 'Keras'*. <https://CRAN.R-project.org/package=keras> R package version 2.11.1.
- [3] Joseph J. Allaire and Yuan Tang. 2022. *tensorflow: R Interface to 'TensorFlow'*. <https://CRAN.R-project.org/package=tensorflow> R package version 2.11.0.
- [4] Isabelle Archambault, Michel Janosz, Véronique Dupéré, Marie-Christine Brault, and Marie Mc Andrew. 2017. Individual, social, and family factors associated with high school dropout among low-SES youth: Differential effects as a function of immigrant status. *British Journal of Educational Psychology* 87, 3 (2017), 456–477. <https://doi.org/10.1111/bjep.12159>
- [5] Kimberly E. Arnold and Matthew D. Pistilli. 2012. Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. 267–270. <https://doi.org/10.1145/2330601.2330666>
- [6] Richard Arum, Jacquelynne S. Eccles, Jutta Heckhausen, Gabe Avakian Orona, Luise von Keyserlingk, Christopher M. Wegemer, Charles E. Wright, and Katsumi Yamaguchi-Pedroza. 2021. A framework for measuring undergraduate learning and growth. *Change: The Magazine of Higher Learning* 53, 6 (2021), 51–59. <https://doi.org/10.1080/00091383.2021.1987810>
- [7] Lovenoor Aulck, Dev Nambi, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. 2019. Mining University Registrar Records to Predict First-Year Undergraduate Attrition. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*. 9–18.
- [8] Cédric Beaulac and Jeffrey S. Rosenthal. 2019. Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education* 60, 7 (2019), 1048–1064. <https://doi.org/10.1007/s11162-019-09546-y>
- [9] Johannes Berens, Kerstin Schneider, Simon Gortz, Simon Oster, and Julian Burghoff. 2019. Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining* 11, 3 (2019), 1–41. <https://doi.org/10.5281/zenodo.3594771>
- [10] Matthew L. Bernacki, Michelle M. Chavez, and P. Merlin Uesbeck. 2020. Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education* 158 (2020), 103999. <https://doi.org/10.1016/j.compedu.2020.103999>
- [11] Kelli A. Bird, Benjamin L. Castleman, Zachary Mabel, and Yifeng Song. 2021. Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education. *AERA Open* 7 (2021). <https://doi.org/10.1177/23328584211037630>
- [12] Michael Geoffrey Brown, R. Matthew DeMonbrun, Steven Lonn, Stephen J. Aguilar, and Stephanie D. Teasley. 2016. What and when: the role of course type and timing in students' academic performance. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. 459–468. <https://doi.org/10.1145/2883851.2883907>
- [13] Joana R. Casanova, Antonio Cervero, José C. Núñez, Leandro S. Almeida, and Ana Bernardo. 2018. Factors that determine the persistence and dropout of university students. *Psicothema* 30, 4 (2018), 408–414. <https://doi.org/10.7334/psicothema2018.155>
- [14] Mohamed Amine Chatti and Arham Muslim. 2019. The PERLA Framework: Blending Personalization and Learning Analytics. *International Review of Research in Open and Distributed Learning* 20, 1 (2019). <https://doi.org/10.19173/irrodl.v20i1.3936>
- [15] Gerben W. Dekker, Mykola Pechenizkiy, and Jan M. Vleeshouwers. 2009. Predicting students drop out: A case study. In *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009, Cordoba, Spain*. 41–50.
- [16] Francesca Del Bonifro, Maurizio Gabbriellini, Giuseppe Lisanti, and Stefano Pio Zingaro. 2020. Student Dropout Prediction. Springer, Cham, 129–140. [https://doi.org/10.1007/978-3-030-52237-7\\_11](https://doi.org/10.1007/978-3-030-52237-7_11)

- [17] Dursun Delen. 2010. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems* 49, 4 (2010), 498–506. <https://doi.org/10.1016/j.dss.2010.06.003>
- [18] Stephen L. Desjardins, Dennis A. Ahlburg, and Brian Patrick McCall. 2002. A Temporal Investigation of Factors Related to Timely Degree Completion. *The Journal of Higher Education* 73, 5 (2002), 555–581. <https://doi.org/10.1353/jhe.2002.0042>
- [19] Alejandro Díaz Mujica, María V. Pérez Villalobos, Ana B. Bernardo Gutiérrez, Antonio Cervero Fernández-Castañón, and Julio A. González-Piñeda. 2019. Affective and cognitive variables involved in structural prediction of university dropout. *Psicothema* 31, 4 (2019), 429–436. <https://doi.org/10.7334/psicothema2019.124>
- [20] Sandra L Dika and Mark M D'Amico. 2016. Early experiences and integration in the persistence of first-generation college students in STEM and non-STEM majors. *Journal of Research in Science Teaching* 53, 3 (2016), 368–383. <https://doi.org/10.1002/tea.21301>
- [21] Mica Estrada, Myra Burnett, Andrew G. Campbell, Patricia B. Campbell, Wilfred F. Denetclaw, Carlos G. Gutiérrez, Sylvia Hurtado, Gilbert H. John, John Matsui, Richard McGee, Camellia Moses Okpodu, T. Joan Robinson, Michael F. Summers, Maggie Werner-Washburne, and MariaElena Zavala. 2016. Improving Underrepresented Minority Student Persistence in STEM. *CBE—Life Sciences Education* 15, 3 (2016), e5. <https://doi.org/10.1187/cbe.16-01-0038>
- [22] Lorin P. Fagioli, Rachel Baker, and Gabe Avakian Orona. 2020. The role of non-cognitive variables in identifying community college students in need of targeted supports. *Research in Higher Education* 61 (2020), 725–763. <https://doi.org/10.1007/s11162-020-09588-7>
- [23] Kiran Fahd, Sitalakshmi Venkatraman, Shah J. Miah, and Khandakar Ahmed. 2022. Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Education and Information Technologies* 27, 3 (2022), 3743–3775. <https://doi.org/10.1007/s10639-021-10741-7>
- [24] Christian Fischer, Zachary A. Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. 2020. Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education* 44, 1 (2020), 130–160. <https://doi.org/10.3102/0091732X20903304>
- [25] Bobby J Franklin and Stephen B Trouard. 2016. Comparing dropout predictors for two state-level panels using Grade 6 and Grade 8 data. *The Journal of Educational Research* 109, 6 (2016), 631–639. <https://doi.org/10.1080/00220671.2015.1016601>
- [26] Guang-Hui Fu, Lun-Zhao Yi, and Jianxin Pan. 2019. Tuning model parameters in class-imbalanced learning with precision-recall curve. *Biometrical Journal* 61, 3 (2019), 652–664. <https://doi.org/10.1002/bimj.201800148>
- [27] Sarah Gentrup, Georg Lorenz, Cornelia Kristen, and Irena Kogan. 2020. Self-fulfilling prophecies in the classroom: Teacher expectations, teacher feedback and student achievement. *Learning and Instruction* 66 (2020), 101296.
- [28] Geraldine Gray, Colm McGuinness, Philip Owende, and Markus Hofmann. 2016. Learning Factor Models of Students at Risk of Failing in the Early Stage of Tertiary Education. *Journal of Learning Analytics* 3, 2 (2016), 330–372. <https://doi.org/10.18608/jla.2016.32.20>
- [29] Karin Hartl. 2019. *The Application Potential of Data Mining in Higher Education Management: A Case Study Based on German Universities*. Ph.D. Dissertation. Karlsruhe Institut für Technologie (KIT). <https://doi.org/10.5445/IR/1000096613>
- [30] Arto Hellas, Petri Ihanntola, Andrew Petersen, Vangel V. Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, and Soohyun Nam Liao. 2018. Predicting Academic Performance: A Systematic Literature Review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (Larnaca, Cyprus). Association for Computing Machinery, New York, NY, USA, 175–199. <https://doi.org/10.1145/3293881.3295783>
- [31] Serge Herzog. 2005. Measuring Determinants of Student Return VS. Dropout/Stopout VS. Transfer: A First-to-Second Year Analysis of New Freshmen. *Research in Higher Education* 46, 8 (2005), 883–928. <https://doi.org/10.1007/s11162-005-6933-7>
- [32] Wonjoon Hong, Matthew L Bernacki, and Harsha N Perera. 2020. A latent profile analysis of undergraduates' achievement motivations and metacognitive behaviors, and their relations to achievement in science. *Journal of Educational Psychology* 112, 7 (2020), 1409. <https://doi.org/10.1037/edu0000445>
- [33] Emma Howard, Maria Meehan, and Andrew Parnell. 2018. Contrasting prediction methods for early warning systems at undergraduate level. *The Internet and Higher Education* 37 (2018), 66–75. <https://doi.org/10.1016/j.iheduc.2018.02.001>
- [34] Dirk Ifenthaler, David Gibson, Doreen Prasse, Atsushi Shimada, and Masanori Yamada. 2020. Putting learning back into learning analytics: actions for policy makers, researchers, and practitioners. *Educational Technology Research and Development* (2020), 1–20. <https://doi.org/10.1007/s11423-020-09909-8>
- [35] Terry T. Ishitani. 2016. Time-varying effects of academic and social integration on student persistence for first and second years in college: National data approach. *Journal of College Student Retention: Research, Theory & Practice* 18, 3 (2016), 263–286. <https://doi.org/10.1177/1521025115622781>
- [36] Terry T. Ishitani and Stephen L. Desjardins. 2002. A longitudinal investigation of dropout from college in the United States. *Journal of College Student Retention: Research, Theory & Practice* 4, 2 (2002), 173–201. <https://doi.org/10.2190/V4EN-NW42-742Q-2NTL>
- [37] Sandeep M. Jayaprakash, Erik W. Moody, Eitel J.M. Lauria, James R. Regan, and Joshua D. Baron. 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics* 1, 1 (2014), 6–47. <https://doi.org/10.18608/jla.2014.11.3>
- [38] Lucas M. Jenó, Anne G. Danielsen, and Arild Raaheim. 2018. A prospective investigation of students' academic achievement and dropout in higher education: a Self-Determination Theory approach. *Educational Psychology* 38, 9 (2018), 1163–1184. <https://doi.org/10.1080/01443410.2018.1502412>
- [39] Lorenz Kemper, Gerrit Vorhoff, and Berthold U. Wigger. 2020. Predicting student dropout: A machine learning approach. *European Journal of Higher Education* 10, 1 (2020), 28–47. <https://doi.org/10.1080/21568235.2020.1718520>
- [40] Sotiris B. Kotsiantis, C.J. Pierrakeas, and Panayiotis E. Pintelas. 2003. Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems: 7th International Conference, KES 2003, Oxford, UK, September 2003. Proceedings, Part II* 7. Springer, 267–274. [https://doi.org/10.1007/978-3-540-45226-3\\_37](https://doi.org/10.1007/978-3-540-45226-3_37)
- [41] Zlatko J. Kovacic. 2010. Early Prediction of Student Success: Mining Students Enrolment Data. In *InSITE Conference*. Informing Science Institute, 647–665. <https://doi.org/10.28945/1281>
- [42] Andy Liaw and Matthew Wiener. 2002. Classification and Regression by randomForest. *R News* 2, 3 (2002), 18–22. <https://CRAN.R-project.org/doc/Rnews/>
- [43] Cindi Mason, Janet Twomey, David Wright, and Lawrence Whitman. 2018. Predicting Engineering Student Attrition Risk Using a Probabilistic Neural Network and Comparing Results with a Backpropagation Neural Network and Logistic Regression. *Research in Higher Education* 59, 3 (2018), 382–400. <https://doi.org/10.1007/s11162-017-9473-z>
- [44] Sandra C Matz, Christina S Bukow, Heinrich Peters, Christine Deacons, Alice Dinu, and Clemens Stachl. 2023. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports* 13, 1 (2023), 5705. <https://doi.org/10.1038/s41598-023-32484-w>
- [45] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2023. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. <https://CRAN.R-project.org/package=e1071> R package version 1.7-13.
- [46] José María Ortiz-Lozano, Antonio Rua-Vieites, Paloma Bilbao-Calabuig, and Martí Casadesús-Fa. 2020. University student retention: Best time and data to identify undergraduate students at risk of dropout. *Innovations in Education and Teaching International* 57, 1 (2020), 74–85. <https://doi.org/10.1080/14703297.2018.1502090>
- [47] Andrew Phippen, L. Sheppard, and Steven Furnell. 2004. A practical evaluation of Web analytics. *Internet Research* 14, 4 (2004), 284–293. <https://doi.org/10.1108/10662240410555306>
- [48] L. Reisel and I. Brekke. 2010. Minority Dropout in Higher Education: A Comparison of the United States and Norway Using Competing Risk Event History Analysis. *European Sociological Review* 26, 6 (2010), 691–712. <https://doi.org/10.1093/esr/jcp045>
- [49] Ido Roll and Philip H. Winne. 2015. Understanding, evaluating, and supporting self-regulated learning using learning analytics. *Journal of Learning Analytics* 2, 1 (2015), 7–12. <https://doi.org/10.18608/jla.2015.21.2>
- [50] Rannveig Grom Sæle, Tove Irene Dahl, Tore Sorlie, and Oddgeir Friberg. 2023. Note to first-year university students: Just do it! In the end, the fact that you study may be more important than how you study. *Uniped* 46, 1 (2023), 28–42. <https://doi.org/10.18261/uniped.46.1.4>
- [51] Dalia Abdulkareem Shafiq, Mohsen Marjani, Riyaz Ahamed Ariyaluran Habeeb, and David Asirvatham. 2022. Student retention using educational data mining and predictive analytics: a systematic literature review. *IEEE Access* (2022), 72480–503. <https://doi.org/10.1109/access.2022.3188767>
- [52] Martin Solis, Tania Moreira, Roberto Gonzalez, Tatiana Fernandez, and Maria Hernandez. 2018. Perspectives to predict dropout in university students with machine learning. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWobi)*. IEEE, 1–6. <https://doi.org/10.1109/iwobi.2018.8464191>
- [53] Suhyun Suh, Jingyo Suh, and Irene Houston. 2007. Predictors of categorical at-risk high school dropouts. *Journal of Counseling & Development* 85, 2 (2007), 196–203. <https://doi.org/10.1002/j.1556-6678.2007.tb00463.x>
- [54] Hasan Tanvir and Irene-Angelica Chounta. 2021. Exploring the Importance of Factors Contributing to Dropouts in Higher Education over Time. In *Proceedings of The 14th International Conference on Educational Data Mining*. 502–509.
- [55] Vincent Tinto. 1975. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research* 45, 1 (1975), 89–125. <https://doi.org/10.3102/00346543045001089>
- [56] Stef van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45, 3 (2011), 1–67. <https://doi.org/10.18637/jss.v045.i03>

- [57] W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Springer New York. <https://doi.org/10.1007/978-0-387-21706-2>
- [58] K. Wagner, A. Merceron, and P. Sauer. 2020. Accuracy of a cross-program model for dropout prediction in higher education. In *Companion Proceedings of the 10th International Learning Analytics & Knowledge Conference (LAK 2020)*. 744–749.
- [59] Renzhe Yu, Hansol Lee, and René F. Kizilcec. 2021. Should College Dropout Prediction Models Include Protected Attributes?. In *Proceedings of the Eighth ACM Conference on Learning @ Scale*. ACM, New York, NY, USA, 91–100. <https://doi.org/10.1145/3430895.3460139>
- [60] Zitong Zhao, Pan Deng, and Jianjun Zhou. 2022. Identifying Longitudinal Attendance Patterns through Student Subpopulation Distribution Comparison. In *Proceedings of the 15th International Conference on Educational Data Mining*. 640. <https://doi.org/10.5281/zenodo.6853034>