My Reading List

BI−based Methodology for Analyzing Higher Education: A Case Study of Dropout Phenomenon in Information Systems Courses

André Menolli, Flávio Horita, José Jorge L. Dias, Ricardo Coelho

Proceeding • 2020

MOOC student dropout: pattern and prevention

Yunfan Chen, Ming Zhang

Proceeding • 2017

Broad Learning System for Predicting Student Dropout in Massive Open Online Courses

Shuang Lai, Yuxin Zhao, Yuqing Yang

Proceeding • 2020

University Dropout Prevention through the Application of Big Data

Yeajou Shiau

Proceeding • 2020

Should College Dropout Prediction Models Include Protected Attributes?

Renzhe Yu, Hansol Lee, René F. Kizilcec

Proceeding • 2021

A Dropout Prediction Framework Combined with Ensemble Feature Selection

Dan Ai, Tiancheng Zhang, Ge Yu, Xinying Shao

Proceeding • 2020

Data Analytics for Predicting Dropout

Gioconda Riofrío-Calderón, María-Soledad Ramírez-Montoya, María-José Rodríguez-Conde

Proceeding • 2021

Correlational Analysis Between School Performance and Municipal Indicators in Brazil Supported by Linked Open Data

Bruno Elias Penteado

Proceeding • 2016

BI−based Methodology for Analyzing Higher Education: A Case Study of Dropout Phenomenon in Information Systems Courses

Authors:

André Menolli • **Northeast University of Paraná and State University of Londrina, Brazil**

Flávio Horita • **Federal University of ABC, Brazil**

José Jorge L. Dias • **Federal University of Paraiba, Brazil**

Ricardo Coelho • **Northeast University of Paraná, Brazil**

View in Digital library

Through a Design Science Research method, this paper presents a BI-based methodology to analyze the Information System (IS) courses, in particular the dropout phenomenon and their enrolled students, using open public data from the Anísio Teixeira National Institute for Educational Studies and Research. Data related to Higher Education in Brazil were analyzed and understood, to create a dimensional model. Based on this, a Data Warehouse was created, and the original data was then extracted, transformed and loaded, and a Business Intelligence (BI) environment was deployed and set. This BI environment was utilized for carrying out several descriptive statistics analyses about the dropout phenomenon. Results showed differences in dropout rates in the dimensions of race, type of city where the course is located, type of institution (public or private), teaching modality (on-site or e-learning), year of entry into course, and course period. Thus, this paper increase the knowledge about IS courses and encourages discussion about dropout phenomenon.

MOOC student dropout: pattern and prevention

Authors:

Yunfan Chen • Peking University, Haidian, Beijing, China

Ming Zhang • Peking University, Haidian, Beijing, China

View in Digital library

Massive Open Online Course (MOOC) is a completely new education method which appeared in 2012. MOOC is great for life-long learning and educational resource sharing. Until now, there are more than twenty MOOC platforms online. Some studies pointed out that students, however, can hardly follow a course till the end on MOOCs. For most courses, less than 13% of students could follow. Such a high dropout rate will restrict development of MOOCs in the future. This paper shows an observation of MOOC data. A statistic analysis is applied to students' behavioral data. The result is discussed in the perspective of course design on MOOC platform. A general system for predicting students' dropout is developed. With unsupervised learning, the system can fit on different on-going courses. A discussion of this system with Data Structures and Algorithms course is conducted. We also apply the system to Introduction to Computing course to test scalability. Additionally, based on the research above, some suggestions for instructing students on MOOC are given.

Broad Learning System for Predicting Student Dropout in Massive Open Online Courses

Authors:

Shuang Lai • School of Humanities, Economics and Law, Northwestern Polytechnical University, Xi'an, China

Yuxin Zhao • School of Humanities, Economics and Law, Northwestern Polytechnical University, Xi'an, China

Yuqing Yang • School of Humanities, Economics and Law, Northwestern Polytechnical University, Xi'an, China

View in Digital library

At a time when the number of Massive Open Online Courses (MOOCs) users, courses and participating universities is increasing rapidly, a short-time training and reliable prediction model for MOOC extremely high dropout rates is needed. This paper proposes a MOOC dropout prediction model which is based on Broad Learning System (BLS) for MOOC dropout prediction. The model first maps the input into the feature node layer, then generates the enhanced node layer according to the feature node layer through activation, and finally performs linear transformation by combining the feature layer and the enhancement layer. The output layer is used for dropout prediction. Experiments are carried out on the dataset provided by KDD CUP 2015, and the experimental results show that the BLS significantly reduces training time and has a better prediction of dropouts than other mainstream research methods.

University Dropout Prevention through the Application of Big Data

Author:

[Yeajou Shiau](#) • Zhaoqing University, China

View in Digital library

Publication:

Proceeding

This study explores the reasons for the suspension and dropout of full-time university students at a university in Taiwan and suggests better timing and strategies for student counseling. In this study, the narrative statistical analysis is used to analyze and discuss the sample objects, and then use data mining technology to find characteristic phenomena and classification conditions of the students who are suspended or dropout. Other studies related to dropouts rarely use leading indicators to predict the student dropout probability in real-time, most likely because of the timeliness and availability of student data. Therefore, this study proposes to use daily changes in absence indicators as predictive variables. Through the use of discriminant analysis to construct discriminant functions, the coefficient value of each student's withdrawal from university and early warning threshold for determining withdrawal from university can be presented in real-time in order to effectively provide the student with immediate counseling.

Should College Dropout Prediction Models Include Protected Attributes?

Authors:

[Renzhe Yu](#) • **University of California, Irvine, Irvine, CA, USA**

[Hansol Lee](#) • **Cornell University, Ithaca, NY, USA**

[René F. Kizilcec](#) • **Cornell University, Ithaca, NY, USA**

View in Digital library

Publication:

Proceeding

Early identification of college dropouts can provide tremendous value for improving student success and institutional effectiveness, and predictive analytics are increasingly used for this purpose. However, ethical concerns have emerged about whether including protected attributes in these prediction models discriminates against underrepresented student groups and exacerbates existing inequities. We examine this issue in the context of a large U.S. research university with both residential and fully online degree-seeking students. Based on comprehensive institutional records for the entire student population across multiple years (N = 93,457), we build machine learning models to predict student dropout after one academic year of study and compare the overall performance and fairness of model predictions with or without four protected attributes (gender, URM, first-generation student, and high financial need). We find that including protected attributes does not impact the overall prediction performance and it only marginally improves the algorithmic fairness of predictions. These findings suggest that including protected attributes is preferable. We offer guidance on how to evaluate the impact of including protected attributes in a local context, where institutional stakeholders seek to leverage predictive analytics to support student success.

Should College Dropout Prediction Models Include Protected Attributes?

Renzhe Yu
University of California, Irvine
Irvine, CA, USA
renzhey@uci.edu

Hansol Lee
Cornell University
Ithaca, NY, USA
hl838@cornell.edu

René F. Kizilcec
Cornell University
Ithaca, NY, USA
kizilcec@cornell.edu

ABSTRACT

Early identification of college dropouts can provide tremendous value for improving student success and institutional effectiveness, and predictive analytics are increasingly used for this purpose. However, ethical concerns have emerged about whether including protected attributes in these prediction models discriminates against underrepresented student groups and exacerbates existing inequities. We examine this issue in the context of a large U.S. research university with both residential and fully online degree-seeking students. Based on comprehensive institutional records for the entire student population across multiple years (N = 93,457), we build machine learning models to predict student dropout after one academic year of study and compare the overall performance and fairness of model predictions with or without four protected attributes (gender, URM, first-generation student, and high financial need). We find that including protected attributes does not impact the overall prediction performance and it only marginally improves the algorithmic fairness of predictions. These findings suggest that including protected attributes is preferable. We offer guidance on how to evaluate the impact of including protected attributes in a local context, where institutional stakeholders seek to leverage predictive analytics to support student success.

Author Keywords

Dropout prediction; Predictive analytics; Higher education; Online learning; Algorithmic fairness

CCS Concepts

•**Applied computing → Law, social and behavioral sciences; Education;**

INTRODUCTION

With the rapid development of learning analytics in higher education, data-driven instructional and learning support systems are increasingly adopted in classroom settings, and institution-level analytics systems are used to optimize resource allocation

and support student success on a large scale. A common objective of these systems is the early identification of at-risk students, especially those likely to drop out of college. This type of prediction has significant policy implications because reducing college attrition has been a central task for institutional stakeholders ever since higher education was made accessible to the general public [33]. As of 2018, fewer than two-thirds of college students in the United States graduated within six years, and this share is even smaller at the least selective institutions which serve disproportionately more students from disadvantaged backgrounds [23]. At the same time, the supply of academic, student affairs, and administrative personnel is insufficient to provide just-in-time support to students in need [23]. It is within these resource-strained contexts that predicting dropouts based on increasingly digitized institutional data has the potential to augment the capacity of professionals who work to support student retention and success. Starting with the Course Signals project at Purdue University, an increasing number of early warning systems have explored this possibility at the institutional level [1, 25, 18, 12].

Accurately forecasting which students are likely to drop out is essentially profiling students based on a multitude of student attributes. These attributes often include socio-demographic information that is routinely studied in higher education research. Although the analysis of historical socio-demographic gaps in retention and graduation rates is well established in higher education research [13], it becomes controversial to use these same characteristics when making predictions about the future. For example, is it fair to label a black first-year student as *at risk* based on the higher dropout rate among black students in previous cohorts? The answer may be equivocal [37]. On the one hand, the observed historical gaps capture systematic inequalities in the educational environment of different student groups, which may well apply to future students from the same groups and therefore contribute to similar gaps. In this sense, explicitly using socio-demographic data can result in more accurate predictions and improve the efficiency of downstream interventions and actions based on those algorithmic decisions [34]. On the other hand, from an ethics and equity perspective, the inclusion of socio-demographic variables may lead to discriminatory results if predictive models systematically assign differential predicted values across student groups based on the records of their historical counterparts. When these results are used for decision-making,

stigmas and stereotypes could carry over to future students and reproduce existing inequalities [26, 4].

In this paper, we investigate the issue of using protected attributes in college dropout prediction in real-world contexts. Protected attributes are traits or characteristics based on which discrimination is prescribed as illegal, such as gender, race, age, religion, and genetic information. We examine students in a residential college setting as well as students in fully online degree programs, which have been increasingly represented in formal higher education. In Fall 2018, 16.6% of postsecondary students in the United States were enrolled in exclusively online programs, up from 12.8% in Fall 2012 [35, 38]. The absence of a residential experience exposes students to additional challenges to accountability and engagement, and also makes it harder for faculty and staff members to identify problems with students' well-being and provide timely support. The COVID-19 pandemic has forced most colleges to move instruction online, which will likely increase the importance of online learning in the future of higher education [32]. Predictive analytics are therefore just as useful for online higher education as they are for residential settings for supporting student achievement and on-time graduation. Our findings in both residential and online settings offer practical implications to a broad range of stakeholders in higher education.

By systematically comparing predictive models with and without protected attributes in two higher education contexts, we aim to answer the following two research questions:

1. How does the inclusion of protected attributes affect the overall performance of college dropout prediction?

2. How does the inclusion of protected attributes affect the fairness of college dropout prediction?

This research contributes to the literature on predictive modeling and algorithmic fairness in (higher) education on several dimensions. First, we present one of the largest and most comprehensive evaluation studies of college dropout prediction based on student data over multiple years from a large public research university. This offers robust insights to researchers and institutional stakeholders into how these models work and where they might go wrong. Second, we apply the prediction models with the same features to both residential and online degree settings, which advances our understanding of generalizability across contexts, such as in which environment it is easier to predict dropout and to what degree key predictors differ. Third, we contribute some of the first empirical evidence on how the inclusion of protected attributes affects the fairness of dropout prediction, which can inform equitable higher education policy around the use of predictive modeling.

RELATED WORK

College Dropout Prediction

Decades of research have charted the ecosystem of higher education as a complex journey with "a wide path with twists, turns, detours, roundabouts, and occasional dead ends that many students encounter" and jointly shape their academic and career outcomes [28]. Among the variety of factors that influence students' journey, background characteristics such as demographics, family background, and prior academic history are strong signals of academic, social, and economic resources available to a student before adulthood, which are substantially correlated with college success [11]. For example, ethnic minorities, students from low-income families, and first-generation college students have consistently suffered higher dropout rates than their counterparts [13, 9], and students who belong to more than one of these groups are even more likely to drop out of college. In addition to these largely immutable attributes at college entry, students' experiences in college such as engagement and performance in academic activities are major factors for success. In particular, early course grades are among the best predictors of persistence and graduation, even after controlling for background characteristics [28].

With the advent of the "datafication" of higher education [36], there has been an increasing thrust of research to translate the empirical understanding of dropout risk factors into predictive models of student dropout (or success) using large-scale administrative data [3, 14, 25, 15, 5, 6, 24]. These applications are usually intended to facilitate targeted student support and intervention programs, and the extensive research literature on college success has facilitated feature engineering grounded in theory. For example, Aulck and colleagues [3] used seven groups of freshman features extracted from registrar data to predict outcomes for the entire student population at a large public university in the US. The model achieved an accuracy of 83.2% for graduation prediction and 95.3% for retention. In a more application-oriented study as part of the Open Academic Analytics Initiative (OAAI), Jayaprakash and colleagues [25] developed an early alert system that incorporated administrative and learning management system data to predict at-risk students (those who are not in good standing) at a small private college, and then tested the system at four other less-selective colleges.

While the recent decade has seen a steady growth in prediction-focused studies on college dropout, a large proportion of them are focused on individual courses or a small sample of degree programs [21]. Most of them investigate dropouts at brick-and-mortar institutions. Our study pushes these research boundaries by examining dropout prediction for multiple cohorts of students across residential and exclusively online degree programs offered by a large public university. The breath of our sample is rare in the dropout prediction literature and promises to offer more generalizable insights about the utility and feasibility of predictive models.

Algorithmic Fairness in Education

A central goal of educational research and practice has been to close opportunity and achievement gaps between different groups of students. More recently, algorithmic fairness has become a topic of interest as an increasing number of students are exposed to intelligent educational technologies [26]. Inaccuracies in models might translate into severe consequences for individual students, such as failing to allocate remedial resources to struggling learners. It is more concerning if such inaccuracies disproportionately fall upon students from disadvantaged backgrounds and worsen existing inequalities. In

this context, the fairness of algorithmic systems is generally evaluated with respect to protected attributes following legal terms. The specific criteria of fairness, however, vary and largely depend on the specific application(s) [39].

In the past few years, a handful of papers have brought the fairness framework to real-world learning analytics research. Most of these studies audit whether supervised learning models trained on the entire student population generate systematically biased predictions of individual outcomes such as correct answers, test scores, course grades, and graduation [41, 29, 20, 24, 16, 31]. For example, Yu and colleagues [41] found that models using college-entry characteristics to predict course grades and GPA tend to predict lower values for underrepresented student groups than their counterparts. Other studies have examined biases encoded in unsupervised representations of student writing [2], or go further to refine algorithms for at-risk student identification under fairness constraints [22]. Overall, this area of research is nascent and in need of systematic frameworks specific to educational contexts to map an agenda for future research.

When it comes to strategies to improve algorithmic fairness, a contentious point is whether protected attributes should be included as predictors (features) in prediction models. Most training data from the real world are the result of historical prejudices against certain protected groups, so directly using group indicators to predict outcomes risks imposing unfair stereotypes and reproduce existing inequalities [4]. In educational settings, it may be considered unethical to label students from certain groups as "at risk" from day one, when in fact, these students have demonstrated an exceptional ability to overcome historical obstacles and might therefore be more likely to succeed [37]. This concern motivated the research effort to "blind" prediction models by simply removing protected attributes (i.e. fairness through unawareness) or more complicated statistical techniques to disentangle signals of protected attributes from other features due to their inherent correlation [8]. In contrast, recent work has advocated for explicitly using protected attributes in predictive models (i.e. fairness through awareness) [17]. In particular, Kleinberg and colleagues [27] showed in a synthetic example of college admission that the inclusion of race as a predictor of college success improves the fairness of admission decisions without sacrificing efficiency. Given the well-documented relationship between student background and their educational outcomes, a recent review also suggests that predictive models in education should include demographic variables to ensure that algorithms are value-aligned, i.e., all students have their needs met [34].

To our knowledge, however, there is only limited empirical evidence to support either side of this debate. Our study therefore presents an in-depth examination of the consequences of including or excluding protected attributes on algorithmic fairness of a realistic, large-scale dropout prediction model.

METHODOLOGY

Dataset
We analyze de-identified institutional records from one of the largest public universities in the United States. This broad-access research university serves nearly 150,000 students with an 86% acceptance rate and 67% graduation rate. Its student population is representative of the state in which it is located, which makes it a Hispanic-serving institution (HSI). The university has offered many of the same undergraduate degree programs fully online to over 40,000 students. The dataset we use in this study focuses on undergraduate students and contains student-level characteristics and student-course-level records for their first term of enrollment at the university, including transfer students (except for those who transfer into their senior year). For our prediction task, we only keep students whose first term was in the Fall along with their course-taking records in their first term, including terms between 2012-18 (residential) and 2014-18 (online).

This sample comprises a total of 564,104 residential course-taking records for 93,457 unique students and 2,877 unique courses, and 81,858 online course-taking records for 24,198 unique students and 874 unique courses. The course-taking records include both a student's letter grade and course-level metadata (subject, course number, units, required for major, etc.). Student-level information includes socio-demographic information (age, gender, race/ethnicity, first-generation status, etc.), prior academic achievement (high school GPA, standardized test scores), enrollment information (transfer student status, part-time status, academic major and minor, etc.). These data are representative of what most higher education institutions routinely manage in their student information systems (SIS) [3].

Prediction Target and Feature Engineering
The primary goal of a dropout prediction model is to alert relevant stakeholders to currently enrolled students who are at risk of dropping out of a degree program so that they can reach out and offer support at an early stage. While the general framework of dropout prediction is well established, the exact definition of dropout, or attrition, varies based on the specific context [33]. In our context, we define dropout as not returning to school a year from the first time of enrollment. We only analyze students who first enrolled in Fall, so dropout means not returning in the following Fall. This final operationalization aligns well with retention, one of the two standard metrics of post-secondary student success in national reports of the United States [38, 23].[1]

We use students' background characteristics and academic records in the first enrolled term (Fall) to predict dropout, because it would be beneficial to identify risks as early as possible and institutional records are usually updated and available at the end of each term. Informed by existing research in higher education and learning analytics (see Related Work),

---

[1]The other standard metric is graduation within 100% or 150% of the normative time (i.e. 4 or 6 years for four-year institutions). We do not examine this metric because the span of our dataset is only six years and we do not observe graduation outcomes for all student cohorts.

Table 1. Features used for dropout prediction.

Category	Features
Protected attributes	Gender (binary), first-generation college student (binary), underrepresented minority (URM; binary; defined as not Asian or White), high financial need (binary; FASFA-based expected family contribution under $5,500)
Incoming attributes	Age, high school GPA, math and verbal SAT/ACT scores, transfer student (binary), transferred credits, transfer GPA
Program information	Part-time student (binary), major, minor, STEM major (binary)
Course performance	Total courses enrolled, total units enrolled, percentage of courses that are required, credits received from different types of courses (lecture, seminar, etc.), levels of courses (100, 200, etc.), term GPA, mean and variance in course grades within each session during the term, percentage distribution of letter grades

we construct 58 features from the dataset for both residential and online students. Table 1 summarizes these feature by four categories. We include four protected attributes, which are the most commonly used dimensions along which to examine educational inequalities and set equity goals in policy contexts [9, 10, 23].

Table 2 depicts the student profile in our analysis. The statistics reaffirm that, regardless of format, the institution serves a large proportion of students from historically disadvantaged groups. There are also major differences across formats. In line with the national statistics of exclusively online programs [38], the online sample has a higher concentration of transfer and non-traditional (older, part-time) students, and also higher dropout rates compared to residential students. These characteristics validate that the current analysis is performed on student populations who are most in need of institutional support and allow us to scrutinize the generalizability of our findings across two distinct contexts of higher education.

Table 2. Comparison of online and residential student populations.

	Online	Residential
N	24,198	93,457
Dropout	40.7%	16.9%
Female	60.9%	47.9%
First-gen	42.4%	33.6%
URM	33.1%	34.6%
High need	61.9%	51.3%
Transfer	85.2%	31.8%
Part-time	77.2%	12.9%
Average age	27.1	19.7

Dropout Prediction

To investigate the consequences of using protected attributes in dropout prediction models, we generate two feature sets: the AWARE set includes all features shown in Table 1, while the BLIND set excludes the four protected attributes from the AWARE set. For convenience, we will refer to a specific model by the feature set it uses in the remainder of this paper. Given our binary target variable, the dropout prediction task is formalized as a binary classification problem. As we focus on identifying the effect of including protected attributes, we experiment with two commonly used algorithms – logistic regression (LR) and gradient boosted trees (GBT). We choose

LR because it is a linear additive and highly interpretable classifier that can achieve reasonable prediction performance with well-chosen features. The choice of GBT, on the other hand, is for its ability to accommodate a large number of features, efficiently handle missing values, and automatically capture non-linear interactions between features.

We predict dropping out separately for online and residential students. For each format, we split the data into a training set and a test set based on student cohort: the last observed cohort (6,939 online and 14,275 residential students entering in Fall 2018) constitutes the test set and the remaining cohorts make up the training set (17,259 online and 79,182 residential students). There are two reasons for doing the train-test split by student cohorts. Practically, this split aligns with the real-world application where stakeholders rely on historical data to make predictions for current students [25]. Technically, this approach alleviates the issue of data contamination between the training and test set [19], as the features we use, especially the first-semester records, might be highly correlated within the same cohort but much less so across cohorts.

There are a few additional technical details about model training. First, we tune hyperparameters of the two algorithms by performing grid search over a specified search space and evaluating the hyperparameters using 5-fold cross-validation. Second, we add indicator variables for missing values in course grades, standardized test scores, and academic majors and minors. Third, we apply robust scaling to training features to regulate the influence of outliers. Fourth, because the class imbalance in both datasets can bias the model learning towards the majority class (i.e. non-dropout), we adjust the sample weights to be inversely proportional to class frequencies during the training stage.

The trained classifiers are then applied to the test set to evaluate the performance. The immediate output of each classifier is a predicted probability of dropping out for each student. To make a final binary prediction of dropout, we use dropout rates in the training data to determine the decision thresholds for the test set, such that the proportion of predicted dropouts in the test set matches the proportion of observed dropouts in the training set [6]. Compared to the default of 0.5, this choice of threshold is more reasonable when we rely on the observed history to predict the unknown future in practice.

Table 3. Overall prediction performance of AWARE and BLIND models trained with gradient boosted trees (GBT) and logistic regression (LR).

	GBT			LR		
Metric	AWARE	BLIND	Δ	AWARE	BLIND	Δ
Online (Non-dropout: 59.3%)						
Accuracy	75.8	75.6	0.2	75.2	75.4	-0.2
Recall	67.3	67.1	0.2	66.7	66.8	-0.1
TNR	82.4	82.3	0.1	81.9	82.0	-0.1
Residential (Non-dropout: 83.1%)						
Accuracy	83.9	83.9	0.0	83.6	83.6	0.0
Recall	54.1	54.1	0.0	53.2	53.3	-0.1
TNR	89.1	89.1	0.0	88.9	88.9	0.0

Note: none of the Δ values is statistically significant with $p < 0.1$.

Performance Evaluation

We evaluate prediction performance based on three metrics: accuracy, recall, and true negative rate (TNR). In the context of dropout prediction, recall is the proportion of actual dropouts who are correctly identified, whereas TNR quantifies how likely a student who persists into the second year of college is predicted to persist. To examine the effects of including protected attributes on overall performance, we compute these metrics separately for each model and test whether each metric significantly changes from BLIND to AWARE models, using two proportion z-tests.

We operationalize fairness as the independence between prediction performance, measured by the three metrics above, and protected group membership. This definition of fairness with respect to the three metrics corresponds to the established notions of overall accuracy equality, equal opportunity, and predictive equality, respectively [26]. Specifically, to quantify the fairness of a given model with regard to a binary protected attribute, such as URM, we compute the differences in each of the three metrics between the two associated protected groups, URM and non-URM students. We then compare how much these differences change between BLIND and AWARE models in order to quantify the effect of including protected attributes as predictors on fairness.

RESULTS

Overall Prediction Performance

We first illustrate the effects of including protected attributes on overall prediction performance. Table 3 reports the overall performance of AWARE and BLIND models, trained with GBT and LR algorithms, on the test dataset. The last column under each algorithm reports the percentage point differences in performance between the two models (from BLIND to AWARE). The main finding is that including or excluding protected attributes does affect the performance of the dropout prediction in either context. None of the performance metrics (accuracy, recall, TNR) differs significantly between the BLIND and AWARE models. Additionally, while the more sophisticated GBT algorithm performs better than the simple LR on all metrics, the advantage is comparatively small (less than one percentage point on all metrics). Because of this, we restrict the following analysis to GBT-based models.

Compared to a naïve baseline which simply predicts every student to be the majority class (non-dropout) and achieves an accuracy equal to that majority's share, the predictive models can accurately predict online dropouts with a decent margin. However, the accuracy margin for predicting residential dropouts is fairly small. The other two metrics, which describe the accuracy among dropouts and non-dropouts respectively, achieve a higher value when the corresponding group has a larger share and vice versa. Specifically, the models are able to identify 67.3% of online dropouts and 54.1% of residential dropouts. This latter value is somewhat lower but still comparable to the recall performance in recent prior work on dropout prediction in residential programs [6, 15].

To take a closer look at the model predictions, beyond the three aggregate performance metrics, we examine whether including protected attributes alters the distribution of predicted dropout probabilities. As shown in Figure 1, the distributions are highly similar across the models which further validates the limited marginal impact of protected attributes. An additional insight from these plots is that dropouts might be much more heterogeneous than non-dropouts in terms of the features in Table 1, as their predicted probabilities are highly spread out, especially in residential settings where the majority of dropouts are assigned a small dropout probability. This pattern is consistent with the lower recall performance shown in Table 3.

This finding appears to conflict with prior research that demonstrates the critical role of demographic and background characteristics for student success in higher education [28]. In an effort to better understand our result, we explore two mutually compatible hypotheses inspired by the algorithmic fairness literature. One hypothesis is that dropping out, the prediction target, is not sufficiently correlated with protected attributes, and thus adding the latter to a dropout prediction model would not improve performance much. To test this, we fit separately for each enrollment format in the test data a logistic regression model that predicts dropout using all the possible interaction terms between the four protected attributes. We find that, even though a few coefficients are statistically significant, the adjusted McFadden's R^2 is as small as 0.006 for either format, lending support to our hypothesis.

The second hypothesis is that protected attributes are already implicitly encoded in the BLIND feature set, and adding them directly does not add much predictive power. We test this by fitting four logistic regressions for each format which use the BLIND feature set to predict each of the four protected attributes. Based on the adjusted McFadden's R^2, we find that only gender can plausibly be considered encoded in the other features (0.159 for online and 0.187 for residential). This lends partial support to our second hypothesis.

Fairness of Prediction

We further examine how the inclusion of protected attributes might affect the fairness of dropout predictions. As mentioned in the previous section, for each of the four protected attributes, we first measure fairness by the group difference in a chosen performance metric. For example, a prediction model that achieves the same accuracy on male and female students is considered fair in terms of accuracy (0% difference). Following this construction, Figure 2 visualizes these
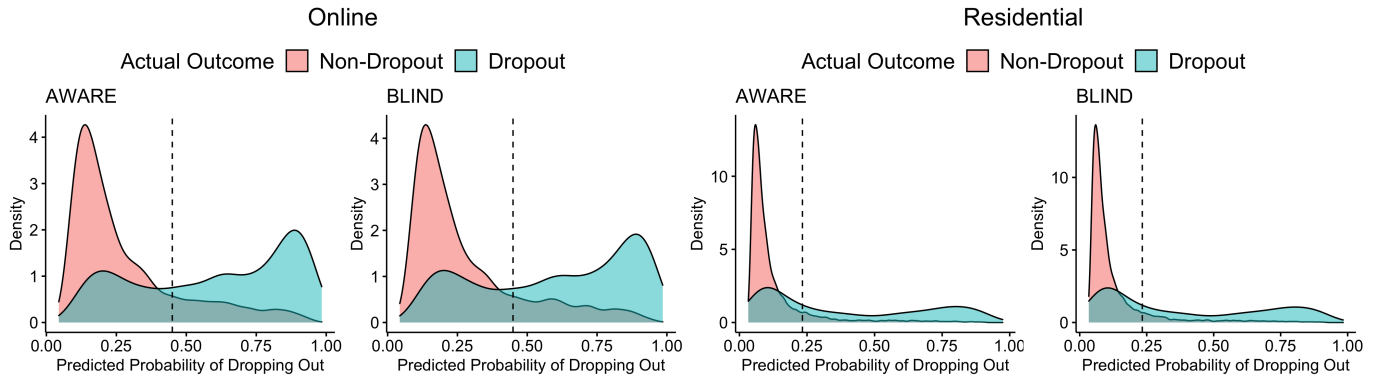
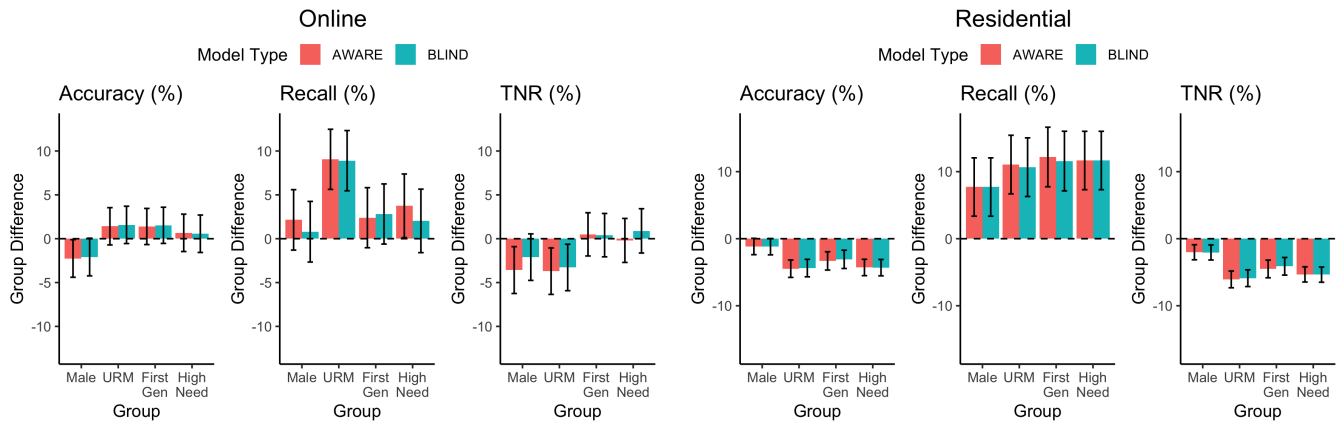**Figure 1. Distribution of predicted dropout probability**

Figure 2. Fairness of AWARE and BLIND models in terms of accuracy (left), recall (middle), and TNR (right). Positive group differences (y-axis) indicate higher values for the listed groups compared to their corresponding reference groups. Group differences closer to zero reflect higher levels of fairness. Error bars indicate 95% confidence intervals.

fairness results of the AWARE and BLIND models for each of the four protected attributes in terms of the three metrics. Each bar in a subplot depicts the difference in that metric between the labeled group and their counterpart (e.g., male - female). The closer the bar is to zero, the fairer that model prediction is. Overall, the figure shows that both the AWARE and BLIND models are unfair for some protected attributes and some metrics, but fair for others. This lack of universal fairness is expected given the many dimensions of protected attributes, models, and metrics. However, for residential students, the model consistently exhibits unfairness across all protected attributes and metrics, especially in terms of recall. The inclusion or exclusion of protected attributes does not in general lead to different levels of fairness in terms of any metric in any enrollment format, as all adjacent error bars in the figure exhibit a high degree of overlap.

While the aggregated group fairness metrics do not differ with vs. without protected attributes, we take a step further to explore how individual-level changes in model predictions can shed light on the overall change in fairness. We examine changes in the individual ranking of predicted dropout probability among all predicted students (test set) from BLIND to AWARE model. Figure 3 plots the distribution of this ranking change for each protected group, where higher values

represent moving up in the assigned risk leaderboard when protected attributes are included for prediction.

We find that overall the ranking change is centered around zero, but there are observable group differences in certain cases. In the online setting, the AWARE model tends to move up females and students without a high financial need on the dropout risk leaderboard simply based on their identity. Similarly, continuing-generation college students are moved up more in residential settings compared to their first-generation counterparts. We argue that these group differences suggest improved fairness if the group going up more in the ranking spectrum has lower dropout rates in reality, and vice versa. To formally evaluate this reasoning, we conduct a series of t-tests between pairs of protected groups on their ranking change. We also compute Cohen's d to gauge the standardized effect size. Comparing Table 5 which describes these results and Table 4 which presents the actual dropout rates of each group, we find that moving from BLIND to AWARE causes students from advantaged (lower dropout rates) groups to be assigned relatively higher risk rankings compared to their disadvantaged (higher dropout rates) reference groups, and that this effect size is larger when the two paired groups have larger gaps in dropout rates. Thus, adding protected attributes to the model is working against existing inequities to a marginal extent instead of reinforcing them.
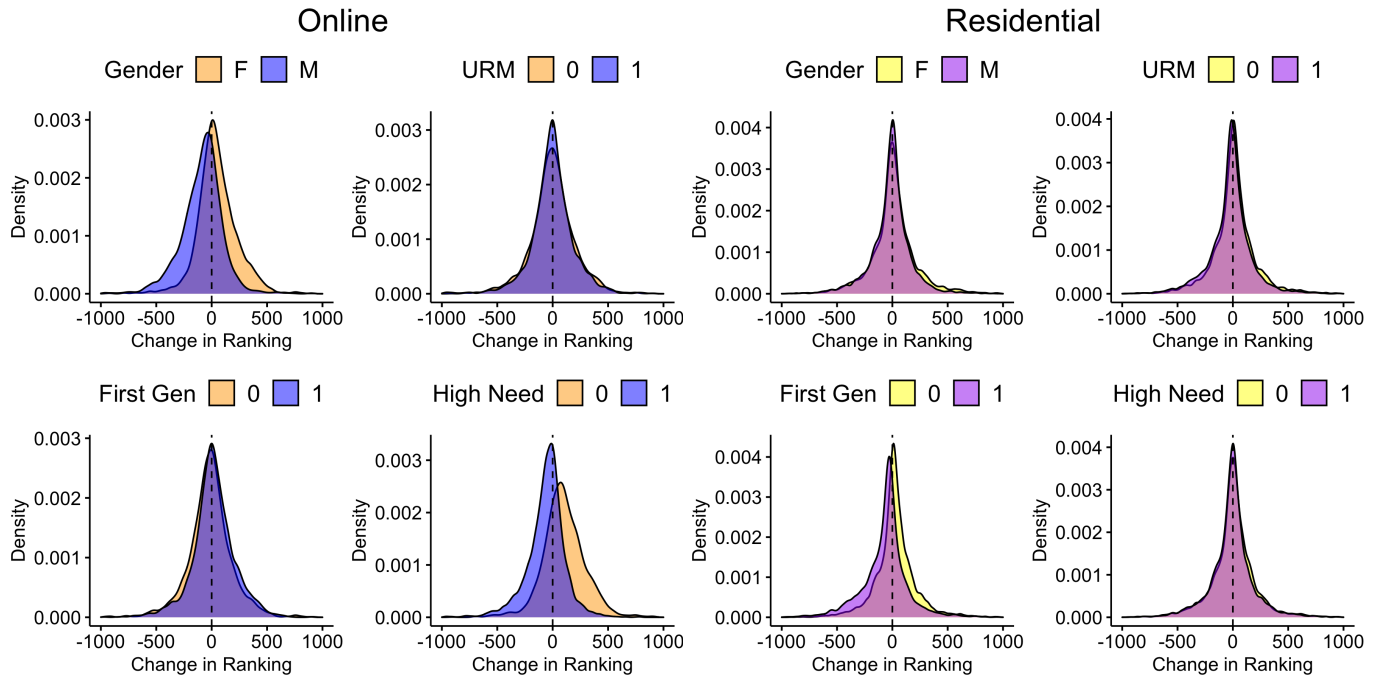
Figure 3. Distribution of change in individual ranking of predicted dropout probability from AWARE to BLIND. One unit increase means going up by one place in the AWARE model compared to the BLIND model.

Table 4. Dropout rates among different protected groups in the test set.

	Dropout rate	
	Online	Residential
Overall	40.7	16.9
Male	49.5	15.6
Female	40.7	14.0
URM	46.6	16.8
Non-URM	42.4	13.7
First-gen	43.7	17.9
Continuing-gen	44.1	13.5
High need	45.8	17.0
Low need	40.5	12.8

Table 5. Welch two-sample t-test results and Cohen's d effect size of individual ranking change. Positive ranking change means increased predicted dropout risks from BLIND to AWARE model.

Group (Avg. ranking change)		Rank Δ	Cohen's d
Online			
Female (50.5)	Male (-88.4)	138.9***	0.71
Non-URM (0.5)	URM (-0.9)	1.4	0.01
Continuing-gen (-13.9)	First-gen (18.6)	-32.5***	0.16
Low need (104.5)	High need (-60.0)	164.5***	0.87
Residential			
Female (15.9)	Male (-15.1)	31.0***	0.13
Non-URM (13.0)	URM (-22.8)	35.8***	0.15
Continuing-gen (27.2)	First-gen (-62.1)	89.3***	0.38
Low need (6.8)	High need (-7.0)	13.8***	0.06

Significance levels: *** $p<0.001$; ** $p<0.005$; * $p<0.01$

DISCUSSION AND CONCLUSION

We set out to answer a simple question: Should protected attributes be included in college dropout prediction models? This study offers a comprehensive empirical examination of how the inclusion of protected attributes affects the overall performance and fairness of a realistic predictive model. We demonstrate this finding across two large samples of residential and online undergraduate students enrolled at one of the largest public universities in the United States. Our findings show that including four important protected attributes (gender, URM, first-generation student, high financial need) does not have any significant effect on three common measures of overall prediction performance when commonly used features (incoming attributes, enrollment information, academic records) are already in the model. Even when used alone without those features, the group indicators defined by the

protected attributes are not highly predictive of dropout, although the actual dropout rates are somewhat higher among minoritized groups. In terms of fairness, we find that including protected attributes only leads to a marginal improvement in fairness by assigning dropout risk scores with smaller gaps between minority and majority groups. However, this trend is not sufficiently large to systematically change the final dropout predictions based on the risk scores, and therefore the formal fairness measures are not significantly different between models with and without protected attributes.

In short, our results suggest limited effects of including protected attributes on the performance of college dropout prediction. This does not point to a clear answer to our normative question and prompts us to further reflect on the focal issue of using protected attributes. Recent work in the broader machine learning community has been in favor of "fairness

through awareness" [17], and has specifically suggested that race-aware models are fairer for student success prediction because they allow the influence of certain features to differ across racial groups [27]. Our findings resonate with these existing studies around fairness but only to a marginal extent. Notably, student groups with historically higher dropout rates are slightly compensated by being ranked lower in predicted dropout risks when protected attributes are used. This compensating effect, however, does not accumulate to statistically significant changes in predicted labels, possibly because the group differences in dropout rates were not sizeable in the past at the institution we study (see Table 4). In other words, protected attributes might have more to contribute to the fairness of prediction in the presence of substantial existing inequalities. Still, the existence of a weak compensating instead of segregating effect justifies the inclusion of these attributes. After all, a major argument for race-aware models, and more generally socio-demographic-aware models, is to capture structural inequalities in society that disproportionately expose members of minoritized groups to more adverse conditions. In addition, the deliberate exclusion of protected attributes from dropout prediction models can be construed as subscribing to a "colorblind" ideology, which has been criticized as a racist approach that serves to maintain the status quo [7].

Another contribution of this work lies in our approach to fairness evaluation. The analyses and visualizations we present are the result of many iterations to arrive at simple yet compelling ways to communicate fairness at different levels of aggregation and across many protected attributes. These methods can be used by those who seek to evaluate model fairness for research and practice. Prior research has mostly focused on evaluating one protected attribute at a time, but in most real-world applications we care about more than one protected attribute. We recommend comparing AWARE against BLIND models in terms of the individual ranking differences by group (Figure 3) as well as the group difference plots for multiple performance metrics and protected attributes (Figure 2). This approach offers a sensitive instrument for diagnosing fairness-related issues in various domains of application, which could easily be implemented in a fairness dashboard that evaluates multiple protected attributes, models, and performance metrics [40]. This will remain a promising line of our future work.

This research has broader implications for using predictive analytics in higher education beyond its contributions to algorithmic fairness. With a common set of institutional features, we achieve 76% prediction accuracy and 67% recall on unseen students in online settings, that is, correctly identifying 67% of actual dropouts with their first-term records. For residential students, we achieve a higher accuracy of 84% but a lower recall of 54%. These performance metrics may seem somewhat lower than in prior studies of dropout prediction, but this might be because most existing studies examine a smaller sample of more homogeneous students, such as students in the same cohort or program [15, 14]. This highlights the general challenge of predicting college dropout accurately. As suggested by the large variance in predicted probabilities for

dropouts (Figure 1), widely used institutional features might not perform well in capturing common signals of dropout. This may point to important contextual factors that our institutional practices are presently overlooking. We view this as a limitation and important next step that will require both an interrogation of the theoretical basis for predictors and close collaboration with practitioners.

Further directions for future research in this area include exploring counterfactual notions of fairness in this context by testing how predictions would differ for counterfactual protected attributes, all else being equal. This would benefit the contemporary education system which relies increasingly on research that provides causal evidence. We would also like to move from auditing to problem-solving by evaluating correction methods for any pre-existing unfairness in predictions to see how the AWARE relative to the BLIND model responds [30]. We hope that this study inspires more researchers in the learning analytics and educational data mining communities to engage with issues of algorithmic bias and fairness in the models and systems they develop and evaluate.

REFERENCES

[1] 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 & Knowledge (LAK '12)*. 267–270. DOI: `http://dx.doi.org/10.1145/2330601.2330666`

[2] Noah Arthurs and AJ Alvero. 2020. Whose Truth is the "Ground Truth"? College Admissions Essays and Bias in Word Vector Evaluation Methods. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. 342–349.

[3] 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.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. `http://www.fairmlbook.org`

[5] 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. DOI: `http://dx.doi.org/10.1007/s11162-019-09546-y`

[6] Johannes Berens, Kerstin Schneider, Simon Görtz, 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. DOI:`http://dx.doi.org/10.5281/ZENODO.3594771`

[7] Meghan Burke. 2018. *Colorblind racism*. John Wiley & Sons.

[8] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*. 3992–4001.

[9] Emily Forrest Cataldi, Christopher T. Bennett, and Xianglei Chen. 2018. *First-Generation Students: College Access, Persistence, and Postbachelor's Outcomes (NCES 2018421)*. Technical Report. National Center for Education Statistics.

[10] Xianglei Chen and Annaliza Nunnery. 2019. *Profile of Very Low and Low-Income Undergraduates in 2015–16*. Technical Report. National Center for Education Statistics.

[11] James S. Coleman. 1988. Social capital in the creation of human capital. *Amer. J. Sociology* 94 (1988), S95–S120.

[12] Shane Dawson, Jelena Jovanovic, Dragan Gašević, and Abelardo Pardo. 2017. From Prediction to Impact: Evaluation of a Learning Analytics Retention Program. In *Proceedings of the 7th International Conference on Learning Analytics & Knowledge*. 474–478. DOI: http://dx.doi.org/10.1145/3027385.3027405

[13] Cristobal de Brey, Lauren Musu, Joel McFarland, Sidney Wilkinson-Flicker, Melissa Diliberti, Anlan Zhang, Claire Branstetter, and Xiaolei Wang. 2019. *Status and Trends in the Education of Racial and Ethnic Groups 2018 (NCES 2019-038)*. Technical Report. National Center for Education Statistics.

[14] 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)*. 41–50.

[15] Francesca Del Bonifro, Maurizio Gabbrielli, Giuseppe Lisanti, and Stefano Pio Zingaro. 2020. Student Dropout Prediction. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education (AIED 2020)*. Springer, 129–140. DOI: http://dx.doi.org/10.1007/978-3-030-52237-7_11

[16] Shayan Doroudi and Emma Brunskill. 2019. Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)*. ACM, 335–339. DOI: http://dx.doi.org/10.1145/3303772.3303838

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226. DOI:http://dx.doi.org/10.1145/2090236.2090255

[18] Manuela Ekowo and Iris Palmer. 2016. *The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis*. Technical Report. New America. 1–31 pages.

[19] Elaine Farrow, Johanna Moore, and Dragan Gašević. 2019. Analysing discussion forum data: a replication study avoiding data contamination. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. 170–179.

[20] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)*. ACM Press, 225–234. DOI: http://dx.doi.org/10.1145/3303772.3303791

[21] Arto Hellas, Petri Ihantola, 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 (ITiCSE 2018 Companion)*. Association for Computing Machinery, 175–199. DOI: http://dx.doi.org/10.1145/3293881.3295783

[22] Qian Hu and Huzefa Rangwala. 2020. Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. 431–437.

[23] Bill Hussar, Jijun Zhang, Sarah Hein, Ke Wang, Ashley Roberts, Jiashan Cui, Mary Smith, Farrah Bullock Mann, Amy Barmer, Rita Dilig, and Nachazel. 2020. *The Condition of Education 2020 (NCES 2020-144)*. Technical Report. National Center for Education Statistics.

[24] Stephen Hutt, Margo Gardner, Angela L. Duckworth, and Sidney K. D'Mello. 2019. Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*.

[25] Sandeep M. Jayaprakash, Erik W. Moody, Eitel J.M. Lauría, 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. DOI: http://dx.doi.org/10.18608/jla.2014.11.3

[26] René F Kizilcec and Hansol Lee. 2020. Algorithmic Fairness in Education. *arXiv preprint arXiv:2007.05443* (2020).

[27] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic Fairness. *AEA Papers and Proceedings* 108 (2018), 22–27. DOI: http://dx.doi.org/10.1257/pandp.20181018

[28] George D. Kuh, Jillian Kinzie, Jennifer A. Buckley, Brian K. Bridges, and John C. Hayek. 2007. Piecing Together the Student success puzzle: Research, Propositions, and Recommendations. *ASHE Higher Education Report* 32, 5 (2007), 1–182. DOI: http://dx.doi.org/10.1002/aehe.3205

[29] Catherine Kung and Renzhe Yu. 2020. Interpretable Models Do Not Compromise Accuracy or Fairness in Predicting College Success. In *Proceedings of the 7th ACM Conference on Learning @ Scale (L@S '20)*. Association for Computing Machinery (ACM), New York, NY, USA, 413–416. DOI: `http://dx.doi.org/10.1145/3386527.3406755`

[30] Hansol Lee and René F Kizilcec. 2020. Evaluation of Fairness Trade-offs in Predicting Student Success. *arXiv preprint arXiv:2007.00088* (2020).

[31] Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Florence, Italy, 1–10. DOI: `http://dx.doi.org/10.18653/v1/W19-4401`

[32] The Chronicle of Higher Education. 2020. *The Post-Pandemic College*. Technical Report.

[33] Timothy J. Pantages and Carol F. Creedon. 1978. Studies of College Attrition: 1950—1975. *Review of Educational Research* 48, 1 (1978), 49–101. DOI: `http://dx.doi.org/10.3102/00346543048001049`

[34] Luc Paquette, Ziyue Li, Ryan Baker, Jaclyn Ocumpaugh, and Alexandra Andres. 2020. Who's learning? Using demographics in EDM research. *Journal of Educational Data Mining* 12, 3 (2020), 1–30. DOI: `http://dx.doi.org/10.5281/ZENODO.4143612`

[35] Julia E. Seaman, I. Elaine Allen, and Jeff Seaman. 2018. *Grade Increase: Tracking Distance Education in the United States*. Technical Report. 1–45 pages.

[36] Neil Selwyn and Dragan Gašević. 2020. The datafication of higher education: discussing the promises and problems. *Teaching in Higher Education* 25, 4 (2020), 527–540. DOI: `http://dx.doi.org/10.1080/13562517.2019.1689388`

[37] Simon Buckingham Shum. 2020. Should predictive models of student outcome be "colour-blind"? `http://simon.buckinghamshum.net/2020/07/should-predictive-models-of-student-outcome-be-colour-blind`. (2020).

[38] Thomas D Snyder, Cristobal de Brey, and Sally A Dillow. 2019. *Digest of Education Statistics 2018 (NCES 2020-009)*. Technical Report. National Center for Education Statistics.

[39] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[40] Kimberly Williamson and René F. Kizilcec. 2021. Learning Analytics Dashboard Research Has Neglected Diversity, Equity and Inclusion. In *Proceedings of the 8th ACM Conference on Learning @ Scale*.

[41] Renzhe Yu, Qiujie Li, Christian Fischer, Shayan Doroudi, and Di Xu. 2020. Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. 292–301.

A Dropout Prediction Framework Combined with Ensemble Feature Selection

Authors:

[Dan Ai](#) • School of Computer Science And Engineering, Northeastern University, Shenyang, China

[Tiancheng Zhang](#) • School of Computer Science And Engineering, Northeastern University, Shenyang, China

[Ge Yu](#) • School of Computer Science And Engineering, Northeastern University, Shenyang, China

[Xinying Shao](#) • School of Computer Science And Engineering, Northeastern University, Shenyang, China

View in Digital library

Publication:

In recent years, with the rapid development of large-scale open online courses, low completion rate and high dropout rate have been important challenges for open online courses. Therefore, it is necessary to make effective prediction and timely intervention to ensure the completion of the course. Some of the traditional prediction models only use the features extracted manually from students' clickstream data, which is too subjective to guarantee the quality of features and affect the prediction accuracy. Others generate features automatically with finer granularity, but the problem of feature redundancy appears. In order to solve this problem, this paper proposes a comprehensive dropout prediction framework of MOOCs students. The framework can automatically extract features from clickstream data, and filter features with an integrated feature selection strategy based on clustering and weighted MaxDiff, and finally predict. Experiments show that the model can effectively improve the accuracy of prediction of dropout.

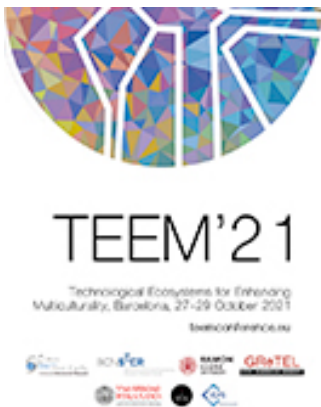# Data Analytics for Predicting Dropout

Authors:

Gioconda Riofrío-Calderón • Universidad Técnica Particular de Loja, Ecuador

María-Soledad Ramírez-Montoya • Tecnológico de Monterrey, Mexico

María-José Rodríguez-Conde • Universidad de Salamanca, Spain
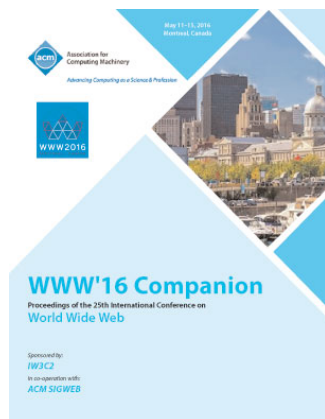
View in Digital library

No abstract available.

Correlational Analysis Between School Performance and Municipal Indicators in Brazil Supported by Linked Open Data

Author:

Bruno Elias Penteado • University of São Paulo, São Carlos, Brazil

View in Digital library

Publication:

Proceeding

The advance in quality of public education is a challenge to public managers in contemporary society. In this sense, many studies point to the strong influence of socioeconomical factors in school performance but it is a challenge to select proper data to perform analyses on this matter. In tandem, it has happening a growth in provision of big quantities of educational indicators data, but in isolate cases, and by different agencies of Brazilian government. For this work, we use both education and economic indicators for analysis. The following socioeconomical indicators were selected: municipal human development index (MHDI), social vulnerability index (SVI), Gini coefficient and variables extracted from DBpedia, as part of the connection of this data to the Web of data: GDP per capita and municipal population. These data were used as independent variables to look into their correlations with Brazilian Basic Education Development Index (IDEB) performances at municipal level, supported by the application of linked open data principles. OpenRefine was used to extract the data from different sources, convert to RDF triples and then the mapping of the variables to existing ontologies and vocabularies in this domain, aiming at the reuse of existing semantics. The correlational analysis of the variables showed coherence with the literature about the theme, with significative magnitude between IDEB performances and the indicators related to income and parent education (SVI and HDI), besides moderate relations with the other varibles, except for the municipal population. Finally, the consolidated dataset, enriched by information extracted DBpedia was made available by a SPARQL endpoint for queries of humans and software agents, allowing other applications and researchers to explore the data from other platforms.