Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data

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

Qiujie Li
New York University
New York, NY, USA
ql16@nyu.edu

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

Shayan Doroudi
University of California, Irvine
Irvine, CA, USA
doroudis@uci.edu

Di Xu
University of California, Irvine
Irvine, CA, USA
dix3@uci.edu

ABSTRACT

In higher education, predictive analytics can provide actionable insights to diverse stakeholders such as administrators, instructors, and students. Separate feature sets are typically used for different prediction tasks, e.g., student activity logs for predicting in-course performance and registrar data for predicting long-term college success. However, little is known about the overall utility of different data sources across prediction tasks and the fairness of their predictions with respect to different subpopulations. Using data from over 2,000 college students at a large public university, we examined the utility of institutional data, learning management system (LMS) data, and survey data for accurately and fairly predicting short-term and long-term student success. We found that institutional data and LMS data both have decent predictive power, but survey data shows very little predictive utility. Combining institutional data with LMS data leads to even higher accuracy than using either alone. In terms of fairness, using institutional data consistently underestimates historically disadvantaged student subpopulations more than their peers, whereas LMS data tend to overestimate some of these groups more often. Combining the two data sources does not fully neutralize the biases and still leads to high rates of underestimation among disadvantaged groups. Moreover, algorithmic biases affect not only demographic minorities but also students with acquired disadvantages. These analyses serve to inform more cost-effective and equitable use of student data for predictive analytics applications in higher education.

Keywords

Predictive analytics; Machine Learning; Higher education; Fairness; Student data

Renzhe Yu, Qiujie Li, Christian Fischer, Shayan Doroudi and Di Xu "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)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 292 - 301

1. INTRODUCTION

The most common application of learning analytics in higher education is using predictive modeling to understand critical factors contributing to student success, or to identify students who need support in a timely manner. Predictive analytics have been used within a course [2] or while using tutoring software [38]. They have also been used to optimize student success in the longer term, for example to predict graduation rates [3] or to make course recommendations [26]. Different data sources can be used to build these predictive models, with varying trade-offs. For example, when making predictions at the course level, log data from learning management systems (LMS) are often used. These systems allow for automated and scalable recording of hundreds of learner actions in every single minute, but they require robust and efficient data management systems. When making longer-term predictions, on the other hand, institutions can use data typically stored in student information systems (SIS), including prior academic history, standardized test scores, and demographic information. While this data source might be readily available to college administrators, it might be more difficult to access, due to ethical concerns or logistic barriers, for individual instructors or researchers trying to build such models for particular use cases. In some cases, both data sources are further combined with assessments or surveys that measure students' metacognitive abilities or other non-cognitive attributes that might predict college success [35]. However, collecting and managing these data is often costly for institutions if they are not already doing so. Given all these trade-offs, it is necessary to examine the utility of different student data sources for building predictive analytics-based solutions to guide instructors, administrators and education policy makers on the costs and benefits of utilizing different data sources.

To date, research that systematically compares data sources and predictions is underrepresented in the literature [14]. To respond to this call for research, this study evaluates the usefulness of three common student data sources for two representative prediction tasks. These three data sources, including institutional data, LMS data, and survey data, are all widely used across research settings and have been shown to predict various measures of college success. Given the different use cases of short-term and long-term predic-

tions as discussed above, we construct two success measures: individual course grades (short-term success) and yearly average GPA (long-term success). The usefulness of each data source is determined by its contribution to overall prediction accuracy and to prediction fairness across student subpopulations. The focus on fairness arises from the concern that predictive models trained on the entire student population may perform systematically worse on selected subpopulations than other others, which may have unintended negative effects for vulnerable students [6]. For instance, if models are less confident in identifying struggling students among an already underrepresented group, this bias may eventually amplify existing achievement gaps.

In short, our research aims to identify what combinations of student data (a) more accurately predict different success measures; and (b) more fairly predict these measures. The remainder of this paper is organized as follows: Section 2 summarizes the related work on college success prediction and fairness of predictive models; Section 3 describes the data and methods we use to construct and evaluate prediction models; Section 4 presents the results from various predictions; Section 5 reflects on the findings and discusses the practical implications for stakeholders; Section 6 concludes the study with limitations and future work.

2. RELATED WORK

2.1 Predicting College Success Using Student Characteristics

Although college is a complicated ecosystem with numerous factors shaping student outcomes, prior research has identified several groups of student characteristics across institutional data, LMS data, and survey data that consistently predict commonly used measures of success.

2.1.1 Personal Background - Institutional Data

Student success in higher education is often stratified by students' demographic, socioeconomic and academic background prior to college experience. For example, college graduation rates substantially differ by students' race/ethnicity. National data indicates that Hispanic students are 15% less likely to graduate college within six years than their white counterparts, and this gap is 25% between black and white students [33]. Such inequalities are particularly pronounced in STEM fields, where even more underrepresented students drop out of their college careers [1]. Also, student performance prior to entering college (e.g., on standardized tests) has often been found to strongly predict college performance across different subpopulations [9]. These overall trends suggest that what happens before college remains predictive of student success in higher education settings. Of course, this could be due to a variety of factors, such student background being correlated with patterns of historical and institutionalized oppression as well as other barriers that students from different backgrounds might face both before and during college.

2.1.2 Learning Behavior - LMS Data

In contrast to latent psychological states, learning behavior is a more extrinsic and observable predictor of academic success [7]. Behavioral patterns capture variations in college experience that may be orthogonal to students' incoming characteristics, allowing for insights into the mechanism of academic success at a day-to-day granularity. With the prevalence of digital learning platforms, learning behavior can be authentically recorded in the form of clickstream data. These time-stamped data record learner's interactions with LMSs. This allows researchers to create measures that look into the "black box" of study behaviors [5]. For example, how students allocate their study time is a consistent predictor of performance. Those who have more regular engagement patterns and who space out their study effort (instead of cramming) are more likely to be high-achieving [27]. Similarly, students who strategically regulate their learning effort (e.g., starting from exercise-oriented tactics and moving to other tactics based on encountered challenges) perform equally well but with less effort, compared to simply hard-working students [23].

2.1.3 Non-Cognitive Abilities - Survey Data

There is emerging evidence that non-cognitive factors, such as personality traits, task values and self-efficacy, are associated with positive academic outcomes even after controlling for cognitive factors measured by intelligence tests as well as various background characteristics [8]. Among these factors, researchers seem to have reached consensus that self-regulated learning skills are essential because unlike in K-12 schooling, college students have the flexibility as well as responsibility to actively and constantly monitor, reflect on, and adjust their motivation, cognition, and study behavior [37]. To better describe and measure a student's ability to regulate their learning process, [29] divided it into three subcomponents with two cognitive components (the use of cognitive strategies and the use of metacognitive strategies) and one non-cognitive component (resource management, including skills of time and study environment management, effort regulation, peer learning, and help seeking). A systematic literature review focused on online learning contexts found consistent evidence that resource management skills, especially time management skills and effort regulation skills, are predictive of performance [10]. While new technologies are creating novel measurement tools for these intangible qualities, the "ground truth" mostly comes from validated surveys.

2.2 Comparison of Different Data Sources

Previous work has examined combining various data sources for predictive analytics in higher education. For example, [2] combined institutional data, course performance data and LMS data to predict students' within-course success. However, there has been little work comparing the impact of various data sources on student success. [3] compared the impact of different types of institutional variables, including demographic variables, prior academic achievement, student majors, and academic achievement in college courses on predicting graduation and re-enrollment rates. [36] compared the impact of virtual learning environment (VLE) data, course assessment data, and a demographic variable on predicting whether a student's performance will drop in a course and whether a student will pass or fail a course. They generally found that using VLE data in conjunction with assessment data was seemingly better than using either alone. In what is perhaps the closest study to ours, [35] compared the impact of learning behavioral features, student background, and non-cognitive features measured

by a socio-emotional skill assessment on predicting within-course success.Our study differs from theirs in that we look at long-term outcomes as well as short-term outcomes, we analyze the fairness of predictive models, and we fit models that span across several courses.

2.3 Fairness of Predictive Analytics in Education

In recent years, the fairness and biases of machine learning algorithms and systems have developed into a focused research area in the general machine learning research community[1]. Research efforts encompass developing statistical measures of fairness, evaluating existing algorithms/systems, and correcting for biases in algorithmic pipelines, among others. As fairness is a concept rooted in a variety of disciplines, it has been a consensus that there is no single "correct" definition of fairness. Rather, what is fair is highly dependent on the specific application scenarios [6]. As such, contextualizing the fairness research in different fields is critical to improving real-world applications.

In earlier education research, there has been a focus on heterogeneous effects across student subpopulations in the contexts of testing [34], observational studies [39] and program evaluation [31]. These earlier perspectives resonate with the current theme of fairness, but as the adoption of predictive analytics systems in education for high-stakes purposes has a comparatively shorter history, formalized research on fairness in such contexts has been somewhat limited. Among the handful of empirical papers that have directly evaluated this aspect of predictive analytics in education, [13] showed through a simulation study that misspecified student models in intelligent tutoring systems could leave "slow" learners at lower mastery levels than "faster" learners; [16] examined the ROC curves from MOOC dropout prediction models, and identified significant gaps between gender groups through slicing analysis; and [19] used college application materials to predict on-time graduation and, employing the same slicing analysis, concluded that their model could make fair predictions across five sociodemographic groups.

As [6] points out, while the biases of predictive systems may be attributed to unfair algorithms, they can also arise from biased data which "reflect historical prejudices against certain social groups, prevailing cultural stereotypes, and existing demographic inequalities". Therefore, unlike the previous studies described above, this paper examines fairness as an attribute of *data sources* rather than of *algorithms*. We look at fairness with respect to between-groups differences in three metrics: accuracy, false positive rate, and false negative rate. These metrics are among the many fairness metrics that have been proposed in the literature [6]. For example, having an equal false negative rate between subgroups has been called "equality of opportunity" in the context of giving everyone an equal opportunity to receive a positive intervention (e.g., being part of the university's honor roll for having a high GPA) [17].

3. DATA AND METHODS

3.1 Data Sources

Following Section 2.1, this study compares the three widely available data sources in higher education settings: institutional data, Canvas LMS log data, and survey data. Specifically, we drew the sample of all students who enrolled and received final grades in ten fully online, introductory STEM courses taught from 2016 to 2018 at a large, public research university in the United States. Six of the courses were in public health while the remaining four were distributed across biology, chemistry and physics. These courses were the subject of a large research project, where our research team administered a series of standard survey questions about students' motivation, self-regulation and other psychological constructs before, during and/or after each course. Therefore, we had valid survey data across multiple courses. Also, looking at online courses ensured that LMS data can provide holistic representations of learning behavior. A total of 2,244 students were in the original dataset, and after data cleaning as described below in Section 3.2, the final sample size was 2,093. Traditionally underrepresented groups in STEM fields made up a large portion of the sample: 72% were female, 48% came from low-income families, 54% were first generation college students, 33% were underrepresented minorities (URM)[2], and 13% were transfer students.

3.2 Features and Outcomes

From each of the three data sources, we constructed a separate feature set in line with the literature. Table 1 gives a summary of these features. *Institutional features* included student demographics and academic achievement prior to college. *Click features* were derived from the LMS data and only included general measures of behavioral engagement to accommodate the variances in course design. Specifically, for each student in each course, we calculated the total number of clicks and total time spent over the first half of the course period. Time spent was calculated as the time lapse between adjacent click events. For the last click event of a student (with no subsequent event) or exceptionally lengthy lapses, we set a heuristic value of 90 seconds. The click counts and time spent were also broken down by categories, which were defined based on the URLs that click events pointed to, including "portal", "tasks", "content", "communication", "performance" and "miscellaneous." Restricting to the first half of course period speaks to the scenario of early identification of at-risk students for instructors. *Survey features* included four constructs of self-regulated learning skills and self-efficacy [29] from pre-course surveys launched during the first week of these courses. The completion rates of these surveys ranged from 65% to 93% across the ten courses. All survey items were adapted from Motivated Strategies for Learning Questionnaire (MSLQ), a popular questionnaire to measure self-regulation skills in online learning [30]. Each of the four constructs was measured by the average of corresponding survey items (Table 2).

As for outcomes, we defined two success measures. *Short-term success* was defined as a binary indicator of whether a student's final course grade was above the class median. Predicting this within-course outcome aligns with the needs of instructors to recognize struggling students in a timely manner [15]. Similarly, *long-term success* was defined as

[1]https://facctconference.org/

[2]This includes African American, Hispanic, and Native American students.

Table 1: Features derived from the three data sources

Institutional	Click	Survey
Female	Total clicks	Effort regulation
Transfer	Total clicks by category	Time management
Low income	Total time	Environment management
First-gen	Total time by category	Self-efficacy
URM	(All above for the first 5 weeks)	
SAT total score		
High school GPA		

Table 2: Details of survey features. Each feature was calculated as the average of its associated items.

Feature	Items (5-point Likert scale)
Effort regulation	I often feel so lazy or bored when I study that I quit before I finish what I planned to do (reverse coded). I work hard to do well in courses even if I don't like what I am doing. When coursework is difficult, I give up or only study the easy parts (reverse coded). Even when course materials are dull and uninteresting, I manage to keep working until I finish.
Time management	I keep a record of what my assignments are and when they are due. I plan my work in advance so that I could turn in my assignments on time.
Environment management	I usually work in a place where I can read and work on assignments without distractions. I can ignore distractions around me when I study.
Self-efficacy	I'm certain I can master the skills taught in this course. I'm certain I can figure out how to learn even the most difficult course material. I can do almost all the work in class if I don't give up.

whether a student's average GPA in the year that followed the course was above the median of their classmates in that course. Predicting this longer-term outcome is of interest to academic advisors and institutional policymakers because it can help them make appropriate policy changes early in students' academic careers to increase student success and graduation rates [22]. We used class medians to construct these outcomes instead of certain grade thresholds in order to better compare short-term and long-term results.

We examined all possible combinations of the three feature sets ($2^3 - 1 = 7$) regarding their ability to predict the two success measures. Therefore, a total of 14 binary classification problems were formulated. To fairly compare the prediction performance of these feature sets, students with missing values on more than 25% of all the individual features in Table 1 were dropped, which accounted for the decrease in sample size from 2,244 to 2,093. All continuous numerical features were standardized by centering to the median and scaling according to the interquartile range (IQR) to better handle outliers. For the remaining missing values, we performed multivariate imputation, i.e., modeling each feature with missing values as a function of other features.

3.3 Predictive Models

For each classification problem, we employed three common classification algorithms: logistic regression, support vector machines (SVM), and random forests. Course-level leave-one-group-out cross validation was used. In other words, the algorithm looped through the ten courses, and in each iteration used one course as the test set for the model trained on the remaining nine courses. Predicted values for each course were then put together from the ten iterations to evaluate the overall prediction performance. As our focus was the predictive power of different feature sets instead of models, we chose the classifier that produced the highest F-score for each combination of feature set and outcome. Because we used median splits to construct outcomes, class imbalance was not a concern and therefore no resampling was performed. The entire predictive modeling process was implemented using the scikit-learn Python library [28].

3.4 Evaluation

We evaluated the prediction results via three metrics. Accuracy measures the overall predictive power of the features used. False positive rate (FPR) reflects the probability of missing out "at-risk" students or "overplacing" students. False negative rate (FNR), on the other hand, captures the chances of "underplacing" students [32]. These metrics can shed light on potential consequences of using certain data source(s) in different applications. From there, we can compare the utility of different data sources in a holistic manner.

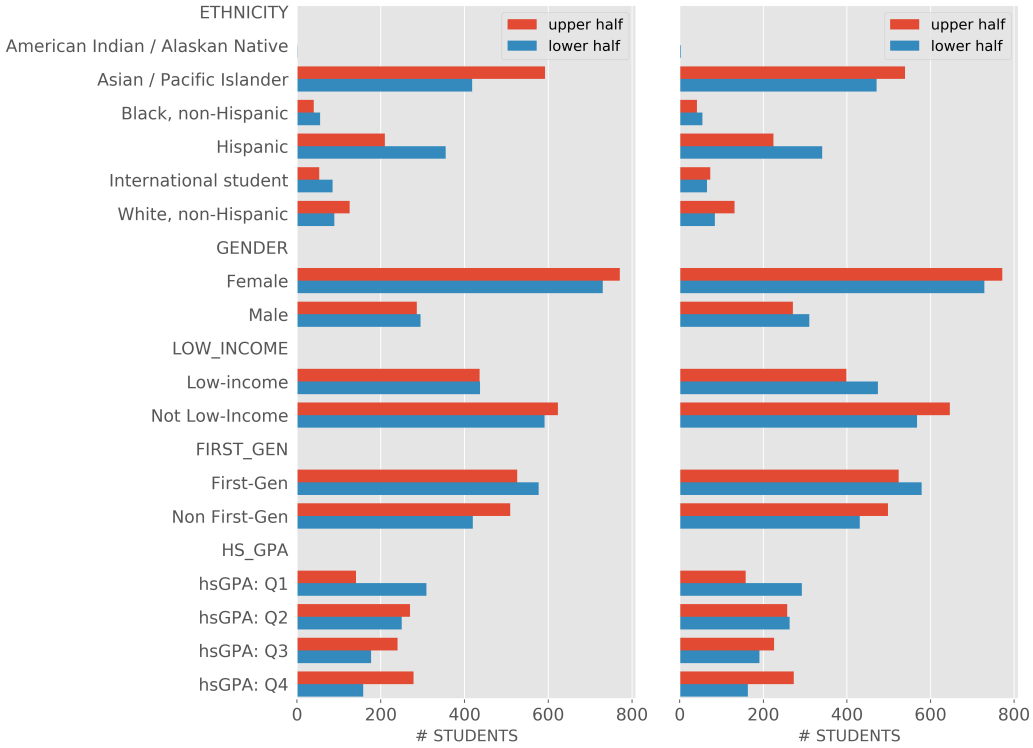
We further evaluated each data source's contribution to the fairness of prediction results. Fairness was conceptualized as the performance parity across student subpopulations when the prediction was performed on the entire student sample. Specifically, we focused on an array of historically disadvantaged subpopulations and compared each of them with a corresponding reference group on the three metrics. For example, we compared the accuracy, FPR and FNR within Latinx students with those within white students. Figure 1a and 1b plot the group size and outcome distribution of these selected groups, where the last group under each category was the reference group.

Statistically, we computed the following disparity metrics for each disadvantaged group g:

$$acc\_disparity = acc_{ref}/acc_g \quad (1)$$

$$fpr\_disparity = fpr_g/fpr_{ref} \quad (2)$$

$$fnr\_disparity = fnr_g/fnr_{ref} \quad (3)$$



(a) Short-term success (b) Long-term success

Figure 1: Outcome distribution within different student subpopulations. Short-term success: whether a student's final course grade was above the class median. Long-term success: whether a student's average GPA in the following academic year was above the class median.

and separately tested whether each of this disparities was significantly larger than 1 using one-sided two proportion z-test. The larger these ratios were, the more this student group was "discriminated against" by the prediction model. We used the less flexible one-sided test because of the consistent evidence that traditionally underrepresented groups experience more inequities than their counterparts in academic settings [1]. All these ratios combined would characterize the comparative utility of different data sources for fair predictions of college success.

4. PREDICTIVE UTILITY OF DIFFERENT DATA SOURCES

4.1 Overall Prediction Performance

Table 3 presents the prediction results on our full student sample across different feature and outcome combinations. In each column, the best-performing model is in bold to indicate which feature set(s) best predicted the corresponding outcome in the column header in terms of the given metric. Among the final sample of 2,093 students, 1,062 (50.7%) had short-term and 1,048 (50.1%) had long-term outcomes above their class median[3]. These numbers serve as the naïve baselines of prediction accuracy where all the students were simply predicted to be in the upper half (majority class).

[3]The slight deviation from 50% was due to the drop of students with too much missing information on predictors, as described in Section 3.2.

When the three data sources were used separately, institutional features and click features both achieved an overall accuracy of around 0.6 for either short-term or long-term outcomes, which was significantly higher than the baseline ($p < 0.001$ for all four cases). Specifically, institutional features appeared to be slightly more predictive of short-term success and click features predicted long-term success a little better, but neither of these comparisons was statistically significant. On the contrary, survey features had much weaker predictive utility because they predicted both outcomes with significantly lower accuracy than the worse of the other two features ($p < 0.001$ for short term and $p = 0.005$ for long term). When these feature sets were combined in different ways, we mostly saw improvement in the overall accuracy. The combination of institutional and LMS data led to the most noticeable accuracy increase in predicting both outcomes ($\Delta = 0.052, p < 0.001$ for short term and $\Delta = 0.037, p = 0.014$ for long term), evidencing complementary signals of student success in these two data sources. Survey data provided limited marginal utility as adding survey features to other feature sets never led to a statistically significant increase in accuracy and sometimes even had negative effects. However, the highest accuracy in predicting the short-term outcome was achieved when all three feature sets were used together.

Given the tradeoff between false positives and false negatives, overall best-performing feature sets did not necessarily

Table 3: Prediction performance on the entire student sample ($N = 2,093$). The best result in each column was in bold. Short: predicting whether a student's final course grade was above the class median; long: predicting whether a student's average GPA in the following academic year was above the class median.

Feature	Accuracy		FPR		FNR	
	Short	Long	Short	Long	Short	Long
Institutional	0.618	0.599	0.467	0.412	**0.299**	0.389
Click	0.602	0.613	0.485	0.385	0.313	0.389
Survey	0.534	0.557	0.599	0.385	0.336	0.502
Institutional+Click	0.670	**0.650**	0.351	**0.330**	0.310	0.370
Institutional+Survey	0.633	0.608	0.398	0.397	0.337	0.386
Click+Survey	0.609	0.604	0.431	0.457	0.353	0.335
Institutional+Click+Survey	**0.675**	0.638	**0.348**	0.402	0.303	**0.323**

have the lowest error rates. Among the three cases using a single data source, institutional features had both the lowest FPR and the lowest FNR for the short-term outcome ($p = 0.402$ for FPR and $p < 0.001$ for FNR compared to the second lowest). The same features also tied with click features for the lowest FNR in predicting the long-term outcome, while the latter led to the lowest FPR in the long term (tied with survey features). Combining these two data sources significantly lowered FPR ($\Delta = -0.116, p < 0.001$ for the short term and $\Delta = -0.055, p = 0.009$ for the long term) but not FNR. As for survey data, the patterns of error rates were more complicated than of overall accuracy. When used alone, survey features mostly led to higher error rates than the other two feature sets, except for FPR in the long term. On the other hand, adding survey features to other feature sets largely decreased FNR for long-term and FPR for short-term success predictions despite the fact that these metrics were exceptionally high in the case of using survey data alone.

4.2 Fairness of Predictions

Following Section 3.4, we computed and tested the extent to which each disadvantaged student subpopulation suffered discriminatory predictions (i.e., algorithmic bias) compared to their reference group under each combination of feature set and outcome. Figure 2a and 2b illustrate these results for short-term and long-term success prediction, respectively. Each cell colors a bias against a certain student subpopulation in a specific model. Darker cells suggest larger biases and crossed out cells represent those that were statistically significant ($p < 0.05$) after correcting for multiple testing within each background attribute. Subpopulations with fewer than 10 students were omitted as the error rates were less reliable.

Overall, there was no feature set that was entirely free from biased predictions. Across both outcomes, institutional features consistently led to higher FNR within various disadvantaged student subpopulations than within their peers. In other words, these students were more likely to be *underestimated* by the prediction model. This finding resonates with previous research that being aware of protected attributes (e.g., ethnicity) might induce identity-based biases in predictive analytics [6]. Adding other features to institutional ones alleviated some of these biases only in a marginal sense. That is, inclusion of institutional features seemed to largely determine the discriminatory behaviors of the model. Identity-blind LMS data was a fairer data source as the number of discriminated subpopulations was smaller. Compared to their reference groups, click features on their own significantly *overestimated* female students for both outcomes and Asian, Hispanic and first-generation college students for the long-term outcome. Survey data turned out to be neither accurate nor fair. When used alone, survey features led to significant biases against certain subpopulations across all metrics and outcomes. When combined with other feature sets, they did little to offset existing biases in most cases, except when they were used together with click features to predict long-term success. However, this latter case may suggest that survey data had equally low predictive utility for long-term success across different student subgroups.

The plots also allowed for insights into the extent to which different student subpopulations were exposed to algorithmic biases across different scenarios. Ethnic minorities, students from low-income families and first-generation college students were more prone to underestimation. Female students were more likely to be overestimated than male students especially in the long term. Moreover, international students and students with lower high school GPAs suffered both more underestimation and less accurate predictions compared to their peers. Note that unlike other variables in the plots, high school GPA is an acquired attribute. Hence, our evidence of algorithmic bias implied that a student can be stigmatized due not only to their demographic attributes but to their past (academic) experience as well.

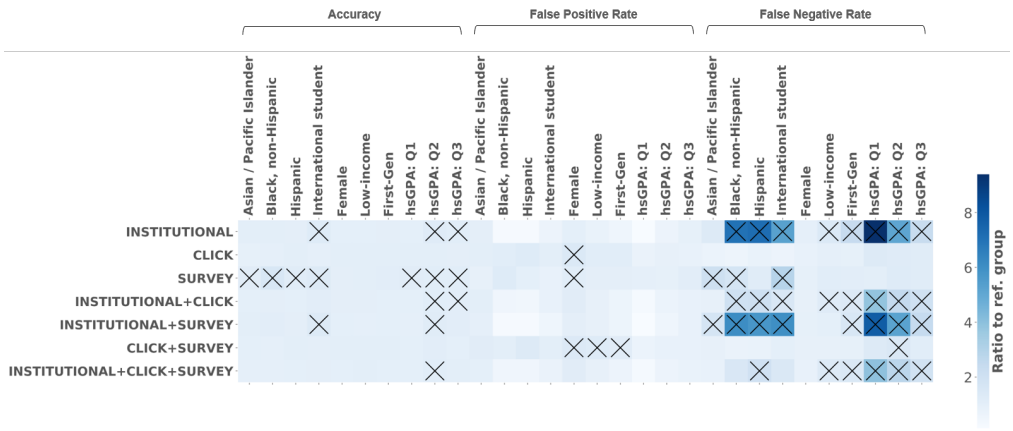
4.2.1 A Closer Look into Institutional Data

Reflecting on the consistent biases against disadvantaged student subpopulations when using institutional data, we also tested if removing a specific institutional feature (e.g., gender) would eliminate the bias against the corresponding disadvantaged group (e.g., female). Surprisingly, all the results looked qualitatively similar regardless of which feature we removed. This suggested the intersectionality of minority identities, i.e., a student from one disadvantaged group tended to have another disadvantaged characteristic as well. As such, simply removing individual background variables would not necessarily make the predictions fairer.

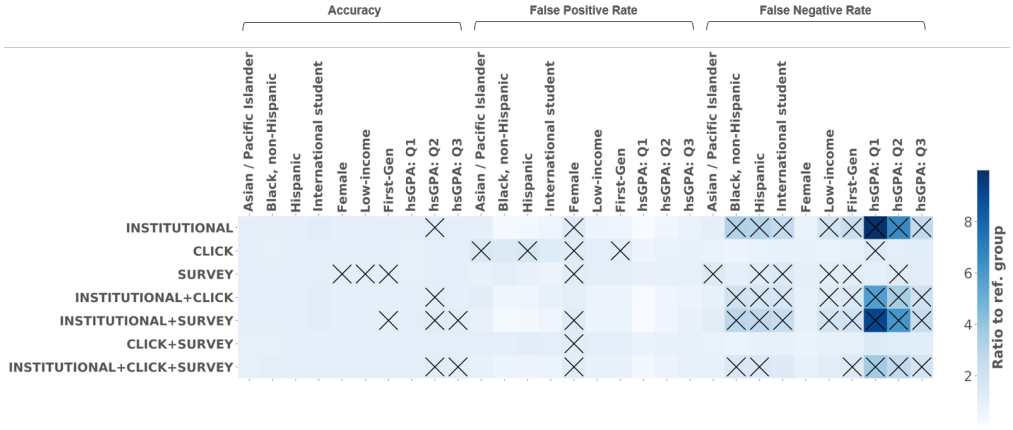
5. DISCUSSIONS

5.1 Reflections on the Results

Our results shed light on the predictive validity of different sources of student data on college success. Our overall results agree well with those of [35], where features from an



(a) Short-term success



(b) Long-term success

Figure 2: Illustration of prediction fairness. Each cell represents the algorithmic bias against a historically disadvantaged student subpopulation (compared to the corresponding reference group) in the specific scenario. Crosses represent statistically significant biases ($p < 0.05$) after correcting for multiple testing. Short-term success: whether a student's final course grade was above the class median. Long-term success: whether a student's average GPA in the following academic year was above the class median.

assessment of socio-emotional skills were least predictive of course success, which is similar to the ineffectiveness of our survey data. On the other hand, they found that models using institutional variables and clickstream features performed better and comparably to one another, as we did. They also discovered that combining clickstream behaviors with socio-emotional skills outperformed institutional data alone, which we also saw with the FNR for the long-term outcome. Interestingly, they did not find additional predictive utility of higher-level behaviors (sequential features) from clickstream data, which we did not further investigate.

The limited ability of pre-course survey data to accurately predict either short-term or long-term success may suggest that self-reported measures of self-regulated learning are not key factors of online learning processes or performance. However, as suggested by previous research [12], it may also suggest that students tend to overestimate their use of learning strategies in online courses. This is likely because students make estimations of their future behaviors based on memories of similar past events that are usually unreliable [21]. Thus, more research is needed to understand how to help students provide valid data of their learning skills as well as other psychological attributes in surveys [25].

When it comes to fairness, several interesting trends emerge. First, predictions using institutional data, which had the lowest FNR overall, were actually discriminatory when it comes to FNR for both outcomes. In particular, institutional data discriminated against students from underrepresented minority groups, low-income students, first-generation college students, and students with low high school GPA. This suggests that these models tend to disproportionately label students from these subpopulations as having below-median performance. In order to achieve higher overall accuracy, these models appear to be using a heuristic of classifying students as above or below median based on the majority class within the subpopulations that they belong to (see Figure 1). Therefore, one of the main sources of unfairness may just be the original class imbalance in different student subpopulations. When this imbalance results from historical inequities, the model will simply replicate those

inequities and produce unfair predictions.

On the other hand, we found that using click features tended to be fair with respect to FNR, but instead somewhat discriminatory with respect to FPR, for several student subpopulations. Contrary to the discrimination brought by institutional features, this form of discrimination could occur just because the model is blind to individual background. More specifically, students coming from different backgrounds may on average exhibit similar learning behaviors, but their likelihood to succeed might differ due to factors that correlate with their socio-economic status. Since the click features do not have access to students' background information, they may predict that students from disadvantaged backgrounds are likely to succeed at a disproportionately high rate.

One specific and possibly counterintuitive trend is seen when it comes to gender biases. While none of the feature sets discriminated against female students in terms of FNR, almost all of the feature sets discriminated against them in terms of FPR for at least one of the two outcomes. In fact, female students tend to have higher GPA than their male peers in the dataset (see Figure 1). This reinforces the inference that for institutional features, the models classify students into the majority class of their subpopulations in order to maximize accuracy. On the other hand, the fact that using only LMS and/or survey data is also biased against female students in terms of FPR might be due to something else. This suggests that female students might (a) exhibit different click behaviors and survey responses from men, which tend to be predictive of better performance; or (b) have different baseline levels of engagement (e.g., likelihood of clicking on LMS pages) independent of their likelihood of success. If the former is true, click behaviors and/or survey responses could act as a weak proxy for gender, even though gender is not encoded in these features.

5.2 Practical Implications

In general, prediction errors are inevitable, but it is important to be aware of and minimize potential misplacement that may result in severe negative consequences. Below, we discuss three major scenarios where prediction models are used for educational decision making and the implications of our findings in these cases.

First, higher education has a long history of screening applicants for desirable educational opportunities such as merit-based scholarships, where the award is based on the prediction of student future performance. In this case, underestimating student performance may limit their educational development. While institutional data is one of the most widely used data sources for these purposes, our results suggest that institutional data alone might be more likely to underestimate achievement of students from disadvantaged background as compared to their peers. Moreover, these systematic biases do not go away easily even when other common data sources are added. Therefore, it is important for policymakers to cautiously employ predictive analytics for selecting students since it may result in unfair exclusion of already disadvantaged students from critical educational opportunities and access to social mobility through education [18].

In community college settings, institutional data has also been used to evaluate students' readiness for college-level courses and assign students into remediation [32], as well as to understand the impact of remedial and preparatory courses on subsequent college success [24]. Put in this scenario, our results would suggest that students from historically disadvantaged subpopulations are more likely to be misplaced into remediation than their counterparts when they are actually capable of taking advanced courses. While remedial courses are designed to help academically underprepared students, they also increase students' cost and may delay student progression towards their degree goal [4]. For both this and the previous application scenarios, a potential algorithmic solution might be setting separate thresholds for different subpopulations to ensure fairness, as [20] suggested.

Finally, in the recent research and practice of online learning, LMS data have been commonly used to predict student performance and identify at-risk students [36]. Students who are identified as being at risk of low performance or dropout will often be placed into light-touch or optional academic support, such as receiving email reminders and tutoring services [11]. In this context, it might be more concerning to overestimate student performance and ignore students in need than to underestimate student performance and place them to educational resources that they could opt out of. Our findings indicate that compared to males, female students would be especially likely to experience overestimation and therefore would not receive academic resources that they need. In this case, incorporating institutional data into the prediction might not be as problematic in order to leave no student behind.

6. CONCLUSION

In this paper, we responded to the call for research to evaluate and compare the utility of common student data sources (i.e., institutional data, LMS data and survey data) for building predictive analytics applications in the context of higher education [14]. We aimed to find out what data sources and their combinations predicted short-term and long-term college success both accurately and fairly across different student subpopulations. Our results suggest that overall, institutional data and LMS data on their own have decent predictive utility for either instructors' or policymakers' needs to identify students in need. Using them together further strengthens that predictive power. Survey data alone poorly predicts student success and only marginally helps alleviate some of the prediction errors in the presence of other data sources. With regard to fairness, institutional data consistently leads to higher false negative rate (underestimation) within historically disadvantaged students subpopulations than within their peers. LMS data, on the other hand, tends to overestimate some of these disadvantaged groups (e.g., female students) more often than their counterparts and these biases would be overridden by institutional data when the latter is added. Survey data makes very limited contribution to fair predictions. Interestingly, all sources of student data tend to overestimate female students who perform better than male students on average in our case. Also, students with lower prior achievement are no less affected by underestimation than underrepresented demographic groups.

These results combined suggest that using multiple data

sources in college success prediction is beneficial for institutional stakeholders from both technical and ethical perspectives. Specifically, given the infancy and decent predictive utility of LMS data, institutions should feel encouraged to invest in the infrastructure to store, manage and analyze such data and integrate LMS-based behavioral measures into the routines of institutional research. On the other hand, utilizing multiple data sources still cannot guarantee fair predictions of college success especially for students who have less competitive academic records and who are historically disadvantaged in higher education. Therefore, it is advisable to combine the intelligence of experienced practitioners and data-driven applications for decision-making in the wild, in hopes of minimizing the risk that students are unfairly excluded from their optimal pathways due to biased algorithms or human judgement.

Our work has a few limitations which point to meaningful future work. First, the scope of our feature sets was limited and not representative of the full potential of different data sources. For example, for survey features we only used measures of self-regulation, but there are other psychological constructs that play equally important roles in learning processes. Therefore, our findings should be taken as a proof of concept in terms of systematically evaluating different data sources. Future work will extend the current piece to more comprehensive data sources that institutions have good control over [19, 3] and to broader feature sets informed by existing research. Second, while we briefly reflected on the prediction results and practical implications, we did not formally examine how the biases illustrated in Figure 2 permeate through the predictive analytics pipeline. Future work will examine this aspect more thoroughly, as well as how to convey these sources of bias to stakeholders for more prudent decision-making on student data usage.

7. ACKNOWLEDGMENT

This study is supported by the National Science Foundation (Grant Number 1535300).

8. REFERENCES

[1] S.-A. A. Allen-Ramdial and A. G. Campbell. Reimagining the pipeline: Advancing stem diversity, persistence, and success. *BioScience*, 64(7):612–618, 2014.

[2] K. E. Arnold and M. D. Pistilli. Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 267–270, 2012.

[3] L. Aulck, D. Nambi, N. Velagapudi, J. Blumenstock, and J. West. Mining University Registrar Records to Predict First-Year Undergraduate Attrition. In *The 12th International Conference on Educational Data Mining (EDM)*, pages 9–18, Montréal, Canada, 2019.

[4] T. Bailey, D. W. Jeong, and S.-W. Cho. Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, 29(2):255 – 270, 2010. Special Issue in Honor of Henry M. Levin.

[5] R. Baker, D. Xu, J. Park, R. Yu, Q. Li, B. Cung, C. Fischer, F. Rodriguez, M. Warschauer, and P. Smyth. The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: opening the black box of learning processes. *International Journal of Educational Technology in Higher Education*, 17:1–24, 2020.

[6] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[7] G. Beattie, J.-W. P. Laliberté, C. Michaud-Leclerc, and P. Oreopoulos. What sets college thrivers and divers apart? A contrast in study habits, attitudes, and mental health. *Economics Letters*, 178:50–53, may 2019.

[8] G. Beattie, J.-W. P. Laliberté, and P. Oreopoulos. Thrivers and divers: Using non-academic measures to predict college success and failure. *Economics of Education Review*, 62:170–182, feb 2018.

[9] E. P. Bettinger, B. J. Evans, and D. G. Pope. Improving college performance and retention the easy way: Unpacking the ACT exam. *American Economic Journal: Economic Policy*, 5(2):26–52, may 2013.

[10] J. Broadbent and W. L. Poon. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *Internet and Higher Education*, 27:1–13, 2015.

[11] S. P. M. Choi, S. Lam, K. C. Li, and B. T. M. Wong. Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions. *Journal of Educational Technology & Society*, 21(2):273–290, 2018.

[12] M. K. DiBenedetto and H. Bembenutty. Within the pipeline: Self-regulated learning, self-efficacy, and socialization among college students in science courses. *Learning and Individual Differences*, 23:218–224, 2013.

[13] S. Doroudi and E. Brunskill. Fairer but not fair enough: On the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, pages 335–339, Tempe, AZ, USA, mar 2019. ACM.

[14] C. Fischer, Z. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, and M. Warschauer. Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44:130–160, 2020.

[15] D. Forteza, T. Harfield, J. Whitmer, and A. Dietrichson. What does it take to predict student risk? Evaluating LMS data to determine readiness for predictive modeling. Technical report, Blackboard Inc., 2017.

[16] J. Gardner, C. Brooks, and R. Baker. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, pages 225–234, Tempe, AZ, USA, 2019. ACM Press.

[17] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[18] R. Haveman and T. Smeeding. The role of higher education in social mobility. *The Future of Children*, 16(2):125–150, 2006.

[19] S. Hutt, M. Gardner, A. L. Duckworth, and S. K. D'Mello. Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications. In *The 12th International Conference on Educational Data Mining (EDM)*, pages 79–88, Montréal, Canada, 2019.

[20] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. Algorithmic Fairness. *AEA Papers and Proceedings*, 108:22–27, 2018.

[21] Q. Li, R. Baker, and M. Warschauer. Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education*, page 100727, 2020.

[22] Y. Luo and Z. A. Pardos. Diagnosing University Student Subject Proficiency and Predicting Degree Completion in Vector Space. In *Proceedings of the Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, LA, USA, 2018.

[23] W. Matcha, D. Gašević, N. A. Uzir, J. Jovanović, and A. Pardo. Analytics of Learning Strategies: Associations with Academic Performance and Feedback. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, pages 461–470, Tempe, AZ, USA, 2019. ACM Press.

[24] H. Nguyen, L. Wu, C. Fischer, G. Washington, and M. Warschauer. Increasing success in college: Examining the impact of a project-based introductory engineering course. *Journal of Engineering Education*, 2020.

[25] J. L. Osterhage, E. L. Usher, T. A. Douin, and W. M. Bailey. Opportunities for self-evaluation increase student calibration in an introductory biology course. *CBE—Life Sciences Education*, 18(2):ar16, 2019.

[26] Z. A. Pardos, Z. Fan, and W. Jiang. Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction*, pages 1–39, feb 2019.

[27] J. Park, R. Yu, F. Rodriguez, R. Baker, P. Smyth, and M. Warschauer. Understanding Student Procrastination via Mixture Models. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM)*, Buffalo, NY, United States, 2018.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[29] P. R. Pintrich and E. V. De Groot. Motivational and Self-Regulated Learning Components of Classroom Academic Performance. *Journal of Educational Psychology*, 82(1):33–40, 1990.

[30] P. R. Pintrich, D. A. F. Smith, T. Garcia, and W. J. McKeachie. A manual for the use of the motivated strategies for learning questionnaire (mslq). Technical report, Ann Arbor, MI, 1991.

[31] M. C. Schippers, A. W. Scheepers, and J. B. Peterson. A scalable goal-setting intervention closes both the gender and ethnic minority achievement gap. *Palgrave Communications*, 1(1):1–12, 2015.

[32] J. Scott-Clayton. Do High-Stakes Placement Exams Predict College Success? 2012.

[33] D. Shapiro, A. Dundar, F. Huie, P. Wakhungu, A. Bhimdiwala, and S. Wilson. Completing College: A State-Level View of Student Completion Rates (Signature Report No. 16a). Technical report, National Student Clearinghouse Research Center, Herndon, VA, 2019.

[34] R. L. Thorndike. Concepts of culture-fairness. *Journal of Educational Measurement*, 8(2):63–70, 1971.

[35] J. Whitmer, S. S. Pedro, R. Liu, K. E. Walton, J. L. Moore, and A. A. Lotero. The Constructs Behind the Clicks. Technical report, ACT, Inc., 2019.

[36] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek. Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 145–149, 2013.

[37] C. A. Wolters. Self-regulated learning and college students' regulation of motivation. *Journal of educational psychology*, 90(2):224, 1998.

[38] J. Xie, A. Essa, S. Mojarad, R. S. Baker, K. Shubeck, and X. Hu. Student Learning Strategies and Behaviors to Predict Success in an Online Adaptive Mathematics Tutoring System. In *Proceedings of the 10th International Conference on Educational Data Mining*, pages 460–465, Wuhan, China, 2017.

[39] D. Xu and S. S. Jaggars. Performance Gaps Between Online and Face-to-Face Courses: Differences Across Types of Students and Academic Subject Areas. *The Journal of Higher Education*, 85(5):633–659, 2014.