Interpretable Models Do Not Compromise Accuracy or Fairness in Predicting College Success*

Catherine Kung[1] , **Renzhe Yu**[1†]
[1]University of California, Irvine
{kungc1, renzhey}@uci.com

Abstract

The presence of "big data" in higher education has led to the increasing popularity of predictive analytics for guiding various stakeholders on appropriate actions to support student success. In developing such applications, model selection is a central issue. As such, this study presents a comprehensive examination of five commonly used machine learning models in student success prediction. Using administrative and learning management system (LMS) data for nearly 2,000 college students at a public university, we employ the models to predict short-term and long-term academic success. Beyond the tradeoff between model interpretability and accuracy, we also focus on the fairness of these models with regard to different student populations. Our findings suggest that more interpretable models such as logistic regression do not necessarily compromise predictive accuracy. Also, they lead to no more, if not less, prediction bias against disadvantaged student groups than complicated models. Moreover, prediction biases against certain groups persist even in the fairest model. These results thus recommend using simpler algorithms in conjunction with human evaluation in instructional and institutional applications of student success prediction when valid student features are in place.

1 Introduction

With the increasingly powerful data infrastructure and visionary strategic plans of institutions, the use of predictive analytics in higher education has seen its rapid growth over the past decade in promoting student success through data-driven applications [Ekowo and Palmer, 2016]. While it is traditionally challenging for educators to attend to every single student across various matters regarding students' learning and life experience, predictive analytics can move them closer to this goal by identifying students who might benefit the most from certain resources being evaluated, such as one-on-one conversations, supplemental training programs and scholarships. In practice, a variety of machine learning models are available for predictive analytics, each with certain pros and cons, making model selection a critical issue for both technical and non-technical stakeholders. To gauge the tradeoffs between different models in both instructional and institutional scenarios, some systematic evaluation of model performance needs to be in place, but such effort is largely underrepresented in previous literature. In an effort to guide instructors, administrators and policymakers in making more informed decisions regarding the use of machine learning models for predicting student success, the current study aims to present a comprehensive examination of the utility of various predictive models under both technical and ethical considerations.

Following the oft-studied tension between interpretability and predictive power of machine learning models, we investigate classical models that are more straightforward to interpret as well as black-box models that have more complicated structures. Comparing the overall prediction performance of these models, we aim to determine whether simpler or more complex models are a better fit for predicting short-term and long-term student outcomes. Furthermore, while predictive analytics can create opportunities of more tailored policies and pedagogies, the processes by which these opportunities are created can amplify existing inequities or create new ones [Barocas *et al.*, 2019]. As such, we borrow the research framework of algorithmic fairness to evaluate the contribution of different models to fair college success predictions. We hope that these examinations will shed light on the usefulness of predictive models in a comprehensive manner, so that various stakeholders can be better informed of the common choices available to them and optimize their decisions in practice.

2 Related work

Earlier educational research and more recent learning analytics research have identified significant predictors of academic success for college students, such as students' background characteristics [Ishitani, 2006] and learning behavior [Park *et al.*, 2018]. For example, a student's family background can be associated with the academic resources they can secure prior to college, which will predict timely graduation comparison; learning behavior captured through learn-

*This is the reformatted version of a published paper. The published version can be found here.

†Contact Author

ing management systems (LMS) can reveal students' study habits, such as procrastination, which correlate well with academic performance. When these insights are leveraged in data-driven applications, the predictive power becomes more of interest, which varies across models and is conditional on predictors. One aspect of model choice is the tradeoff between the accuracy and interpretability of machine learning models. While complex models may better capture the underlying patterns within data, they are more susceptible to overfitting, less computationally efficient and harder to interpret than simpler models [Rudin, 2019]. Our study follows this line of inquiry as well as some earlier studies that investigated deployed educational systems [Jayaprakash *et al.*, 2014], with a focus on the comparison of prediction performance across algorithms.

As predictive analytics are increasingly being used to make decisions that influence people's lives, the topic of fairness has recently made its way into the machine learning community. Fairness-aware machine learning algorithms generate predictive outcomes that are non-discriminatory for people based on their sensitive attributes include race, sex, socioeconomic status, etc. [Friedler *et al.*, 2019] Although there is a wealth of criteria by which an algorithm is regarded to be fair, in the case of (binary) classification, they generally fall into three formulations, including independence, separation and sufficiency [Barocas *et al.*, 2019]. The criterion of independence requires that the sensitive attribute and the predicted probabilities be independent of each other. Separation requires that the true positive rate and the false positive rate experienced by all groups in the sensitive attribute be equal. The final fairness criterion is sufficiency, which is fulfilled when, of those with equal predicted probabilities, the distribution of actual classes is orthogonal to sensitive attributes. Due to the comparatively short history of predictive analytics solutions in higher education, formal examination of algorithmic fairness in this context has been rather limited [Hutt *et al.*, 2019]. This study aims to enrich and inspire discussion on this topic, following the foregoing framework.

3 Data and method

We analyzed a population of 1,971 students who had been enrolled in ten fully online, introductory STEM courses offered at a public, four-year university from 2016 to 2018. Building upon our previous work that investigated the same context to understand the predictive utility of different data sources [Yu *et al.*, 2020], we collected administrative data and Canvas LMS data and extracted an array of features (predictors) for each student. From administrative data, we included student demographics (age, gender, transfer status, family income level, first-generation status, underrepresented minority) and academic history (SAT total score, high school GPA, cumulative college GPA) information. From LMS data, we calculated total clicks, total clicks by category, total time on page and total time by category for the first two weeks of each course. To reflect the diversity of student success prediction tasks in practice, two illustrative prediction targets were considered for each student in each course: short-term success, defined as whether a student received a course final score above the median of the class they were in; long-term success, defined as whether a student's average GPA for the year following the course was above the median of that class. We chose the class median instead of more practical thresholds (e.g., letter grade C) primarily for comparability between short-term and long-term predictions. For both targets, we employed five commonly used machine learning algorithms: logistic regression, support vector machines (SVM), random forest, decision tree, and Naïve Bayes. Course-level leave-one-out cross validation was performed to iteratively get predicted labels within each course.

To evaluate the utility of these models, we first measured the overall predictive power of each model by accuracy, which was calculated according to the following formula:

$$acc = P\left(\hat{Y} = Y\right) \tag{1}$$

where $\hat{Y}$ is the predicted label and Y is the actual target value.

The models were further evaluated for fairness based on the three fairness criteria. For this purpose, we included five sensitive attributes: ethnicity, gender, low income status, first generation college student status and high school GPA quartile within their class. For each attribute, we identified one reference group (e.g., white students for ethnicity) and compared them with every other more disadvantaged group(s) (e.g., Latinx students) on their prediction results. To evaluate independence, we tested the following null hypothesis for each non-reference group (g_i):

$$H_0 : P\left(\hat{Y} = 1 \middle| G = g_i\right) = P\left(\hat{Y} = 1 \middle| G = g_{ref}\right) \tag{2}$$

For separation, the following null hypothesis was tested (equal false positive rates):

$$H_0 : P\left(\hat{Y} = 1 \middle| Y = 0, G = g_i\right) = P\left(\hat{Y} = 1 \middle| Y = 0, G = g_{ref}\right) \tag{3}$$

Sufficiency involved a more complicated design. Students were split into 5 bins (b_j) based on their raw predicted scores (probabilities) and the following null hypothesis was tested:

$$H_0 : P\left(\hat{Y} = 1 \middle| S \in b_j, G = g_i\right) = P\left(\hat{Y} = 1 \middle| S \in b_j, G = g_{ref}\right) \tag{4}$$

We applied two-proportion z-test with Bonferroni correction for all the tests above, with the overall significance level set to 0.1. A model would fail to meet a criterion for a sensitive attribute if any associated test within that attribute rejected the null hypothesis. Putting together all these results, we then qualitatively examined which model(s) contributed to fairer predictions.

4 Results and discussion

A total of 10 models (5 models$\times$2 targets) were created and systematically compared. The overall accuracy of each model is shown in Table 1. Although the general machine learning literature acknowledges the tradeoff between model interpretability and prediction accuracy, our results indicate that logistic regression, the most interpretable model of the five, performed marginally better than any other model in terms of

Model	Short-term	Long-term
Logistic Regression	0.701	0.724
SVM	0.693	0.712
Random Forest	0.696	0.723
Decision Tree	0.667	0.714
Naïve Bayes	0.659	0.644

Table 1: Overall accuracy across models predicting college success

predicting both short-term and long-term success (p¿0.1 for both targets when compared to the second best-performing model

Regarding the fairness of each model, a summary of results from various statistical tests can be found in Figure 1. Within each cell, a colored square denotes a sensitive attribute for which the model failed to satisfy the corresponding criterion. Overall, the results indicate that simpler models such as logistic regression and Naïve Bayes do not necessarily compromise fairness, as can be told from the comparatively small number of discriminated attributes for these models. Below we will look slightly deeper into the fairness aspects behind these summary tables. We draw on results from logistic regression given its competitive performance on both overall accuracy and fairness.

Model	Independence	Separation	Sufficiency
Logistic Regression	E F H	E F H	E H
SVM	E F H	E F H	E H
Random Forest	E I F H	E F H	E
Decision Tree	E I F H	E G H	E G H
Naïve Bayes	E G F H	E G H	E H

(a) Short-term

Model	Independence	Separation	Sufficiency
Logistic Regression	E I H	H	E I H
SVM	E G I F H	E I F H	
Random Forest	E I F H	F H	E H
Decision Tree	E I F H	H	E I F H
Naïve Bayes	E G H	H	E H

(b) Long-term

Figure 1: Fairness of models predicting college success. Each square denotes an attribute that failed to pass the "fairness test". Legend: ethnicity (E), gender (G), low income status (I), first-generation college student status (F), high school GPA quartile (H).

Logistic regression failed to achieve independence for ethnicity and high school GPA on both targets. While the statistical nature of independence requires that the prediction model close existing gaps of the target across different groups, we observed the opposite behavior. The model predicted fewer students from already disadvantaged groups to be in the upper half (positive class) than they actually were in the original dataset. As an example, Figure 2 illustrates that while more students whose high school GPA was in the third and fourth quartile were predicted to be in the positive class in comparison to their true labels, it is the other way around for those whose high school GPA was in the lower quartiles. In other words, the gaps were widened by the model. This may be an issue as students from disadvantaged groups are more likely to struggle yet less likely to be identified for additional resources necessary for them to succeed.

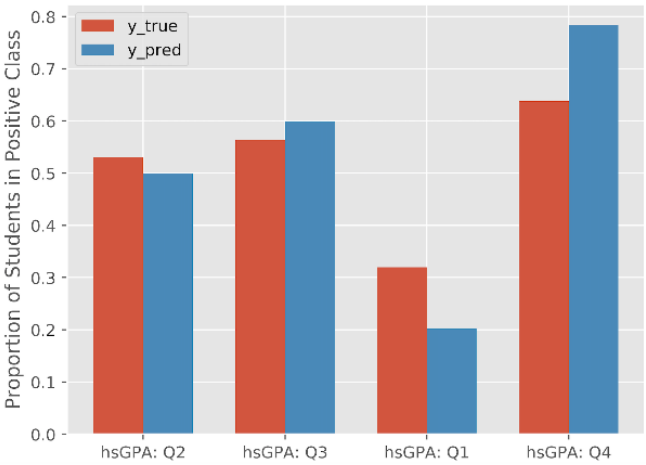

Figure 2: Comparison of true and predicted labels from logistic regression predicting short-term success, by high school GPA quartiles

From the perspective of separation, logistic regression produced bias against three out of five sensitive attributes on short-term success and only one attribute on long-term success. We further found that Asian/Pacific Islander students, students in the fourth high school GPA quartile and non-first-generation college students had the highest false positive rates within their corresponding attributes. That is, the model was overly confident in students from these academically "advantaged" groups than in their disadvantaged peers. This again indicates that the model reinforced existing inequities reflected in the dataset instead of closing the gaps.

When predicting both targets, the logistic regression model was well-calibrated by all sensitive attributes except for ethnicity and high school GPA. While sufficiency often comes for free [1], which was mostly true in our case, we observed that deviations often occurred on both ends of the bins. Specifically, student from the most disadvantaged group might have higher positive rates than other groups in the first bin or lower positive rates in the last bin. For example, Figure 3 shows that deviations in the calibration curves for higher school GPA quartiles were primarily seen for students in the first quartile, specifically in the last bin. This further con-

firms the gap-widening behavior of even the fairer prediction model.

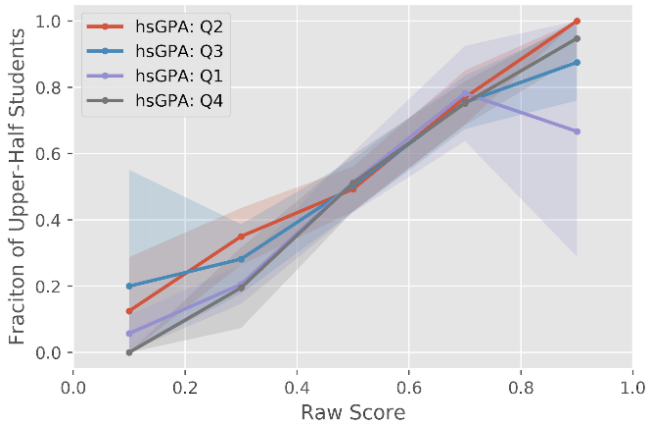

Figure 3: Calibration curve from logistic regression model predicting short-term success, by high school GPA quartile

5 Conclusion

This study presents a systematic comparison of commonly used machine learning models in the context of predicting student success in higher education. Building upon a set of demographic, cognitive and behavioral features, the comparison was done along two dimensions – overall accuracy and fairness – which are important to various stakeholders due to the technical and ethical concerns in practice. The results showed that interpretable models such as logistic regression does not compromise accuracy or fairness compared to more complicated models such as random forest. This implies that it may not be necessary to use complex, black-box models for college success prediction when validated student features are in place. Simpler models can be much more cost-effective for both instructional and institutional stakeholders because they allow for ease of interpretation and efficient computation without sacrificing performance. On the other hand, even in the fairest prediction model, algorithmic biases persist, especially against ethnically minority and academically underprepared students. As such, it is necessary to combine human evaluation and machine predictions in the process of decision-making to ensure that students who are most in need of academic support receive the necessary resources to succeed.

Acknowledgement

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

References

[Barocas *et al.*, 2019] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. 2019.

[Ekowo and Palmer, 2016] Manuela Ekowo and Iris Palmer. The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis. Technical report, New America, 2016.

[Friedler *et al.*, 2019] Sorelle A. Friedler, Sonam Choudhary, Carlos Scheidegger, Evan P. Hamilton, Suresh Venkatasubramanian, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pages 329–338. Association for Computing Machinery, Inc, jan 2019.

[Hutt *et al.*, 2019] Stephen Hutt, Margo Gardner, Angela L. Duckworth, and Sidney K. D'Mello. 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)*, Montréal, Canada, 2019.

[Ishitani, 2006] Terry T. Ishitani. Studying attrition and degree completion behavior among first-generation college students in the united states. *The Journal of Higher Education*, 77(5):861–885, 2006.

[Jayaprakash *et al.*, 2014] Sandeep M. Jayaprakash, Erik W. Moody, Eitel J.M. Lauría, James R. Regan, and Joshua D. Baron. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics*, 1(1):6–47, may 2014.

[Park *et al.*, 2018] Jihyun Park, Renzhe Yu, Fernando Rodriguez, Rachel Baker, Padhraic Smyth, and Mark Warschauer. Understanding Student Procrastination via Mixture Models. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*, pages 187–197, Buffalo, NY, United States, 2018.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, may 2019.

[Yu *et al.*, 2020] 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)*, pages 292–301, 2020.