Yet Another Predictive Model? Fair Predictions of Students' Learning Outcomes in an Online Math Learning Platform

Chenglu Li
University of Florida
Gainesville, USA
li.chenglu@ufl.edu

Wanli Xing*
University of Florida
Gainesville, USA
wanli.xing@coe.ufl.edu

Walter Leite
University of Florida
Gainesville, USA
walter.leite@coe.ufl.edu

ABSTRACT

To support online learners at a large scale, extensive studies have adopted machine learning (ML) techniques to analyze students' artifacts and predict their learning outcomes automatically. However, limited attention has been paid to the fairness of prediction with ML in educational settings. This study intends to fill the gap by introducing a generic algorithm that can orchestrate with existing ML algorithms while yielding fairer results. Specifically, we have implemented logistic regression with the Seldonian algorithm and compared the fairness-aware model with fairness-unaware ML models. The results show that the Seldonian algorithm can achieve comparable predictive performance while producing notably higher fairness.

CCS CONCEPTS

• **Applied computing** → **Interactive learning environments**;
• **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

fair machine learning, predictive analytics, online math learning

ACM Reference Format:
Chenglu Li, Wanli Xing, and Walter Leite. 2021. Yet Another Predictive Model? Fair Predictions of Students' Learning Outcomes in an Online Math Learning Platform. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3448139.3448200

1 INTRODUCTION

Online learning has received great popularity in K-12 and higher education as instruction increasingly migrates from conventional methods [24]. The closing of physical schools worldwide due to the COVID-19 pandemic has further attracted people's attention to online learning [5]. To support online learners at a large scale, extensive studies have adopted machine learning (ML) techniques to analyze students' artifacts automatically and predict their learning outcomes [30, 32, 33]. These techniques empower knowledge discovery at the big data level and characterizations in individual interaction with instructional tools.

However, limited attention has been paid to the fairness of prediction with ML in educational settings [14]. Studies have shown that ML models can be biased by demographic factors such as genders [4, 37]. In the study of Bolukbasil et al., the researchers showed that natural language processing (NLP) models trained with Google News article (a widely adopted dataset for NLP models) demonstrated various stereotypes towards genders. For example, women were highly correlated with stereotyped professions such as homemakers, while men were coupled with computer programmers. Similar biases have also been found in hiring and finance [3, 15], where participants with specific demographic backgrounds were more likely to get employed or be granted loans. No exceptions would be warranted for ML applications in education. For instance, an ML algorithm may mislabel African-Americans as a high risk of failure at nearly twice the rate it mislabeled white students. In another example, Riazy et al. [27] reported that machine learning models could extensively favor students without disabilities if no careful examination of model fairness were conducted.

With the increasingly wide adoption of learning analytics in data-driven decision systems, fairness issues arising from ML algorithms in educational settings should be examined [21]. Moreover, there is an urgent call to help teachers understand students' learning status in online contexts. Therefore, this study aims to explore the possibility of building an early warning system (EWS) that could fairly predict students' learning outcomes in an online math learning platform. Specifically, we have compared a cutting-edge ML algorithm to enhance fairness with three commonly used algorithms for educational predictions. The ML models were built to predict if a student is likely to pass or fail an in-course assessment. This study aims to initialize a foundation work for the EWS that could treat over a million active students on the platform fairly.

2 BACKGROUND

Algorithms might not be biased by nature, but the data fed to algorithms can be [38]. The undesired behaviors of algorithms that reflect humans' hidden values are algorithmic bias [2]. Fairness in algorithms aims to avoid such a bias that creates discriminatory or unjust results [36]. Shin and Park [28] suggested three components of algorithmic fairness: indiscrimination, impartiality, and accuracy. The accuracy in their study is not regarding what users want but avoiding socially and politically incorrect consequences. In the meantime, fairness can be factual or perceived [23], where factual fairness is measured quantitatively with metrics and perceived fairness is perceptions from individuals. This study intends to address the factual fairness with ML algorithms that can limit

*Corresponding Author

LAK21, April 12–16, 2021, Irvine, CA, USA

ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448200

inherent data biases' effects on gender and race while achieving desired predictive accuracy.

Predictive analytics that predicts students' learning outcomes based on behavioral data is a widely used approach in learning analytics research [26]. The viability of using ML models for predictive analytics has been extensively explored in educational settings [31, 34, 35]. For example, Xing and Du [31] used deep learning to predict students' drop-out status temporally based on students' learning management system (LMS) activities. The results suggested that the model could accurately predict students' drop-out status and provide teachers with actionable intelligence. However, to the best of our knowledge, most research with predictive analytics focused on the predictive performance of different algorithms while ignoring their fairness. A few recent studies have demonstrated the effectiveness of algorithm fairness from a predictive analytics perspective [20, 23, 27]. Nonetheless, most of them focused on fairness evaluation of existing algorithms instead of adapting algorithms to achieve better fairness. This study intends to fill the gap by introducing a generic algorithm that can orchestrate with existing ML algorithms while yielding fairer results.

3 METHODS

3.1 Research Context and Data

This study uses upper-level high school students' data on Algebra 2 from Algebra Nation, an online math learning platform developed by Study Edge. There are 14 sections for the course of Algebra 2, with each section having an end-of-section assessment of 10 items. A correction rate above 60% of an assessment is treated as pass. Since this study aims to build a predictive model that can fairly provide students with early warnings, only students who registered in September 2017 and had completed at least one assessment were selected. Data generated by the participants in the academic year of 2017-2018 were extracted because Algebra Nation's use usually follows the academic calendar. The data consists of 2,761 assessments and 717,402 click-stream data entries by 484 students. The click-stream data recorded students' interactions with all the pages and resources (e.g., lecturing videos, reviewing videos, and discussion board).

3.2 Features

Previous studies have suggested that students' behavioral data can offer significant prediction power to their performance [10, 13, 31]. While other studies have adopted feature engineering to achieve better predictive performance, it is out of this study's scope. Therefore, we used a flat feature structure, which was used in most previous studies [31]. Table 1 illustrates what each feature stands for. The combination of assessment id and student id served as one unit and features were computed for each unit, where only behavioral data recorded before a unit were retained. Overall, we used the features to predict a binary outcome on whether a student is about to fail an assessment.

3.3 Equalized Odds

In this study, we used equalized odds as the metric to inform us of models' fairness. While there are several fairness metrics used in previous research, most failed to comprehensively address fairness when it comes to learning. For example, the widely used demographic parity is defined such that the probability of yielding a positive prediction should be the same across protected groups (e.g., grouped with genders or races). However, demographic parity is not ideal in that (1) it only ensures aggregated fairness while ignoring individual fairness and (2) it does not reflect the true tendency of a disadvantaged group [14, 17, 19]. For example, female students might have the tendency to enroll in various courses to find the best fit and thus tend to drop out at the start of courses after final course schedules have been planned [21]. If constrained with demographic parity, a predictive model of students' dropout status in this context would not be able to respect the fact that female students tend to have a higher dropout rate. To remedy the disadvantages of demographic parity, Hardt et al. [17] proposed the metric of equalized odds, which is defined as

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1) \quad (1)$$

$$P(\hat{Y} = 1|A = 0, Y = 0) = P(\hat{Y} = 1|A = 1, Y = 0) \quad (2)$$

, where $\hat{Y}$ is the predicted outcome of the model, Y is the binary outcome from the dataset, and A is the comparison group (e.g., female vs. male). Essentially, equalized odds are satisfied when a model yields equal false positive rate (FPR, Equation 2) and true positive rate (TPR, Equation 1) across groups. In reality, it is difficult to have the same FPR and TPR in different groups. Therefore, we have defined a score function of equalized odds as

$$\gamma(\hat{Y}) = max(|FPR_{A_b} - FPR_{A_c}|, |TPR_{A_b} - TPR_{A_c}|) - \epsilon \quad (3)$$

, where A_b and A_c are groups for comparison and ϵ is a user-defined threshold. Scores equal to or smaller than 0 are preferred since they mean the FPR and TPR of comparison groups are equal to or smaller than the threshold ϵ.

3.4 Seldonian Algorithm

To ensure the predictive model's fairness, we adopted the Seldonian algorithm to achieve better fairness with satisfactory performance. The Seldonian algorithm [29] was created to define a series of procedures to help machine learning algorithms behave desirably. Instead of being a specific machine learning (ML) algorithm, the Seldonian algorithm serves more as a framework to construct fair ML algorithms. Therefore, different tasks (e.g., regression and classification) and existing ML algorithms (e.g., logistic regression) can all orchestrate with the Seldonian algorithm [25, 29].

Figure 1 demonstrates the implementation of the Seldonian algorithm. In the algorithm, we will first define n tuples of constraints (g_i, δ_i), where g_i is a constraint function and $1-\delta_i$ is the confidence level that the constraint will be met. For example, if we aim to build a model that would output an equalized odds score less than 0.02 between females and males with δ being 0.01, we can have one constraint as

$$g(\theta) = max(|FPR_{female} - FPR_{male}|, |TPR_{female} - TPR_{male}|) - 0.02$$

, where FPR is the false positive rate, and TPR is the true positive rate. Undesirable behavior is produced if and only if $g(\theta) > 0$. This means that we are restricting the model to have a close false positive rate and true positive rate of females and males (less than 0.02), and the confidence that an undesirable behavior will appear is 0.99 (1 - 0.01).

Table 1: Features used for predictive models and their description

Feature	Description
Video Watch	Frequency of watching videos
Video Pause	Frequency of pausing videos
Video Play	Frequency of clicking the video play button
Video Seek	Frequency of adjusting video progress
Video Completed	Number of videos completed
Correct Answer Review	Frequency of watching videos for reviewing correct answers
Solution Video Watch	Frequency of watching videos for solution walkthrough
Reviewing Video Watch	Frequency of watching reviewing videos for assessment topics
Discussion Post	Number of discussion posts created
Assessment Order	The order of the current assessment among all taken assessments
History Avg Correct Answer	Avg number of correct answers from previous assessments
Previous Correct Answer	Number of correct answers from the last assessment

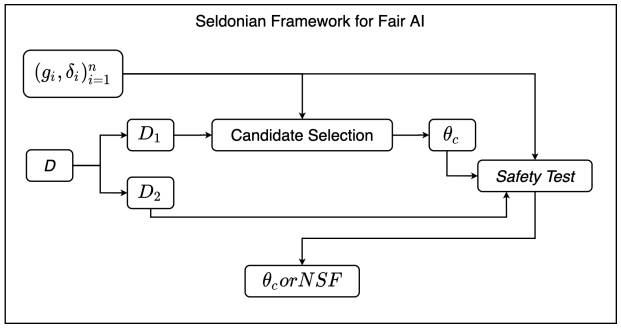

Figure 1: Illustration of the Seldonian algorithm.

D is the training dataset, which will be split into two subsets, D_1 and D_2. Then D_1 will be passed to the Candidate Selection step, which is the model learning phase. The model learning phase shares the main goal with a typical ML algorithm, which is to minimize the model's loss. In the example of binary logistic regression, parameters of the model (e.g., $\beta_0, \beta_1, ..., \beta_n$) will be adjusted to minimize the loss function of

$$J(\theta) = -\frac{1}{m}\sum_{i}^{m}[y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)] \quad (4)$$

, where m is the number of entries of data, y_i is the actual class (0 or 1) of an entry, and $\hat{y}_i$ is the current entry's predicted probability from the logistic regression model [11]. However, what makes the Candidate Selection step different from the learning phase of a typical ML algorithm is that the Candidate Selection step will also try to satisfy a series of fairness constraints defined above while minimizing the model loss within a confidence level. We can use Hoeffding's inequality or bootstrap confidence bound to check if the upper bound of a fairness constraint will be below 0 and thus satisfies a constraint [29]. While confidence bounds computed with Hoeffding's inequality is valid for any distribution, the values tend to be overly conservative [18]. This often requires an impractical amount of data to have a confidence bound that satisfies a constraint. Therefore, we used a bootstrap confidence bound in this study to approximate the bounds with a limited number of data [12]. Bootstrap is a technique commonly used in statistics and machine learning and has been shown to be an effective approach for confidence bounds approximation. Figure 2 shows the procedure of bootstrapping. For example, we can use bootstrap to resample from D_1 with replacement, with each bootstrapped sample having the size of D_1. Statistics will be calculated for each bootstrapped sample (e.g., Equation 3). After repeating the procedure j times, where j needs to be a reasonably large number (e.g., 10,000), we will have j equalized odds scores computed and the equalized odds scores will form the bootstrap distribution, with which we can get approximated confidence bounds. We can then check the upper confidence bound and a penalty for model learning will be given if the upper confidence bound is greater than 0 (recall $g(\theta) > 0$ is undesired).

The Safety Test will use the partitioned dataset D_2 to check whether the parameters candidate θ_c will help a model achieve sufficient confidence such that $g_i(\theta_c) \leq 0$. This step resembles the Candidate Selection step, except that the Safety Test will not adjust model parameters and will solely focus on the constraint checking.

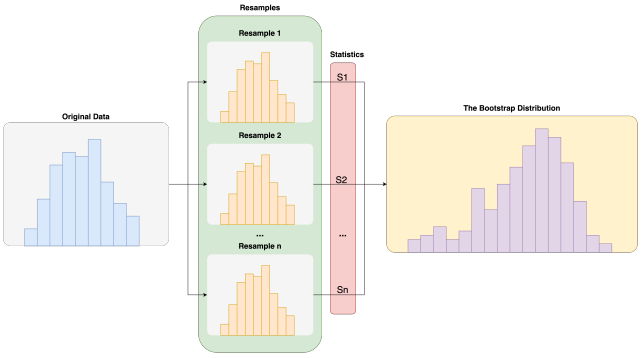

Figure 2: The process to use bootstrap for confidence bounds approximation.

3.5 Comparison Groups

Under-representation of different races and genders have long been reported, especially in the context of STEM education. Therefore,

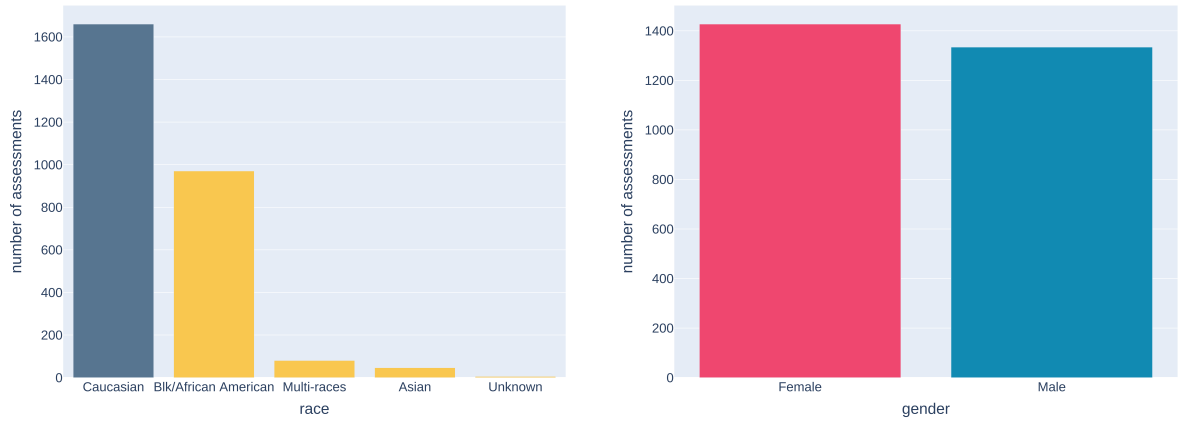

Figure 3: Distribution of race and gender among all the assessments taken.

this study chose to examine prediction fairness in terms of race and gender. Five different values were recorded for races in the dataset: Caucasian, Black or African American, Two or More Races, Asian, and Unknown. The race variable was then coded into a binary variable, with 1 being White students, and 0 being non-White students. Gender was only recorded with binary values: Female and Male. The gender variable was also coded with 1s or 0s, with 1 being male students and 0 being female students. Figure 3 shows the distribution of race and gender among all the assessments taken.

3.6 Model Training and Evaluation

In this study, three popular ML algorithms were used to benchmark with the Seldonian algorithm in terms of prediction fairness of race and gender, and they are logistic regression (LR), support vector machine (SVM), and random forest (RF). We implemented the Seldonian algorithm with logistic regression since its hypothesis function for prediction and loss function for optimization were straightforward to modify. Moreover, studies have suggested LR can achieve comparable performance with other ML algorithms [8]. SVM was selected because of its robust binary classification performance and has been widely used in educational settings [9, 22]. RF was chosen because of its superior performance due to its ensemble features [6, 16, 22].

The data were split into a training (70%) and testing (30%) dataset to better evaluate these algorithms' prediction performance. To find optimized parameters, we conducted a grid search along with 10-fold cross-validation on the benchmark algorithms; Model hyperparameters were searched within a defined space and performances were measured from multiple rounds of cross-validation and averaged over the iterations. For the Seldonian algorithm, we defined 2 pairs of constraints: (1) a minimum loss with 95% confidence to avoid overfitting and (2) a maximum equalized odds score with 95% confidence to avoid unfair behaviors. Different values for the constraints were examined to achieve the ideal results. After model training, model performance was evaluated on the testing dataset with metrics such as accuracy, F-measure, and area under the receiver operating characteristic curve (AUC).

Equalized odds scores were then calculated in terms of race and gender, respectively, on the benchmark models and the Seldonian algorithm. The equalized odds scores were defined as the maximum between the FPR difference and TPR difference between groups (e.g., male vs. female, white vs. non-white students). This is the same as Equation 3 with ϵ set to 0. Specifically, equalized odds scores of the benchmark models were computed first. Then thresholds ϵ for the constraint (2) of the Seldonian algorithm were selected to see if better equalized odds scores can be achieved.

4 RESULTS

4.1 Prediction Performance

Table 2 shows the model performance on predicting students' pass or fail status of the next assessment. The best-performed models are bolded. "*" suggests comparable performance from the Seldonian algorithm. SA stands for the Seldonian algorithm, with the underscored values being the maximum equalized odds required in the constraints. The predictive performance is the same for benchmark models (SVM, RF, and LR) in different comparison groups. Because these models' learning goal is to reduce the model loss and do not get affected by which group is used for comparison. However, for the Seldonian algorithm, predictive performance varies since the learning goal is more than minimizing the model loss but also satisfying the fairness constraints.

For the race, RF is the best performed model in terms of prediction, with an F1 score of 0.81 and an AUC score of 0.82, followed by SVM, whose F1 score is 0.79 and AUC score is 0.81. However, though slightly less performative, two out of the four variants of the Seldonian algorithm can achieve comparable performance in terms of accuracy, F1, and AUC. In general, as we require smaller equalized odds in the fairness constraint, the Seldonian algorithm generates more unsatisfactory performance. It is worth noting that the Seldonian algorithm can achieve similar and even slightly better performance than its direct counterpart, LR.

Similar when using race as the comparison group, for gender, RF achieves the highest F1 (0.81) and AUC (0.82) scores, and the Seldonian algorithm can still yield comparable metrics. However,

Table 2: Models Performance Comparison

Comparison Group	Model	Accuracy	F1	AUC
Race	SVM	0.81	0.79	0.81
	RF	**0.82**	**0.81**	**0.82**
	LR	0.81	0.79	0.74
	$SA_{race_{odds0.25}}$*	0.81	0.78	0.79
	$SA_{race_{odds0.20}}$*	0.80	0.77	0.76
	$SA_{race_{odds0.15}}$	0.78	0.72	0.78
	$SA_{race_{odds0.10}}$	0.77	0.69	0.74
Gender	SVM	0.81	0.79	0.81
	RF	**0.82**	**0.81**	**0.82**
	LR	0.81	0.79	0.74
	$SA_{gender_{odds0.05}}$*	0.79	0.78	0.77
	$SA_{gender_{odds0.01}}$*	0.79	0.79	0.77

the Seldonian algorithm outputs no solution found (NSF) in this case, meaning not enough confidence level can be given using the learned parameters to ensure the fairness. This is why the actual equalized odds scores were higher than the thresholds set in constraints (see Figure 4 right).

4.2 Prediction Fairness

Figure 4 shows models' equalized odds comparisons in terms of race and gender. For the race, SVM has an equalized odds of 0.2371, that of LR is 0.3637, and the best-performed RF has an equalized odds of 0.2626. While for the Seldonian algorithm, the variant with the most comparable performance has an equalized odds of 0.2082 (F1 = 0.78), whose maximum equalized odds constraint was set to 0.25, followed by another competitive variant (equalized odds = 0.1794, F1 = 0.77). While the other two variants have much smaller equalized odds (0.1194 and 0.0673) than the benchmark models, their predictive performances are also less ideal, with F1 being 0.72 and 0.69, respectively. As for gender, SVM has an equalized odds of 0.2431, LR has an equalized odds of 0.3296, and RF has an equalized odds of 0.0675. The two variants of the Seldonian algorithm have equalized odds of 0.0726 (F1 = 0.79) and 0.0856 (F1 = 0.78), respectively. In general, models with the Seldonian algorithm tend to have lower equalized odds while retaining competitive predictive performance. When compared with the fairness-unaware LR, the Seldonian algorithm can predict more fairly with similar or better precision.

5 DISCUSSION AND FUTURE DIRECTIONS

This study shows that the use of the Seldonian algorithm can achieve both desirably fair and predictive performance. Although the choice of thresholds in the fairness constraint seems arbitrary, there exists the potential to incorporate such values as additional hyper-parameters of ML models. Researchers can then adjust the fairness hyper-parameter to balance the accuracy and fairness trade-off. The gains of fairness with the Seldonian algorithm are notably high, especially compared to its direct counterpart LR. For SVM and RF, Seldonian algorithm variants can still achieve comparable predictive performance while being fairer when it comes to race. However, for gender, models with the Seldonian algorithm had slightly higher equalized odds scores compared with RF, suggesting a less fair result. The discrepancy of achieved fairness in race and gender can be explained as follows. The fairness-unaware baselines solely try to minimize training loss. When there is no conflict between loss minimization and fairness, high-performing solutions can also be fair (e.g., RF in gender). The finding aligns with the study by Metevier et al. [25]. Importantly, while the baselines might be fair in some cases, unlike the Seldonian algorithm, these approaches do not provide fairness guarantees. However, when there is a conflict between loss minimization and fairness, the Seldonian algorithm might start to shine. Figure 5 shows the number of passed assessments in different races and genders. For the race, the proportion of passed assessments of White students are much higher than that of non-White students. In contrast, the proportions are similar for female and male students. The skewed data in the race might have caused conflicts between loss minimization and fairness such that high-performing fairness-unaware baseline models failed to retain fairness.

This study is exploratory by nature and intends to start a foundation work for further endeavors. In the future, we plan to incorporate over-sampling techniques such as SMOTE [7]. SMOTE can help with the data asymmetry in the race synthetically, and we can understand if the Seldonian algorithm can still achieve better fairness with SMOTE enhanced data. Meanwhile, we plan to adopt explainable AI (XAI) techniques to compare fairness-unaware LR and Seldonian algorithm to understand why better fairness can be achieved when predictive performance is almost the same. Last, we will compare fair algorithms (e.g., reductions approach [1]) developed before the Seldonian algorithm to understand the affordances of different fair algorithms.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification *(Proceedings of*

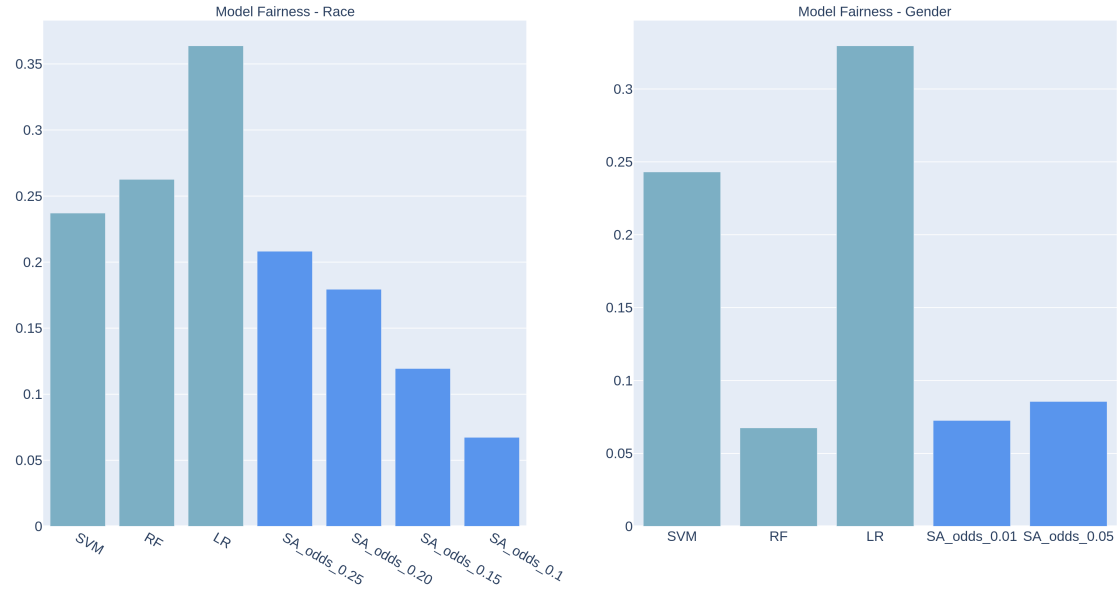

Figure 4: Models' equalized odds comparisons in terms of race and gender. In this graph, lower bars are fairer.

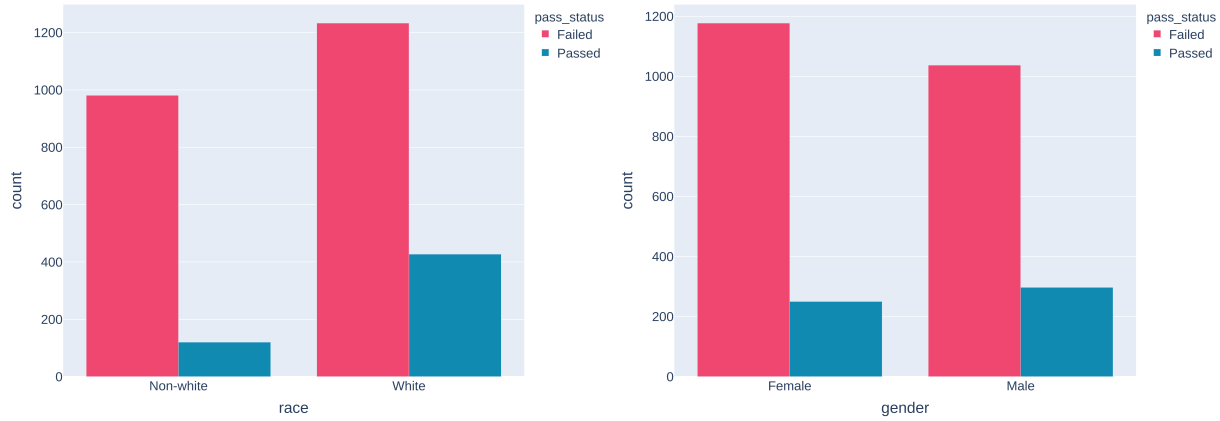

Figure 5: Number of passed assessments in different races and genders.

Machine Learning Research, Vol. 80), Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

[2] David Beer. 2017. The social power of algorithms.

[3] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. 149–159.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[5] Richard Allen Carter Jr, Mary Rice, Sohyun Yang, and Haidee A Jackson. 2020. Self-regulated learning in online learning environments: strategies for remote learning. *Information and Learning Sciences* (2020).

[6] Vo Thi Ngoc Chau and Nguyen Hua Phung. 2013. Imbalanced educational data classification: An effective approach with resampling and random forest. In *The 2013 RIVF International Conference on Computing & Communication Technologies-Research, Innovation, and Vision for Future (RIVF)*. IEEE, 135–140.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[8] Evangelia Christodoulou, Jie Ma, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, and Ben Van Calster. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology* 110 (2019), 12–22.

[9] Kwok Tai Chui, Dennis Chun Lok Fung, Miltiadis D Lytras, and Tin Miu Lam. 2020. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior* 107 (2020), 105584.

[10] Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S McNamara, and Ryan S Baker. 2016. Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the sixth international conference on learning analytics & knowledge*. 6–14.

[11] Stephan Dreiseitl and Lucila Ohno-Machado. 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics* 35, 5-6 (2002), 352–359.

[12] Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American statistical Association* 82, 397 (1987), 171–185.

[13] Mi Fei and Dit-Yan Yeung. 2015. Temporal models for predicting student dropout in massive open online courses. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 256–263.

[14] 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*. 225–234.

[15] Vivian Giang. 2018. The potential hidden bias in automated hiring systems. *The Future of Work. FastCompany. May 8th* (2018).

[16] Julie Hardman, Alberto Paucar-Caceres, and Alan Fielding. 2013. Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm. *Systems Research and Behavioral Science* 30, 2 (2013), 194–203.

[17] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[18] Wassily Hoeffding. 1994. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*. Springer, 409–426.

[19] Wen Huan, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fairness through Equality of Effort. In *Companion Proceedings of the Web Conference 2020*. 743–751.

[20] 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. *International Educational Data Mining Society* (2019).

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

[22] Jiajun Liang, Jian Yang, Yongji Wu, Chao Li, and Li Zheng. 2016. Big data application in education: dropout prediction in edx MOOCs. In *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*. IEEE, 440–443.

[23] Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. Implications of AI (un-)fairness in higher education admissions: the effects of perceived AI (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 122–130.

[24] Richard E Mayer. 2019. Thirty years of research on online learning. *Applied Cognitive Psychology* 33, 2 (2019), 152–159.

[25] Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S Thomas. 2019. Offline Contextual Bandits with High Probability Fairness Guarantees. In *Advances in Neural Information Processing Systems*. 14922–14933.

[26] Alejandro Peña-Ayala. 2014. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications* 41, 4 (2014), 1432–1462.

[27] Shirin Riazy and Katharina Simbeck. 2019. Predictive Algorithms in Learning Analytics and their Fairness. *DELFI 2019* (2019).

[28] Donghee Shin and Yong Jin Park. 2019. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98 (2019), 277–284.

[29] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366, 6468 (2019), 999–1004.

[30] Wanli Xing, Xin Chen, Jared Stein, and Michael Marcinkowski. 2016. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in human behavior* 58 (2016), 119–129.

[31] Wanli Xing and Dongping Du. 2019. Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research* 57, 3 (2019), 547–570.

[32] Wanli Xing, Sean Goggins, and Josh Introne. 2018. Quantifying the effect of informational support on membership retention in online communities through large-scale data analytics. *Computers in Human Behavior* 86 (2018), 227–234.

[33] Wanli Xing, Rui Guo, Eva Petakovic, and Sean Goggins. 2015. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior* 47 (2015), 168–181.

[34] Wanli Xing, Bo Pei, Shan Li, Guanhua Chen, and Charles Xie. 2019. Using learning analytics to support students' engineering design: the angle of prediction. *Interactive Learning Environments* (2019), 1–18.

[35] Wanli Xing, Hengtao Tang, and Bo Pei. 2019. Beyond positive and negative emotions: Looking into the role of achievement emotions in discussion forums of MOOCs. *The Internet and Higher Education* 43 (2019), 100690.

[36] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 1–6.

[37] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).

[38] Franziska Zimmer, Katrin Scheibe, Mechtild Stock, and Wolfgang G Stock. 2019. Fake news in social media: Bad algorithms or biased users? *Journal of Information Science Theory and Practice* 7, 2 (2019), 40–53.