Assessing the Fairness of Graduation Predictions

Henry Anderson
University of Texas at Arlington
henry.anderson@uta.edu

Afshan Boodhwani
University of Texas at Arlington
afshan.boodhwani@uta.edu

Ryan S. Baker
University of Pennsylvania
rybaker@upenn.edu

ABSTRACT

Predictive and data-intensive modeling has rapidly gained prominence in many research fields over the past decade. In recent years, the fairness of analytical models has become an increasingly important question for researchers: might a model systematically underperform on certain demographics? Might its application have different impacts across different demographic groups? Recently, these questions have found their way into educational research as well, where predictive and statistical modeling is used for such purposes as predicting course completion, assisting university recruitment, and proactively offering assistance to students. If such models are potentially unfair, students may remain underserved or suffer potential harm as a result. In this paper, we demonstrate two post-hoc assessments of fairness, applied to existing models predicting student graduation. Our assessments are intended to check for, rather than proactively prevent, algorithmic bias in predictive models. The first assessment investigates whether a selected model is *equitable:* does its performance systematically differ for members of different demographic groups? The second investigates whether the decision to use one model for all individuals in a given dataset is *optimal:* does using one model for all students come at a significant loss in per-group accuracy? By assessing fairness systematically, we hope to reduce to the risk of inequities from predictive analytics in education.

1. INTRODUCTION

As educational research incorporates more automated tools for analyzing and predicting such factors as student behaviors and educational outcomes, the field must grapple with the ethical question of algorithmic fairness. Do our algorithms, powerful as they may seem, encode and reinforce existing societal inequalities? Are the predictive tools we use systematically less accurate for some demographics, and does the use of these tools risk harming members of those demographics?

In this work, we assess a set of machine learning models, trained to predict student graduation at a public, four-year university, using several definitions of fairness. Our analysis centers on race and gender, but in principle, can be extended to include age, international status, or other categorical variables.

Our investigations are driven by two main research questions: RQ 1) Are the models *equitable?* Do they perform considerably worse on students of particular genders or ethnicities? RQ 2) Are the models *optimal?* Could we achieve substantially better accuracy by using models specific to particular genders or ethnicities?

We investigate these questions through comparisons of false positive rates, false negative rates, and overall accuracy measures, since these each represent different potential impacts of these models' use. The primary intended use of the models is to aid university advisers, and faculty, and administrators in structuring student support. This intended use of the models informs our particular definitions and questions around the fairness of these models.

2. RELATED WORK

Researchers have utilized a wide range of approaches to defining fairness. Dwork et al. [4] define fairness as *similar individuals receive similar predictions* (individual fairness). Hardt et al. [7] measures *equalized odds*, which penalizes models for performing well only on the majority of data points, and *equal opportunity,* which requires parity between groups' predictions only where the ground truth is an "advantaged" outcome (e.g., "was admitted to college," "received a promotion"). Feldman et al. [5] propose measures of *group fairness,* where the distribution of errors and impacts should not vary across groups. There are many approaches to controlling for these, and other, fairness measures, as discussed in, e.g., [2, 9].

While the exact definitions and approaches to fairness are quite varied, a common thread is that there are no unambiguously "correct" definitions of fairness, or clear ways to control for all aspects of it. The particular definitions and measures depend on the specific data, analysis, and use case.

3. FAIRNESS FOR OUR MODELS

We select our measures of fairness based on the intended applications of the models, and their potential impact. The models are intended to inform and assist advisers, mentors, faculty, and university administrators in making decisions

about student intervention, student support, and university policies. We focus on the advising and mentoring applications, as this is is the first intended use for the models.

In this setting, we anticipate a difference in the impacts of false negative and false positive errors. False negatives (students on track to graduate who are identified as likely to not graduate) are likely to result in additional contact with advisers, and additional assistance being made available. There is little harm to the student. False positives (students not on track to graduate, but who are predicted as likely to graduate) mean a student may not receive assistance from advisers. The potential harm is much greater. We thus use both false positive and false negative rates as metrics for fairness, per RQ 1, but we are more concerned with possible disparities in the false positives.

The overall accuracy of the models may also vary across our populations. Large variations put certain populations at higher risk of being systematically underserved due to lower-quality predictions, and thus less reliable information being presented to advisers and mentors. Therefore we test if selecting one model, trained on all students, is optimal, per RQ 2, by comparing its overall accuracy to the population specific models.

4. DATA

In this paper, we assess the fairness of models predicting whether a student will graduate within six years, initially presented in a conference poster [1]. In this section, we briefly describe the data on which those models were trained; more details are available in the original presentation.

The dataset utilized was from a large, publicly-funded, R1 research university in the southern United States. It contains data on 14,706 first time in college (FTIC) undergraduate students, all of whom were admitted in Fall semesters between 2006-2012 (inclusive), and were enrolled full-time. The data contains one entry per student, summarizing their first three enrolled semesters (Fall, Spring, and Summer). The final feature set covers academic performance (e.g. GPA, credit hours completed), financial information (e.g. scholarships, unmet need), pre-admission information (e.g. SAT/ACT scores), and extra-curricular activities (e.g. Greek Life, athletics). A student's first year has been shown to be an important period for determining a student's likelihood of dropping out [8], which while not quite the inverse of graduation, is a closely related outcome.

Tables 1 and 2 shows basic descriptive statistics for the dataset. Some of the populations have very small Ns. We include these populations in the fairness analyses for completeness, but we caution against drawing meaningful conclusions from them. Similarly, we caution against drawing conclusions for the Multiple Ethnicities and Foreign populations, since these are "catch-all" labels, and represent extremely diverse groups of students.

Table 1: Basic descriptive statistics of the dataset, by self-reported ethnicities.

Population	N	Graduation rate
American Indian	44	27.27%
Asian	2091	61.74%
African American	2092	38.48%
Foreign	296	52.03%
Hispanic/Latino	3805	43.97%
Multiple Ethnicities	499	48.30%
Hawaiian/Pacific Islander	22	54.55%
Ethnicity Not Specified	85	35.29%
White	5772	44.51%
TOTAL	14706	46.15%

Table 2: Basic descriptive statistics of the dataset, by self-reported gender.

Population	N	Graduation rate
Female	7613	50.06%
Male	7092	41.96%
Gender Unknown	1	0.00%
TOTAL	14706	46.15%

5. METHODOLOGY

5.1 Model Building

We investigate the fairness of five separate models, each trained on our dataset to predict whether a student will graduate within 6 years of first enrolling at the university. Early versions of these models (except Random Forest) have been presented in poster form [1]: linear kernel Support Vector Machines (SVM), Decision Trees, Random Forests, Logistic Regressions, and scikit-learn's Stochastic Gradient Descent classifier (SGD[1]).

Each model was trained on 80% of the full dataset, with 20% held out for testing. The train-test split was conducted such that the proportions of students in each demographic category were as close as possible across the folds, and the within-group and overall graduation rates were as similar as possible. Model parameters were selected using 5-fold cross-validation within the training set. Ethnicity and gender features were omitted when training the models.

5.2 Assessing Equity

We measure the equity of our models via their false positive and false negative predictions on the held-out testing set. Each comparison is made using a one-versus-rest approach: *Male* versus *non-Male*, *White* versus *non-White*, etc. To compare false positive and false negative rates, we assign each student a label of 1 (indicating a false positive/false negative prediction) or 0 (true positive/true negative). The populations are compared using a χ^2 test on the resulting binary-valued vectors, with Benjamini & Hochberg's post-hoc correction [3] applied within each combination of population and fairness metric, across algorithms.

[1]scikit-learn's `SGDClassifier` model uses gradient descent to construct a hyperplane classifier. We refer to this classifier, not the general numeric optimization method, in this paper. See the scikit-learn documentation for further details: `https://scikit-learn.org/stable/modules/sgd.html`

Table 3: Results of the χ^2 tests on false negatives and false positives, by demographic. Differences shown are the overall rates for students in the listed demographic, minus the overall rates for students not in the listed demographic. Benjamini & Hochberg [3] corrected p-values are in parentheses; p-values less than 0.05 are in bold. * = $N < 500$.

	Population	DT	RF	SVM	LR	SGD
False Negatives	American Indian*	0.051 (0.496)	0.036 (0.598)	0.055 (1.652)	0.055 (0.559)	0.055 (0.826)
	Asian	0.014 (0.311)	0.000 (0.997)	0.013 (0.880)	0.013 (0.239)	0.013 (0.440)
	African American	0.011 (0.729)	0.025 (0.188)	-0.001 (1.193)	0.002 (1.423)	-0.001 (0.954)
	Foreign*	0.009 (0.731)	-0.022 (2.226)	0.013 (1.458)	0.013 (0.744)	0.013 (0.972)
	Hispanic/Latino	**0.020 (0.046)**	**0.030 (0.007)**	0.010 (0.223)	0.013 (0.181)	0.011 (0.186)
	Multiple Ethnicities*	-0.001 (0.943)	-0.006 (0.992)	0.023 (1.066)	0.023 (0.555)	0.013 (0.801)
	Hawaiian/Pacific Islander*	-0.041 (1.613)	-0.055 (2.950)	-0.036 (0.831)	-0.037 (1.105)	-0.036 (0.664)
	White	**-0.030 (<0.001)**	**-0.035 (<0.001)**	**-0.020 (0.006)**	**-0.023 (0.002)**	**-0.020 (0.005)**
	Female	-0.005 (0.565)	0.000 (0.966)	**-0.019 (0.024)**	**-0.017 (0.030)**	**-0.017 (0.025)**
	Male	0.006 (0.560)	0.000 (0.974)	**0.019 (0.023)**	**0.017 (0.029)**	**0.017 (0.025)**
False Positives	American Indian*	0.103 (0.366)	0.122 (1.290)	0.105 (0.438)	0.108 (0.562)	0.110 (0.810)
	Asian	-0.024 (1.139)	0.013 (0.496)	-0.024 (0.581)	-0.018 (0.448)	-0.018 (0.592)
	African American	0.001 (0.956)	-0.037 (0.245)	-0.018 (0.602)	-0.021 (0.722)	-0.012 (0.653)
	Foreign*	0.060 (1.078)	0.013 (0.967)	0.046 (0.564)	0.049 (0.778)	0.001 (0.985)
	Hispanic/Latino	-0.021 (0.304)	-0.019 (0.247)	-0.019 (0.222)	-0.028 (0.346)	-0.023 (0.329)
	Multiple Ethnicities*	-0.024 (0.870)	-0.024 (1.253)	-0.032 (2.005)	-0.019 (0.764)	-0.017 (0.656)
	Hawaiian/Pacific Islander*	0.030 (0.861)	0.049 (3.794)	0.032 (1.059)	0.035 (1.392)	0.037 (2.057)
	White	**0.029 (0.038)**	**0.032 (0.023)**	**0.040 (0.012)**	**0.043 (0.010)**	**0.039 (0.009)**
	Female	-0.018 (0.309)	-0.008 (0.711)	-0.021 (0.648)	-0.020 (0.359)	-0.004 (0.753)
	Male	0.019 (0.300)	0.008 (0.698)	0.021 (0.627)	0.020 (0.347)	0.004 (0.741)

Table 4: Model scores, trained on all students, evaluated against the held-out testing set. Values in parentheses are the standard error, calculated as in [6].

Model	AUC
Decision Tree	0.798 (0.008)
SVM	0.805 (0.008)
Logistic Regression	0.807 (0.008)
Random Forest	0.800 (0.008)
SGD	0.814 (0.008)

5.3 Assessing Optimality

To assess the optimality of our models (RQ 2), we compare how much accuracy is gained or lost by building separate models for each population, based on AUC ROC scores. For each population in our dataset, we compare the AUC ROC scores (calculated only on the test set) of the models when trained on all students to the AUC ROC scores of the models when trained only on that population.

6. RESULTS

6.1 Model Performance

The overall metrics, evaluated against all students in the testing set, are reported in Table 4. The models achieve consistently good performance, with high AUC ROC and F1 scores. These numbers are a good baseline of performance; if the models' performance drops or rises considerably for any demographic, that can be taken as an indication of algorithmic bias and the need for population-specific models. AUC ROC standard errors are computed according to Hanley & McNeil's method [6].

6.2 RQ 1: Model Equity

Table 3 shows the results of our investigation into RQ 1. The majority of the comparisons do not show significance at $p = 0.05$ after correction. Notable exceptions are White students, with consistently higher false positive and lower false negative rates across all models; Hispanic/Latino students, with consistently higher false negative rates for the tree-based models; and Male students, who have consistently higher false negative rates for hyperplane-based models (SVM, SGD, and Logistic Regression).

Since the impacts of false positives are likely to be more harmful than false negatives, we find the consistently higher false positive rate for White students to be more noteworthy than the false negative results.

6.3 RQ 2: Model Optimality

Table 5 shows the AUC scores and AUC standard errors on each population for two models: one trained on the entire dataset and tested only on students in the listed demographic group ("Whole Population Model"), and one trained only on students in the listed demographic group ("Population Specific Model"). The only instance where the performance differed by more than the standard error is the Logistic Regression model for Asian students, which saw a *decrease* in performance when using the population-specific model. This indicates that the current models are optimal in terms of population-specificity: population-specific models do not gain appreciable predictive power for any population in the dataset.

Further, this indicates that the trends identified by the models generalize across populations in the dataset, and thus, the models' performance on all students benefits from access to

Table 5: The results of comparing the AUC ROC scores for models trained on all students ("Whole Population Model," abbreviated "WPM") versus just one demographic ("Population Specific Model," abbreviated "PSM"). AUCs with non-overlapping standard error (SE) intervals are in bold.

		DT		RF		SVM		LR		SGD	
Population		PSM	WPM	PSM	WPM	PSM	WPM	PSM	WPM	PSM	WPM
African American	AUC	0.788	0.796	0.799	0.804	0.818	0.830	0.805	0.830	0.799	0.830
	SE	0.024	0.024	0.023	0.023	0.022	0.022	0.023	0.022	0.023	0.022
American Indian	AUC	0.857	0.661	0.679	0.661	0.464	0.661	0.589	0.661	0.589	0.661
	SE	0.135	0.182	0.180	0.182	0.187	0.182	0.188	0.182	0.188	0.182
Asian	AUC	0.747	0.762	0.729	0.744	0.746	0.769	**0.710**	**0.766**	0.753	0.769
	SE	0.023	0.023	0.024	0.024	0.024	0.023	**0.025**	**0.023**	0.023	0.023
Female	AUC	0.795	0.800	0.803	0.798	0.814	0.815	0.812	0.816	0.798	0.811
	SE	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
Foreign	AUC	0.770	0.712	0.763	0.796	0.677	0.729	0.658	0.729	0.620	0.781
	SE	0.060	0.066	0.061	0.057	0.068	0.064	0.070	0.064	0.072	0.059
Hispanic/Latino	AUC	0.800	0.799	0.786	0.790	0.810	0.814	0.798	0.819	0.806	0.819
	SE	0.017	0.017	0.017	0.017	0.016	0.016	0.017	0.016	0.016	0.016
Male	AUC	0.777	0.793	0.796	0.801	0.787	0.791	0.792	0.794	0.790	0.804
	SE	0.013	0.012	0.012	0.012	0.013	0.013	0.012	0.012	0.013	0.012
Multiple Ethnicities	AUC	0.816	0.818	0.812	0.826	0.809	0.807	0.807	0.797	0.723	0.807
	SE	0.043	0.043	0.043	0.042	0.044	0.044	0.044	0.045	0.051	0.044
Hawaiian/Pacific Islander	AUC	0.667	0.750	0.750	0.750	0.833	0.750	0.833	0.750	0.417	0.750
	SE	0.265	0.239	0.239	0.239	0.201	0.239	0.201	0.239	0.287	0.239
White	AUC	0.803	0.805	0.806	0.809	0.803	0.800	0.800	0.802	0.805	0.805
	SE	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013

the full population at training time. However, since the models primarily identified GPA and student credit hours obtained as the most important predictors [1], this trend may be specific to these variables.

7. DISCUSSION AND FUTURE WORK

We have demonstrated an approach to assessing fairness that derives definitions of fairness directly from the use cases of the models in question. We find that our models are not perfectly equitable, as the term is defined in RQ 1. However, the differences are generally small (under 5%), and with the important exception of White students, are not consistent across all models. We encourage any end-users of models that display some unfair tendencies to be cautious and mindful of the potential impact. These differences are relatively small, but it has still not been established what level of unfairness should be considered acceptable in a model with real-life implications. Perfect fairness is ideal, but difficult to achieve. Our models are, though, very optimal across groups, in terms of our definition in RQ 2. No model, or population, saw a meaningful change in per-group performance when trained only on one population.

The most important avenue for future work on this subject is investigating how the implementation of models such as these (e.g. making them available to advisers) will affect student outcomes, and whether the slight unfairness observed here will translate to real-world differences. The assessment of fairness we have performed is an attempt to pre-empt such effects, but is not a substitute for directly measuring them.

8. REFERENCES

[1] H. Anderson, A. Boodhwani, and R. S. Baker. Predicting graduation at a public R1 university. In *Proceedings of the 9th International Learning Analytics and Knowledge Conference*, 2019.

[2] S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[6] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

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

[8] V. Tinto. Research and practice of student retention: What next? *Journal of College Student Retention: Research, Theory & Practice*, 8(1):1–19, 2006.

[9] I. Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.