Fairness and Bias in Prediction of Students' Dropout in Higher Education

Zhanyang Sun; Haoze Zhu;
Viterbi School of Engineering, University of Southern California, Los Angeles, California 90089, USA

(Dated: May 5, 2025)

Students' dropout has always been a critical issue for higher education, since it usually causes some long-term socioeconomic consequences. While many prior studies have spotted many potential causes and some even proposed predictive modeling for dropout rates, few have addressed the fairness and biases within these models. In this project, we focus on demographic variables including age, race and gender, and investigate the way they affect the prediction accuracy and biases of the models. We choose deep learning models including Multi-Input Neural Network, TabNet, and TabTransformer, with proper preprocessing techniques as well. In addition to the conventional evaluation metrics like accuracy, precision, recall, and F1 score, we use Statistical Parity Difference and Equalized Opportunity to evaluate the fairness of the models. We aim to balance the model's performances and the fairness across protected groups, while evaluating any potential systematic biases. After implementing the models, we find SMOTE improved the model's performance metrics including recall and accuracy. This could potentially mean that the model's performance is improved when the data imbalance problem is addressed.

Keywords: Machine Learning, Natural Language Processing, Fairness, Bias

I. INTRODUCTION

Student Dropout is a persistent challenge in higher education, which brings high-reaching consequences. High dropout rate not only produces great disruptions to individual's growth, but also brings significant costs for the entire households, or even the society. To address this issue, colleges and universities have been developing many early-risk warning system to identify at-risk students. Many predictive modeling techniques have been implemented to better assist with this early-warning mechanism. Such systems was developed by good intentions to enable early interventions, but also raised many ethical concerns regarding the fairness and biases. Since these models usually incorporate demographic variables like gender and age, it inadvertently invoke inequalities towards many disadvantaged groups.

II. RELATED WORKS

Some researches have already found probable causes for students' dropout in higher education. Paura et al. found that students' lack of interest in engineering and ignorance of secondary school knowledge were the main causes for that.[1] Many other researches on this matter have mainly focused on qualitative methods, with few implementing quantitative approaches. We find that even among those who are tackling this matter from a quantitative perspective, large language models are not widely used[2]. Researchers have used various models to predict the students' dropout rate in higher education. Martins et al. pointed out that such data have serious imbalance problems, and boosting algorithms tend to work better than traditional methods.[3] However, they fail to recognize the potential bias and inequalities in these models. Likewise, Efthyvoulos et al. focused on time-series methods such as RNN and LSTM.[4] Their paper mainly focused on the comparison of different models, without really arguing if their approaches have underlying biases.

Admittedly, it is important that we should focus on the models themselves. It is also clear that the fairness of these models should be carefully evaluated. In this project, we want to explore whether the inclusion of age, race, and other socioeconomic factors will affect the accuracy and fairness of student dropout prediction models.

III. METHODS

In this project, our dataset, 'Predict Students' Dropout and Academic Success', is obtained from the UC Irvine Machine Learning Repository[5]. The dataset was created in the study by Martins et al[3] to explore early prediction of student performance using various machine learning techniques in the hope of reducing academic dropout and failure in higher education. Their research concludes that boosting algorithms respond better to the classification task than standard methods, such as logistic regression, support vector machine, decision tree, and random forest. They focused on mitigating class imbalance issues using methods like SMOTE and ADASYN. But even using boosting models with SMOTE has failed to correctly classify the minority classes in the dataset.

The dataset underpins a three-category classification task that differentiates between students who drop out, remain enrolled, or graduate by the end of the normal duration of their course. Our approach begins with a exploratory data analysis to generate correlation heatmaps, use pair plots, and perform statistical tests to identify highly correlated features. Afterwards, we preprocessed the dataset by normalizing continuous variables and encoding categorical features. Our choice of deep learning models for the classification task includes Dense Neural Network (DNN), Multi-Input Neural Network(MINN), and TabNet. To evaluate deep learning models, we used accuracy, precision, recall, and F-1 score as metrics for model performance evaluation and calculated Statistical Parity Difference (SPD) and Equalized Opportunity (EO) as fairness metrics. The deep learning models will be trained on the dataset (using an 80/20 training-test split). After evaluating three models to determine whether there is bias against minority classes, we applied three debiasing methods across three models. We use feature selection and introduce protected variables to avoid proxy bias. We employed a Random Forest classifier to identify the most influential features. First, we calculated Random Forest feature importances using Gini impurity reduction. Then, we computed permutation importance scores by randomly shuffling feature values and measuring the resulting decrease in model performance. After combining two steps to create a comprehensive ranking system, we applied protected features on family information and age to protect privacy, creating a smaller list of features for deep learning model training. The second debiasing method is applying SMOTE in our model to handle the imbalance of certain target groups in the dataset. Furthermore, we developed a robust adversarial training approach to improve model fairness. For DNN model, we generated adversarial examples using gradient approximation. For MINN model, we added controlled Gaussian noise (mean=0, std=0.05) to training data. For TabNet model, we applied similar noise-based perturbation techniques. In summary, we trained baseline models, SMOTE applied models, and adversarial trained models for three deep learning models with full features. And we trained baseline models, SMOTE applied models, and adversarial trained models again for three deep learning models with selected features.

IV. RESULTS

In this section, we systematically present the results for our approach, including a review of the exploratory data analysis results, performances of our deep learning models, and an evaluation of fairness across protected groups.

A. Exploratory Data Analysis

The dataset comprises 4424 instances, with each instance representing a student. The dataset also contains 36 features, including information known at the time of student enrollment (academic path, demographics, and social-economic factors) and students' academic performance at the end of the semester. There are three possible outcomes for the target variable: graduation (49.93%), dropout (32.12%), and continued enrollment (17.95%).

Age is a key demographic variable and it ranges from 17 to 70 years. Also, we found sampling bias in the gender distribution. Gender distribution is skewed toward female group (64.8% vs. 35.2%), as shown in Figure 1. In addition, the target class shows a significant imbalance, with dropouts only constitutes to about 32.1% as demonstrated in Figure 2. It could be caused by sampling bias in gender since female students show higher graduation rate compared to male students.

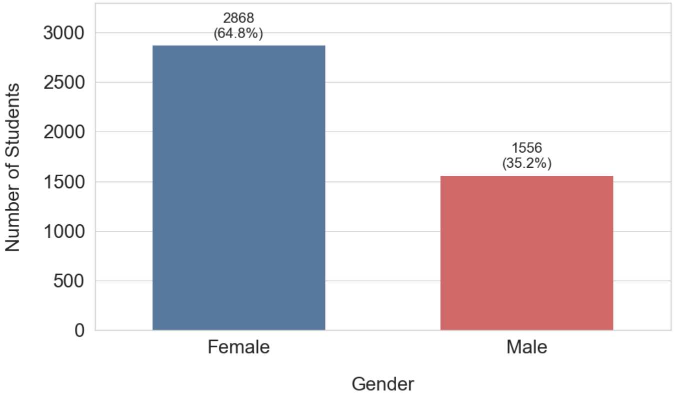

FIG. 1. Gender Distribution

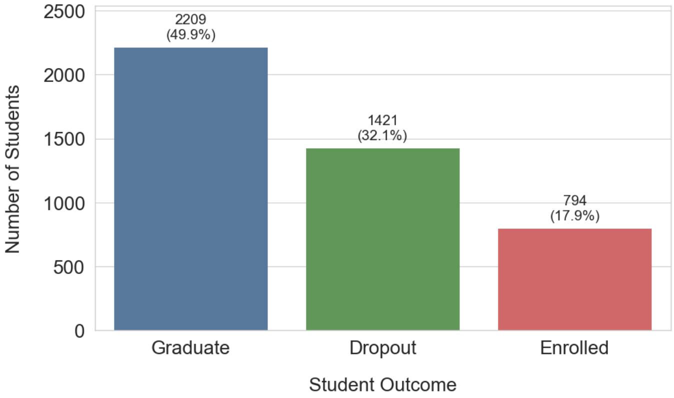

FIG. 2. Target Class Distribution

The proportion of international students is small(2.5%). For academic performances, the average first and second semester grades are 10.64 and 10.23, respectively. And the average approval rates are is 75.06% in the first semester and 71.18% in the second. Socioeconomic factors include scholarships and student debts. EDA results show that 24.8% of students receive scholarships and 11.4% of students report having debts.

We also find that there is a strong correlation between first-semester performance and second-semester outcomes. The prior qualification grades also moderately correlate with academic success. The economic factors, on the other hand, do not exhibit strong correlations with performances and may warrant deeper investigations.

B. Preliminary Modeling Results

Figure 3 demonstrates a preliminary result for the three models trained on full feature sets. Among the three models, we can see the classification metrics are quite similar, with TABNET yielding the best accuracy score of 0.765. These scores serve as a benchmark for later comparisons with different variants of the baseline models.

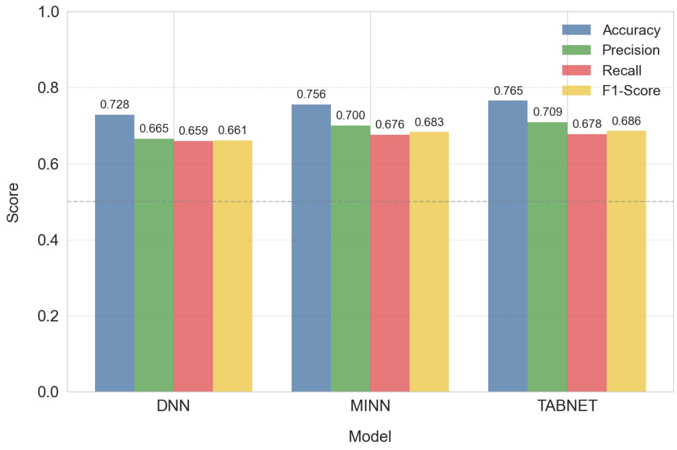

FIG. 3. Performance Metrics for Baseline Models (Full Features)

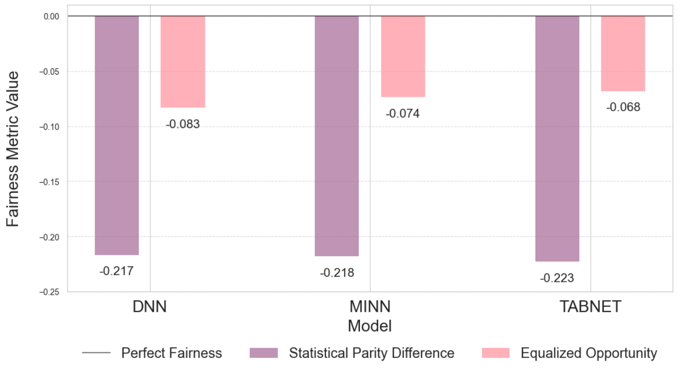

FIG. 4. Fairness Metrics for Baseline Models (Full Features)

Figure 4 illustrates the fairness metrics for the baseline models. All three baseline models showed consistent bias towards one target. The Statistical Parity Difference is around -0.2 and the Equalized Opportunity is around -0.07 to -0.08. These fairness scores call for further analysis of the models.

C. Feature Selection Results

The Random Forest model achieved a classification accuracy of 0.7684 on the test set, demonstrating effective discrimination between student outcome categories. Table I demonstrates combined ranking of top 20 features in the dataset.

TABLE I. Top 20 Features by Random Forest Importance Ranking

Rank	Feature	RF Imp.	Perm. Imp.
1	Curr. units 2nd sem (approved)	0.143	0.149
2	Curr. units 1st sem (approved)	0.092	0.036
3	Curr. units 2nd sem (grade)	0.109	0.029
4	Tuition fees up to date	0.039	0.035
5	Curr. units 2nd sem (evaluations)	0.038	0.009
6	Curr. units 1st sem (evaluations)	0.037	0.004
7	Age at enrollment	0.040	0.003
8	Curr. units 2nd sem (enrolled)	0.021	0.004
9	Mother's occupation	0.026	0.003
10	Admission grade	0.044	0.001
11	Father's occupation	0.029	0.002
12	Course	0.034	0.001
13	Application mode	0.021	0.002
14	Debtor	0.012	0.003
15	Curr. units 1st sem (credited)	0.007	0.004
16	Scholarship holder	0.017	0.002
17	Previous qualification (grade)	0.037	-0.000
18	Curr. units 1st sem (grade)	0.060	-0.002
19	Mother's qualification	0.021	0.001
20	Gender	0.011	0.002

Feature importance analysis shows that semester performance metrics emerged as the strongest predictors of academic outcomes. Specifically, the number of approved curricular units and grade averages from both first and second semesters ranked consistently high across both importance metrics. Socioeconomic indicators such as GDP and unemployment rate also appeared among the top 20 features, suggesting environmental factors play a significant role in student academic success. We choose Age at enrollment, Mother's occupation, Father's occupation, Mother's qualification, and Gender as protected features, leaving a reduced set of 15 features chosen as selected feature to train on three deep learning models.

D. Final Modeling Results

Table II demonstrates the classification performances across three model types (DNN, MINN, TabNet), two feature sets (full, selected), and three training strategies (baseline, SMOTE, adversarial). The metrics include accuracy, precision, recall, and F1-scores.

Among all variant configurations, the best performance one was achieved by the **MINN** model using the **selected feature set** under both the **baseline** and **adversarial** settings, with an accuracy of **0.771**. In general, models using selected features demonstrate similar performances compared to models with full features, and the dimensionality reduction here is effective.

Table III summarizes fairness using two metrics: Statistical Parity Difference (SPD) and Equalized Opportu-

TABLE II. Model Performance Metrics by Configuration

Model	Feat.	Variant	Acc.	Prec.	Rec.	F1
DNN	Full	Adv.	0.720	0.660	0.655	0.657
		Baseline	0.728	0.665	0.659	0.661
		SMOTE	0.734	0.676	0.669	0.672
	Sel.	Adv.	0.733	0.668	0.660	0.662
		Baseline	0.741	0.678	0.670	0.673
		SMOTE	0.721	0.662	0.662	0.661
MINN	Full	Adv.	0.752	0.694	0.676	0.681
		Baseline	0.756	0.700	0.676	0.683
		SMOTE	0.748	0.690	0.674	0.679
	Sel.	Adv.	**0.771**	0.715	0.684	0.693
		Baseline	**0.771**	0.715	0.682	0.691
		SMOTE	0.755	0.700	0.695	0.697
TabNet	Full	Adv.	0.764	0.705	0.685	0.691
		Baseline	0.765	0.709	0.678	0.686
		SMOTE	0.730	0.689	0.695	0.687
	Sel.	Adv.	0.760	0.700	0.677	0.684
		Baseline	0.748	0.687	0.665	0.670
		SMOTE	0.704	0.683	0.689	0.672

nity (EO). From all the variants, it is clear that the **SPD values are negative**. This means the female students are less likely to be predicted to be dropouts, indicating a consistent bias. **MINN with full features and adversarial training** shows an SPD of **-0.243**, which is the biggest one among all variants. The **Equalized Opportunity (EO)** scores were also generally negative. This suggests that models generally do a better job at identifying male dropouts. **TabNet model using selected features with SMOTE** has the least biased score with an EO of **-0.041**.

TABLE III. Fairness Metrics (SPD and EO) by Model and Configuration

Model	Feature Set	Variant	SPD	EO
DNN	Full	Adversarial	-0.219	-0.062
		Baseline	-0.217	-0.083
		SMOTE	-0.210	-0.066
	Selected	Adversarial	-0.189	-0.058
		Baseline	-0.187	-0.043
		SMOTE	-0.188	-0.061
MINN	Full	Adversarial	**-0.243**	-0.094
		Baseline	-0.218	-0.074
		SMOTE	-0.222	-0.081
	Selected	Adversarial	-0.191	-0.061
		Baseline	-0.195	-0.046
		SMOTE	-0.192	-0.056
TabNet	Full	Adversarial	-0.211	-0.066
		Baseline	-0.223	-0.068
		SMOTE	-0.220	-0.082
	Selected	Adversarial	-0.204	-0.055
		Baseline	-0.203	-0.044
		SMOTE	-0.192	**-0.041**

From the different model subsets, we find that adversarial training and SMOTE do help reduce bias. Comparing to the vanilla baseline model results in Figure 4, models with Selected dataset greatly improves the fairness metrics. SMOTE and adversarial training also tend to result in lower absolute SPD and EO with some variations. For example, DNN with the full features variant showed improved fairness with SMOTE, where SPD improves from -0.217 to -0.210. However, TabNet showed minimal improvement from adversarial training. Likewise, feature selection does not demonstrate significant positive effect for all these models.

Overall, MINN has the best accuracy scores, while TabNet with selected features under SMOTE training demonstrated the best fairness trade-off. These results highlights the importance of achieving a balance between predictive performance and fairness in the education field.

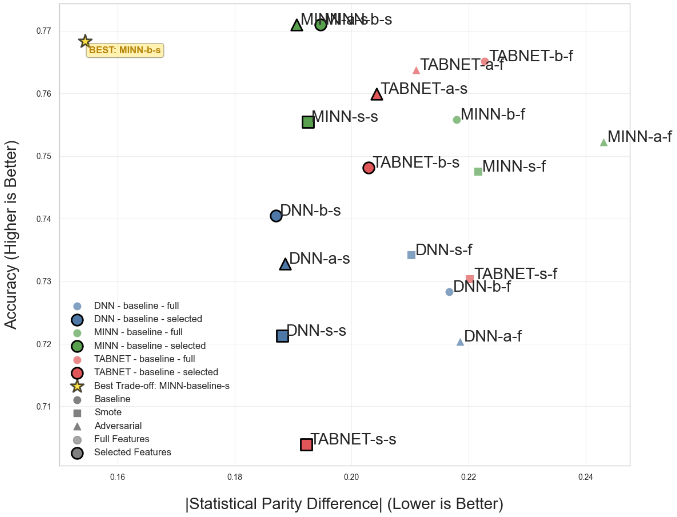

FIG. 5. Accuracy vs. Statistical Parity Difference Trade-off

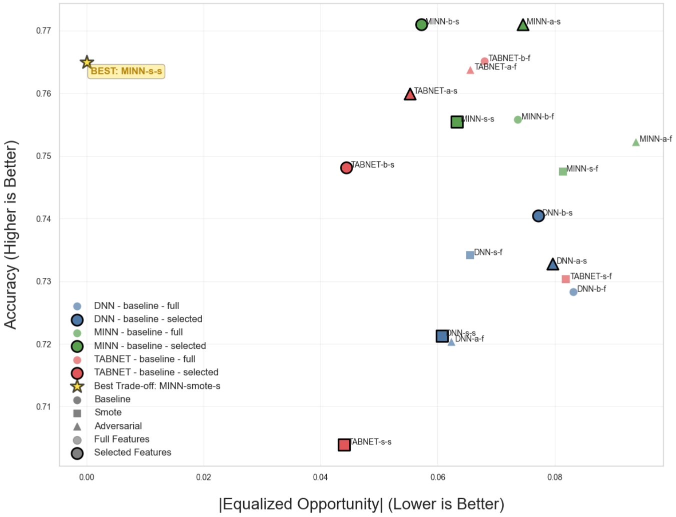

FIG. 6. Accuracy vs. Equalized Opportunity Trade-off

Figure 5 and Figure 6 show the trade-offs between accuracy and fairness metrics. We can see that many high-accuracy models tend to have worse fairness scores. The best model (highlighted by a star in the figure) in such scenario is the MINN model with SMOTE on selected data set. This model achieves the best fairness metric scores while maintaining decent accuracy.

V. CONCLUSION

From the result section, we can clearly see that SMOTE greatly helps the model's performances while sacrificing much fairness. Models show clear improvements in accuracy and recall, indicating solving class imbalance problems can improve the performances. However, the Statistical Parity Difference and Equalized Opportunity all witness noticeable drop, signaling reduced fairness.

After testing 18 different configurations of models, we find that MINN model with selected features and baseline or adversarial training shows the best performance in terms of classification accuracy. The fact that we have better performance means that dimensional reduction through feature selection is a viable way and did not hurt the overall performance.

However, we witness persistent biases for all model variants. All Statistical Parity Difference and Equalized Opportunity values are negative, indicating an underprediction of dropout risks for female students. Such systematic bias could be further analyzed on the sociol-economic reasons. One possible reason could be the fact that there are significant barriers for females in the job markets, and they tend to have stronger need for higher degrees so that they can move upward. Also, female students may be more easily to be socially integrated into campus life and form their own social networks.

In this project, we used Generative AI such as ChatGPT and Copilot to help write and organize the code. The project overall emphasized on the importance of balancing between the performance of models and their fairness. Such trade-offs should be carefully evaluated and people should ill afford to go blindly with the best performing models without any evaluation on fairness metrics.

DATA AVAILABILITY

Data is available at `https://github.com/yourlogin/projectname`.

CODE AVAILABILITY

Code is available at `https://github.com/Haoze-Z/DSCI531_Group_Project_Student_Dropouts`.

[1] L. Paura and I. Arhipova, Cause analysis of students' dropout rate in higher education study program, Procedia - Social and Behavioral Sciences **109**, 1282 (2014), 2nd World Conference on Business, Economics and Management.

[2] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, Predicting student dropout in higher education, in *ICML Workshop on #Data4Good: Machine Learning in Social Good Applications* (2016).

[3] M. V. Martins, D. Tolledo, J. Machado, L. M. Baptista, and V. Realinho, Early prediction of student's performance in higher education: a case study, in *Trends and Applications in Information Systems and Technologies: Volume 1 9* (Springer, 2021) pp. 166–175.

[4] E. Drousiotis, P. Pentaliotis, L. Shi, and A. I. Cristea, Capturing fairness and uncertainty in student dropout prediction – a comparison study, in *Artificial Intelligence in Education*, edited by I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova (Springer International Publishing, Cham, 2021) pp. 139–144.

[5] V. M. M. M. J. Realinho, Valentim and L. Baptista, Predict Students' Dropout and Academic Success, UCI Machine Learning Repository (2021), DOI: https://doi.org/10.24432/C5MC89.