Examining the Algorithmic Fairness in Predicting High School Dropouts

Chenguang Pan
cp3280@columbia.edu

Zhou Zhang
zz2863@columbia.edu

Teachers College, Columbia University

Overview

1. Introduction
2. Data and variables
3. Evaluate the predictive performance
4. Examine the algorithmic fairness on selected models
5. Conclusions

1. Introduction

A 2015 report from *U.S. Department of Education* indicates that over 1.2 million students drop out of high school in the United States each year.

That's a dropout every 26 seconds – or 3300 a day.

To detect the students at risk in advance is critical for preventing the potential dropouts. Various studies explored this topic. However, less attention has been paid on **the fairness of the predicting models**.

1. Introduction

A 2015 report from *U.S. Department of Education* indicates that over 1.2 million students drop out of high school in the United States each year.

That’s a dropout every 26 seconds – or 3300 a day.

To detect the students at risk in advance is critical for preventing the potential dropouts. Various studies explored this topic. However, less attention has been paid on **the fairness of the predicting models**.

1. Introduction

A 2015 report from *U.S. Department of Education* indicates that over 1.2 million students drop out of high school in the United States each year.

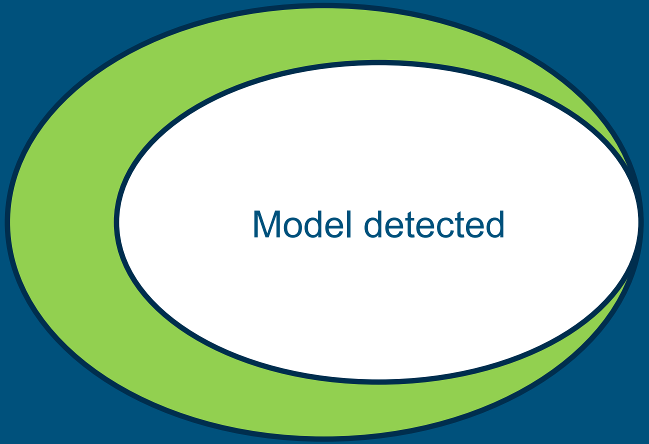
That's a dropout every 26 seconds – or 3300 a day.

To detect the students at risk in advance is critical for preventing the potential dropouts. Various studies explored this topic. However, less attention has been paid on **the fairness of the predicting models**.

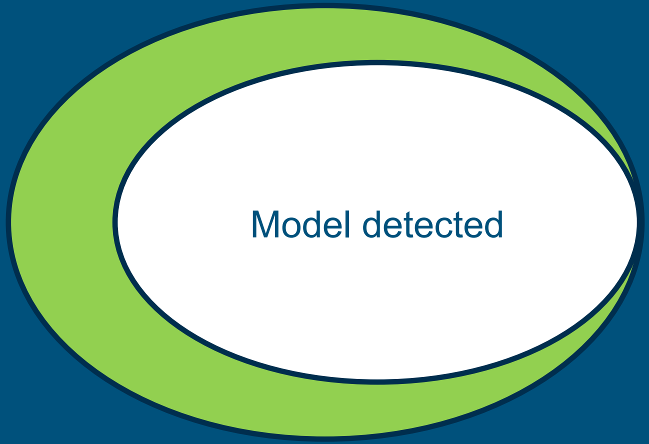
1. Introduction

Algorithmic fairness:

We should expect the model perform equally on different groups after controlling for some factors.



Male students who actually dropped

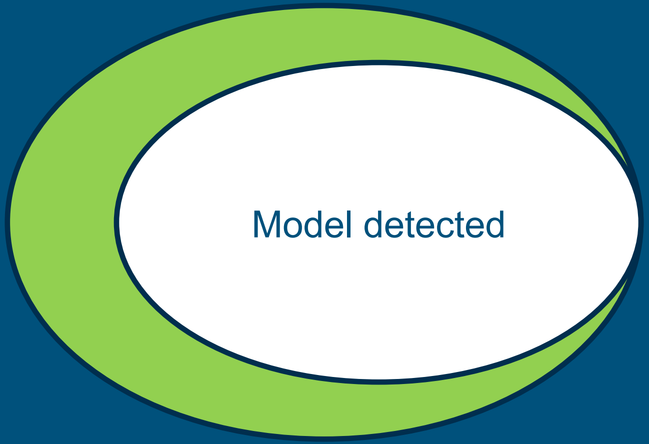


Female students who actually dropped

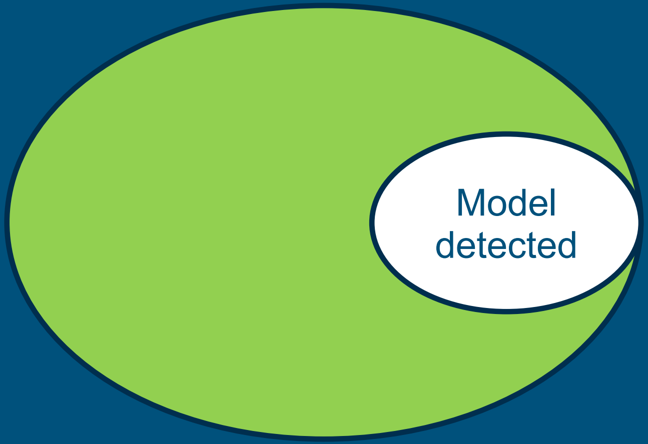
1. Introduction

Algorithmic fairness:

However, some models may perform less well on specific group.



Male students who actually dropped



Female students who actually dropped

1. Introduction

Research Questions:

1. Does a machine learning model inevitably introduce bias in predicting high school dropouts?

2. In what ways does the inclusion or exclusion of protected attributes affect the predictive performance of the model?

3. How does the inclusion or exclusion of protected attributes impact the algorithmic fairness of the model?

1. Introduction

Research Questions:

1. Does a machine learning model inevitably introduce bias in predicting high school dropouts?

2. In what ways does the inclusion or exclusion of protected attributes affect the predictive performance of the model?

3. How does the inclusion or exclusion of protected attributes impact the algorithmic fairness of the model?

1. Introduction

Research Questions:

1. Does a machine learning model inevitably introduce bias in predicting high school dropouts?

2. In what ways does the inclusion or exclusion of protected attributes affect the predictive performance of the model?

3. How does the inclusion or exclusion of protected attributes impact the algorithmic fairness of the model?

2. Data*, variables, and models: dataset

High School Longitudinal Study of 2009 (HSLS:09) is a nationally representative, longitudinal study of more than 21,000 ninth graders in 944 schools who were followed through their secondary and postsecondary years.

*** We have open-sourced the data and code for this project, available at:**

https://github.com/cgpan/HSLSdropout.

2. Data, variables, and models: predictors

Student-level:

- Gender, Race/ethnicities, Age
- Prior academic performance
- Self-expectation
- School engagement and School belonging
- Attitude towards Math and Sci. (ID, Efficacy, Utility, INTERESTS)

Family-level:

- SES (income, parent’s occupation & education, location)
- Parent’s expectation

School-level:

- School type (Public/ Private) and region
- School climate and problem

2. Data, variables, and models: outcome

X4EVERDROP: Ever dropped out of high school, binary coding with 1=Yes.

The processed dataset is at 17332×30 size.

Predictive Mean Matching (PMM) is used to impute the missing data.

However, this is an imbalanced dataset.

2. Data, variables, and models: oversampling

The distribution of the outcome variable:

14618/17332

negative cases.
85% never dropped.

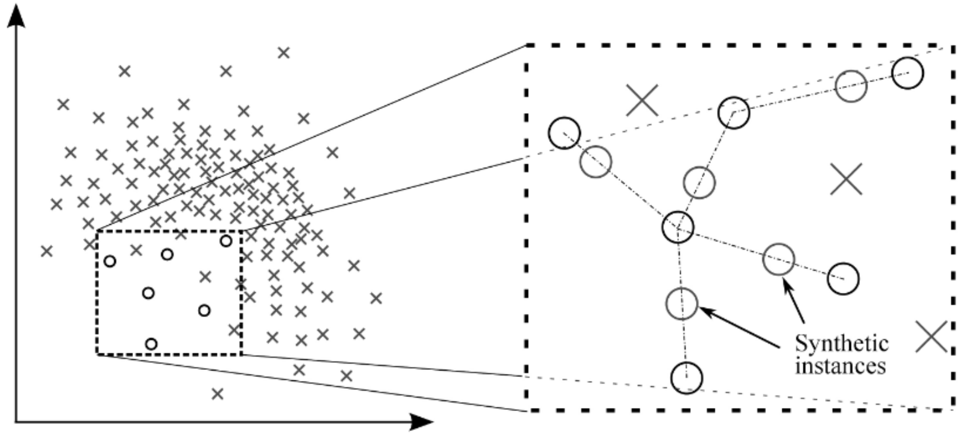
2714/17332

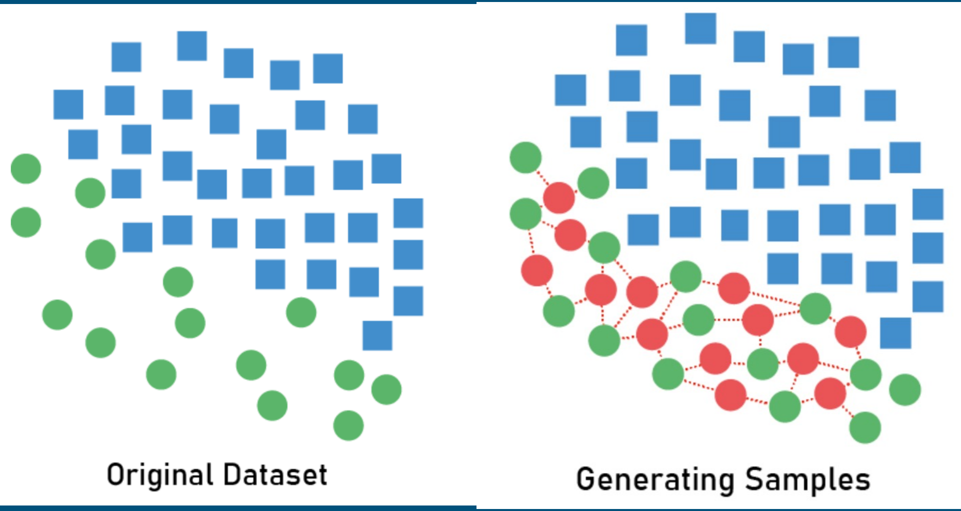
positive cases.
15% dropped.

The primary purpose of the model is to detect the dropouts, to predict the true positive as many as possible.

2. Data, variables, and models: oversampling

The Synthetic Minority Oversampling Technique (SMOTE)
Generate new observations from the minority class by randomly picking a point from the minority and computing K-nearest neighbors.





https://www.datasciencecentral.com/handling-imbalanced-data-sets-in-supervised-learning-using-family/

https://bestnavosk.live/product_details/89620871.html

3. Evaluate the predictive performance

$$Accuracy = \frac{Correctly\ Precited\ Observations}{Toatal\ Observations}$$

$$Recall(Sensitivity) = \frac{True\ Positive}{False\ Positive + True\ Positive}$$

$$Specificity = \frac{True\ Negative}{False\ Negative + True\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$F_\beta\ score = (1 + \beta^2)\frac{Precision \times Recall}{\beta^2 \times Precision + recall}$$

AUC (Area under the ROC Curve)

3. Evaluate the predictive performance

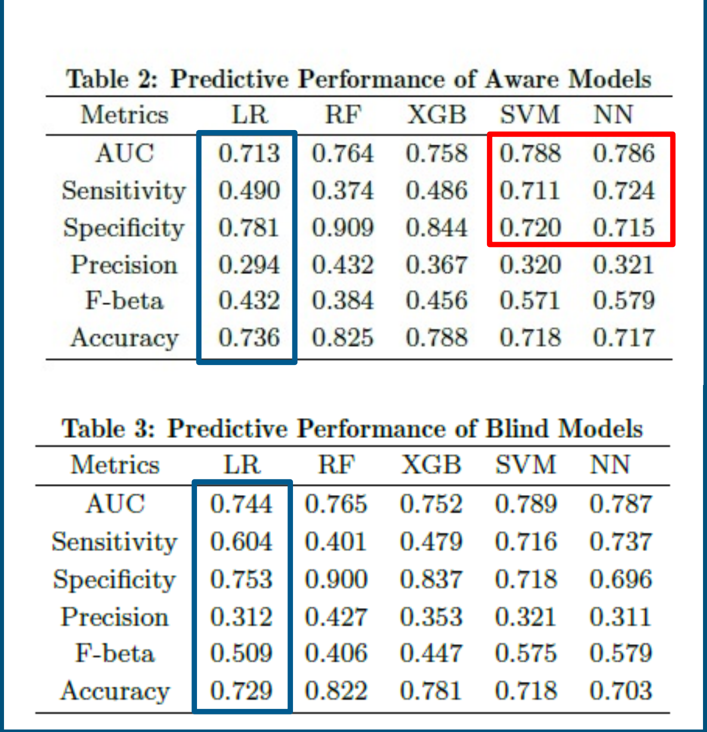


Table 2: Predictive Performance of Aware Models

Metrics	LR	RF	XGB	SVM	NN
AUC	0.713	0.764	0.758	0.788	0.786
Sensitivity	0.490	0.374	0.486	0.711	0.724
Specificity	0.781	0.909	0.844	0.720	0.715
Precision	0.294	0.432	0.367	0.320	0.321
F-beta	0.432	0.384	0.456	0.571	0.579
Accuracy	0.736	0.825	0.788	0.718	0.717

Table 3: Predictive Performance of Blind Models

Metrics	LR	RF	XGB	SVM	NN
AUC	0.744	0.765	0.752	0.789	0.787
Sensitivity	0.604	0.401	0.479	0.716	0.737
Specificity	0.753	0.900	0.837	0.718	0.696
Precision	0.312	0.427	0.353	0.321	0.311
F-beta	0.509	0.406	0.447	0.575	0.579
Accuracy	0.729	0.822	0.781	0.718	0.703

*The threshold is set to be .50 for interpretability concern.

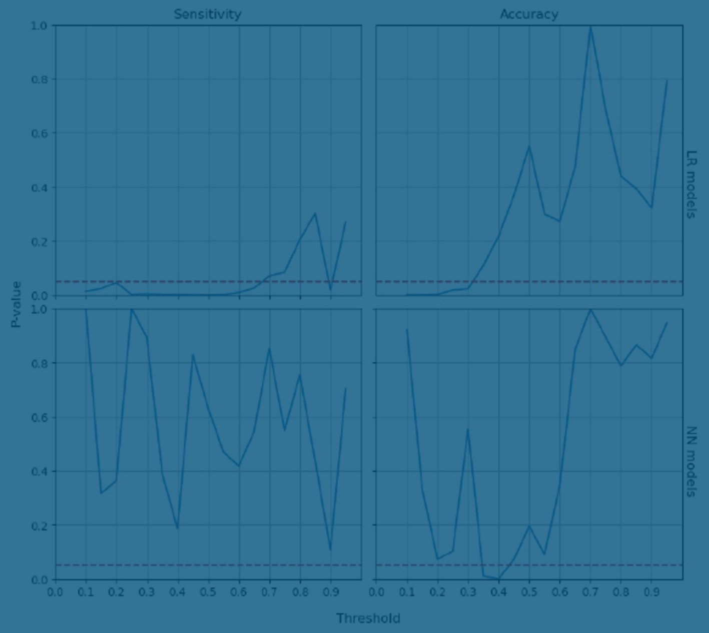


Figure 2: P-values from proportion Z-test on changes in recall and accuracy for LR and NN models. Red dashed line represents the significance level of 0.05. P-values under this line indicate significant difference in corresponding metric between the aware and blind models.

3. Evaluate the predictive performance

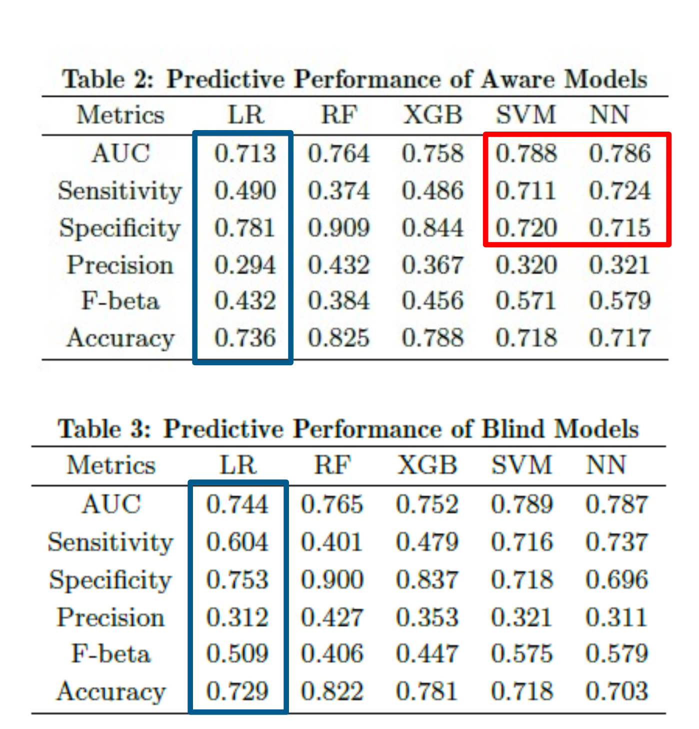


Table 2: Predictive Performance of Aware Models

Metrics	LR	RF	XGB	SVM	NN
AUC	0.713	0.764	0.758	0.788	0.786
Sensitivity	0.490	0.374	0.486	0.711	0.724
Specificity	0.781	0.909	0.844	0.720	0.715
Precision	0.294	0.432	0.367	0.320	0.321
F-beta	0.432	0.384	0.456	0.571	0.579
Accuracy	0.736	0.825	0.788	0.718	0.717

Table 3: Predictive Performance of Blind Models

Metrics	LR	RF	XGB	SVM	NN
AUC	0.744	0.765	0.752	0.789	0.787
Sensitivity	0.604	0.401	0.479	0.716	0.737
Specificity	0.753	0.900	0.837	0.718	0.696
Precision	0.312	0.427	0.353	0.321	0.311
F-beta	0.509	0.406	0.447	0.575	0.579
Accuracy	0.729	0.822	0.781	0.718	0.703

*The threshold is set to be .50 for interpretability concern.

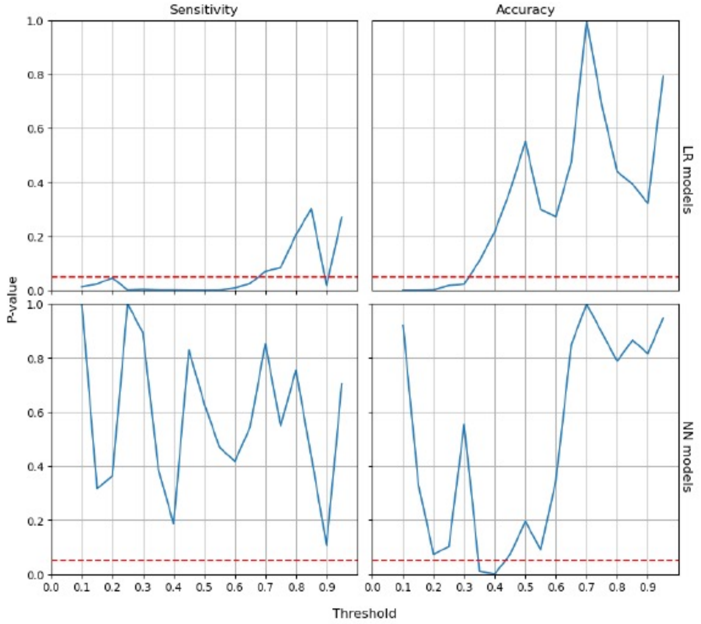


Figure 2: P-values from proportion Z-test on changes in recall and accuracy for LR and NN models. Red dashed line represents the significance level of 0.05. P-values under this line indicate significant difference in corresponding metric between the aware and blind models.

4. Examine the algorithmic fairness on selected models

1. Statistical Disparity:

$$\Pr(D = 1|G = g) = \Pr(D = 1|\, G = g')$$

2. Conditional Statistical Disparity:

$$\Pr(D = 1|L, G = g) = \Pr(D = 1|L,\, G = g')$$

3. Separation/ Error Rate Balance/ Equalized odds:

$$\Pr(D = 1|Y = y, G = g) = \Pr(D = 1|Y = y,\, G = g')$$

4. Sufficiency:

$$\Pr(Y = 1|D = d, G = g) = \Pr(Y = 1|D = d,\, G = g'), \qquad d \in \{0,1\}$$

(Suk & Han, 2023)

4. Examine the algorithmic fairness on selected models

Differential Algorithmic Functioning* (DAF) is motivated by DIF in psychometrics.
A fair algorithm should not make discriminatory decisions based on protected variables after controlling for fair attribute

$$DAF\colon \Pr(D = 1|W, G = g) \neq \Pr(D = 1|W, G = g')$$

$$NonDAF\colon \Pr(D = 1|W, G = g) = \Pr(D = 1|W, G = g')\,, \qquad where\ W\ is\ the\ fair\ attribute.$$

Three methods to detect DAF: Mantel-Haenszel test, Logistic regression, and residual-based

W: SES, Academic performance, School Belonging…

*Suk, Y., & Han, K. T. (2024). A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning.

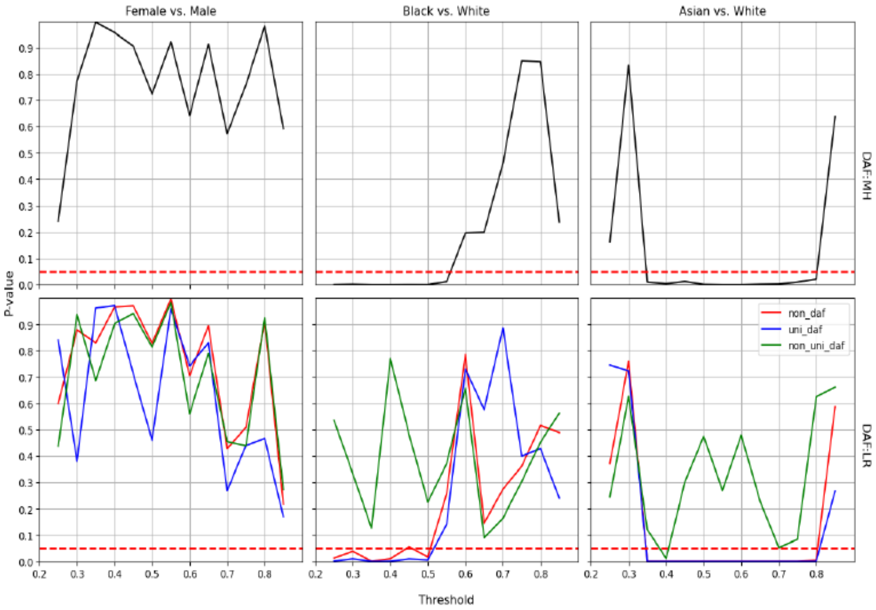


Figure 3: P-values for algorithmic fairness of aware NN model using DAF detection methods. Red dashed line represents the significance level of 0.05. MH method (top three) detects the DAF only, whereas LR (bottom three) method can additionally identify both uniform and non-uniform DAF. For the LR method, initially examine the red line to determine if any part of it falls below the red dashed line. If it does not, there is no need to further check for uniform or non-uniform DAF, as this indicates the absence of DAF. Figure 4, 5, and 6 follow the same logic.

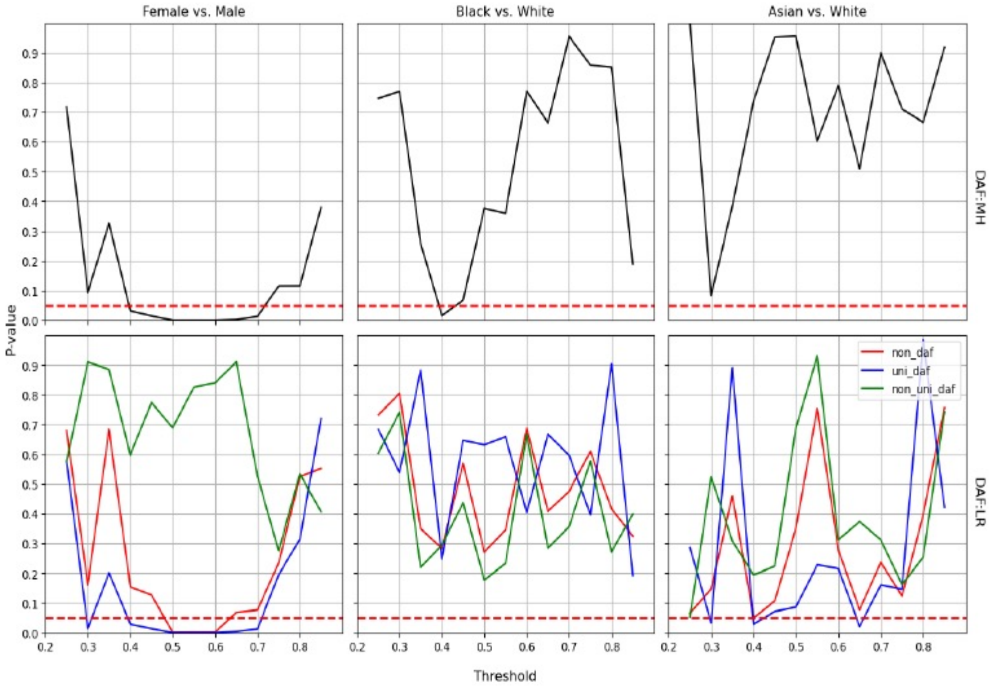


Figure 4: P-values for algorithmic fairness of blind NN model using DAF detection methods.

Algorithmic Fairness for aware and blind NN models.

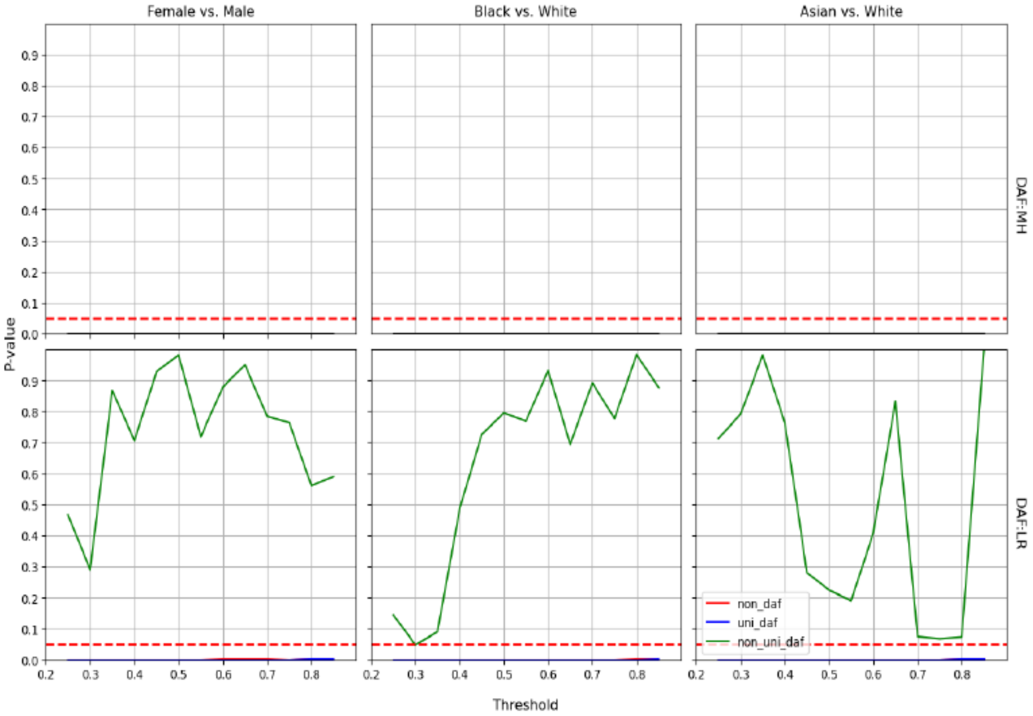


Figure 5: P-values for algorithmic fairness of aware LR model using DAF detection methods. It should be noted that the the black lines in the top three plots and the red and blue lines in the bottom three plots are all below the red dashed line (i.e., p-values are very close to 0), which demonstrates (uniform) DAF across all thresholds (0.25˜0.90).

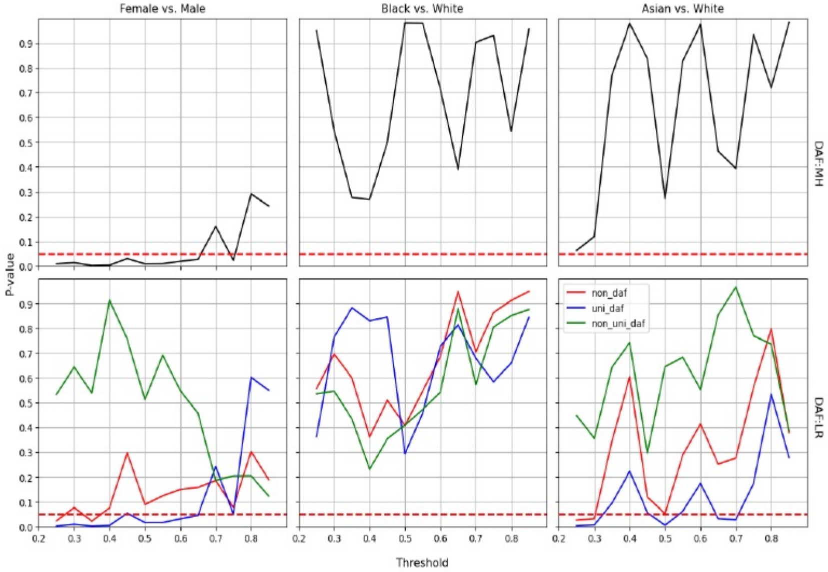


Figure 6: P-values for algorithmic fairness of blind LR model using DAF detection methods.

Algorithmic Fairness for aware and blind LR models.

5. Conclusion

1. Does a machine learning model inevitably introduce bias in predicting high school dropouts?

Conclusion:
The presence of predictive bias is contingent upon the

a) model used,
b) the protected groups involved, and
c) the specific threshold range.

5. Conclusion

2. In what ways does the inclusion or exclusion of protected attributes affect the predictive performance of the model?

3. How does the inclusion or exclusion of protected attributes impact the algorithmic fairness of the model?

5. Conclusion

Conclusion:
The impact of protected variables on predictive performance varies depending on the model. These attributes may have a minimal contribution to the detection of the target, especially in terms of specific metrics like sensitivity or accuracy.

The decision to include or exclude these variables should also take into account their effect on algorithmic fairness. Omitting certain group memberships might reduce predictive bias (as seen in the Black vs. White comparison in the blind NN model), but it could simultaneously exacerbate outcome disparities in other groups (such as the female vs. male comparison in the blind NN model).

5. Conclusion

Conclusion:
The impact of protected variables on predictive performance varies ***depending on the model***. These attributes may have a minimal contribution to the detection of the target, especially in terms of specific metrics like sensitivity or accuracy.

The decision to include or exclude these variables ***should also take into account their effect on algorithmic fairness***. Omitting certain group memberships might reduce predictive bias (as seen in the Black vs. White comparison in the blind NN model), but it could simultaneously exacerbate outcome disparities in other groups (such as the female vs. male comparison in the blind NN model).

5. Conclusion

We recommend that in deciding the retention of a protected attribute, researchers should consider its impact on three key aspects across the range of thresholds they are interested in:

a) predictive performance according to certain metrics,
b) changes in algorithmic fairness, and
c) the practical implications of deploying the models.

6. Future plan

We are currently extending this study to explore potential improvements in predictive performance, conduct a more comprehensive evaluation of algorithmic fairness, develop strategies to reduce bias, and analyze the trade-offs between predictive performance and fairness.

Thank You!

Contact Info:	cp3280@columbia.edu
Data and code:	https://github.com/cgpan/HSLSdropout.