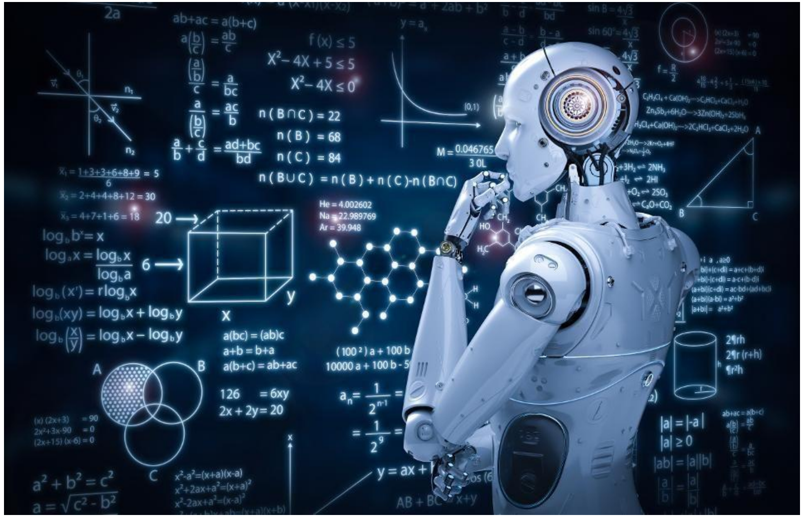
Evaluating fairness constraints to enforce a fair classifier to remove bias in COMPAS

Shiva Omrani, Abia Khan

Introduction

- Algorithmic decision making systems is part of our everyday lives, whether its choosing a successful job applicant or deciding which restaurant to eat.
  - However, can these automated decisions lead to a unintentional lack of fairness?



COMPAS


JAMES RIVELLI
LOW RISK 3
ROBERT CANNON
MEDIUM RISK 6


JAMES RIVELLI
Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking
Subsequent Offenses
1 grand theft
LOW RISK 3
ROBERT CANNON
Prior Offense
1 petty theft
Subsequent Offenses
None
MEDIUM RISK 6

Problem Statement

- Since the COMPAS Algorithm is considered a "black box" and we don't have access to the training data, our technique will evaluate which fairness notion are essential to a fair criminal risk score assessment and reassign the output results based on our post-processing.

Related Work

- ProPublica, Angwin et al. - One of the first studies that discovered COMPAS algorithm was wrong in its predicting, the results were displayed differently for black and white offenders.

- Verma et. Al - A comprehensive review of the most prominent definitions of fairness in the algorithmic classification problem

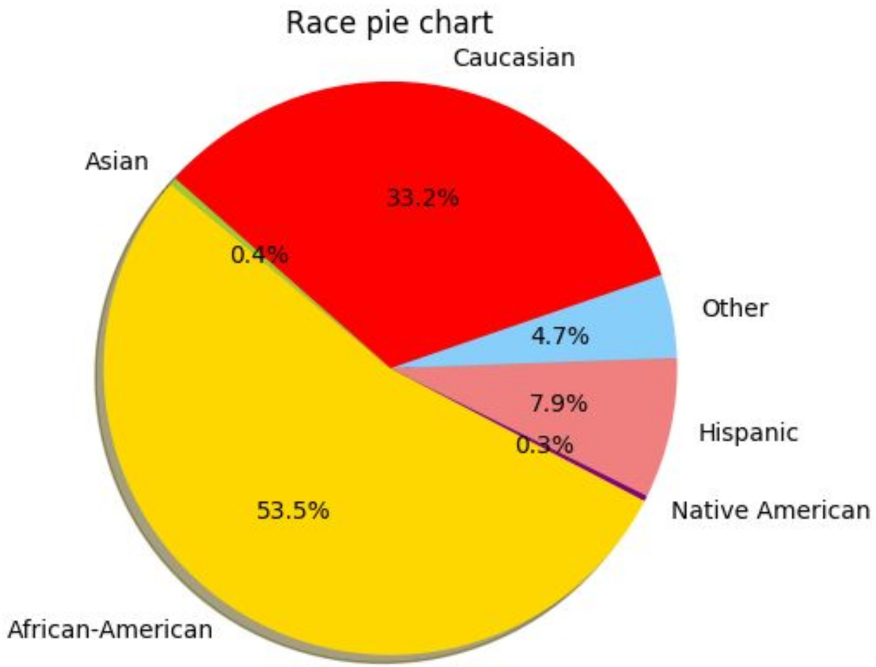
- Zafar et. Al - Proposed a fair classifier formulation to remove disparate mistreatment only on false positive and false negative rates in COMPAS.

Dataset

- Data gathered by ProPublica, contains both COMPAS risk scores and criminal records for each defendant.
- 52 features, 18,310 rows
- Criminal record indicating recidivism after two years from risk assignments acts as the ground truth.
- Risk of violence, risk of recidivism, risk of failure to appear.
- Score scale from 1-10 with 10 being the highest risk.

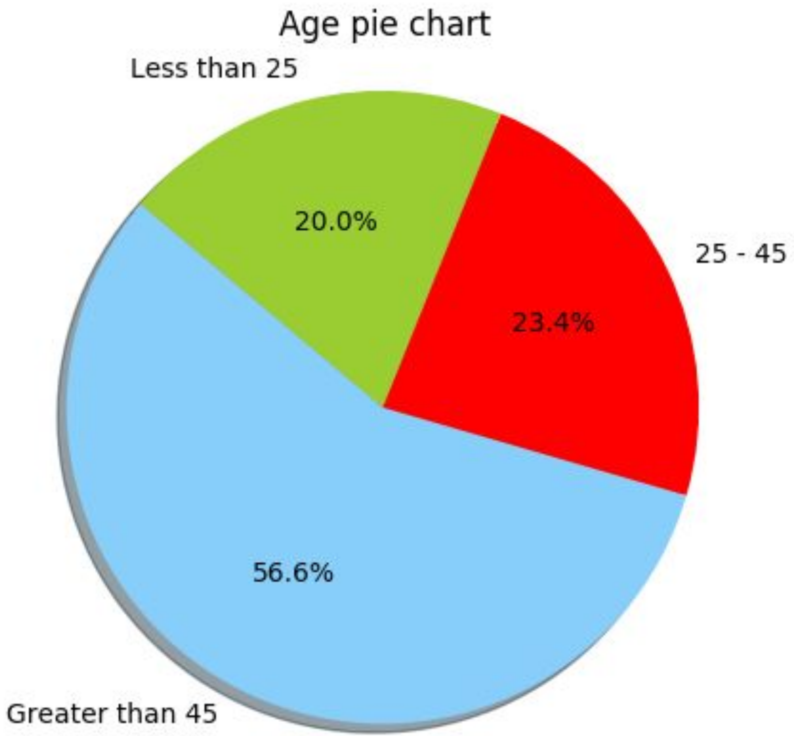
Race Distribution

- African Americans make up the majority.
- Together with caucasian, account for 87%.



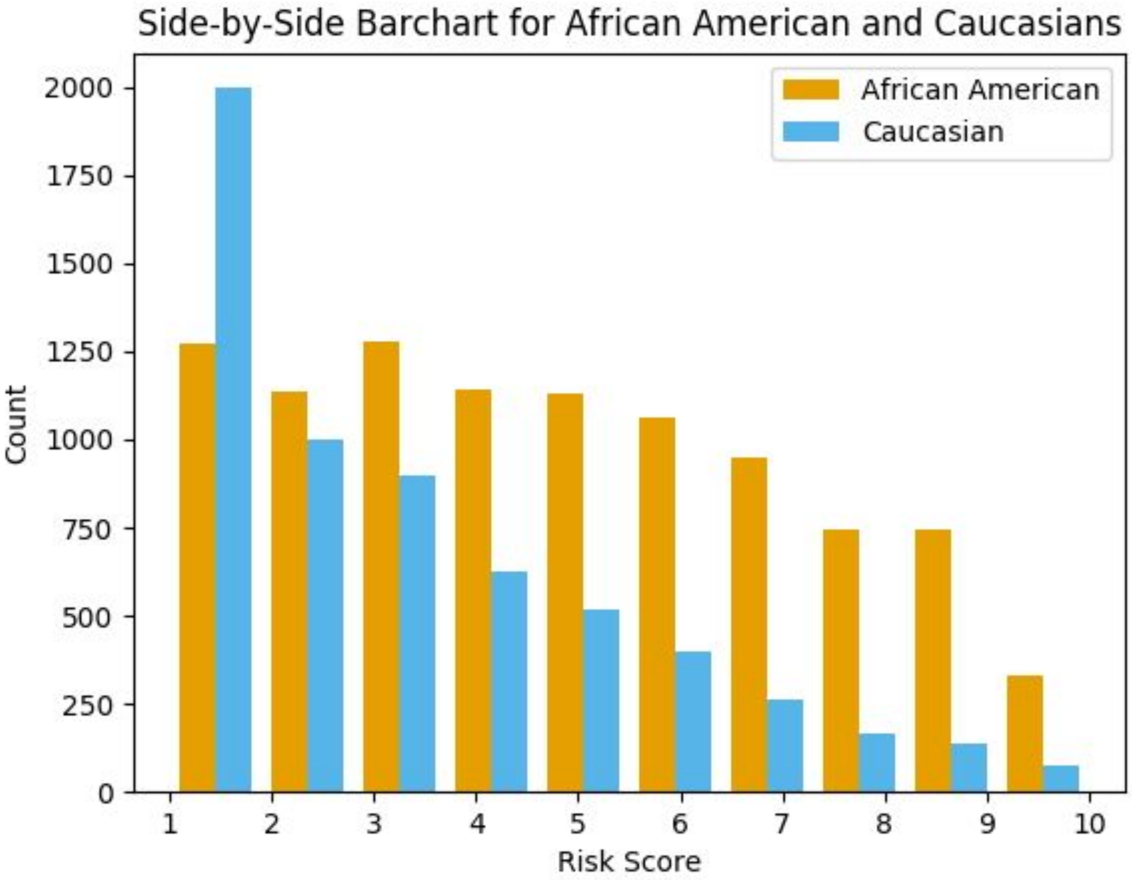
Age Distribution

- Greater than 45 make up the majority.

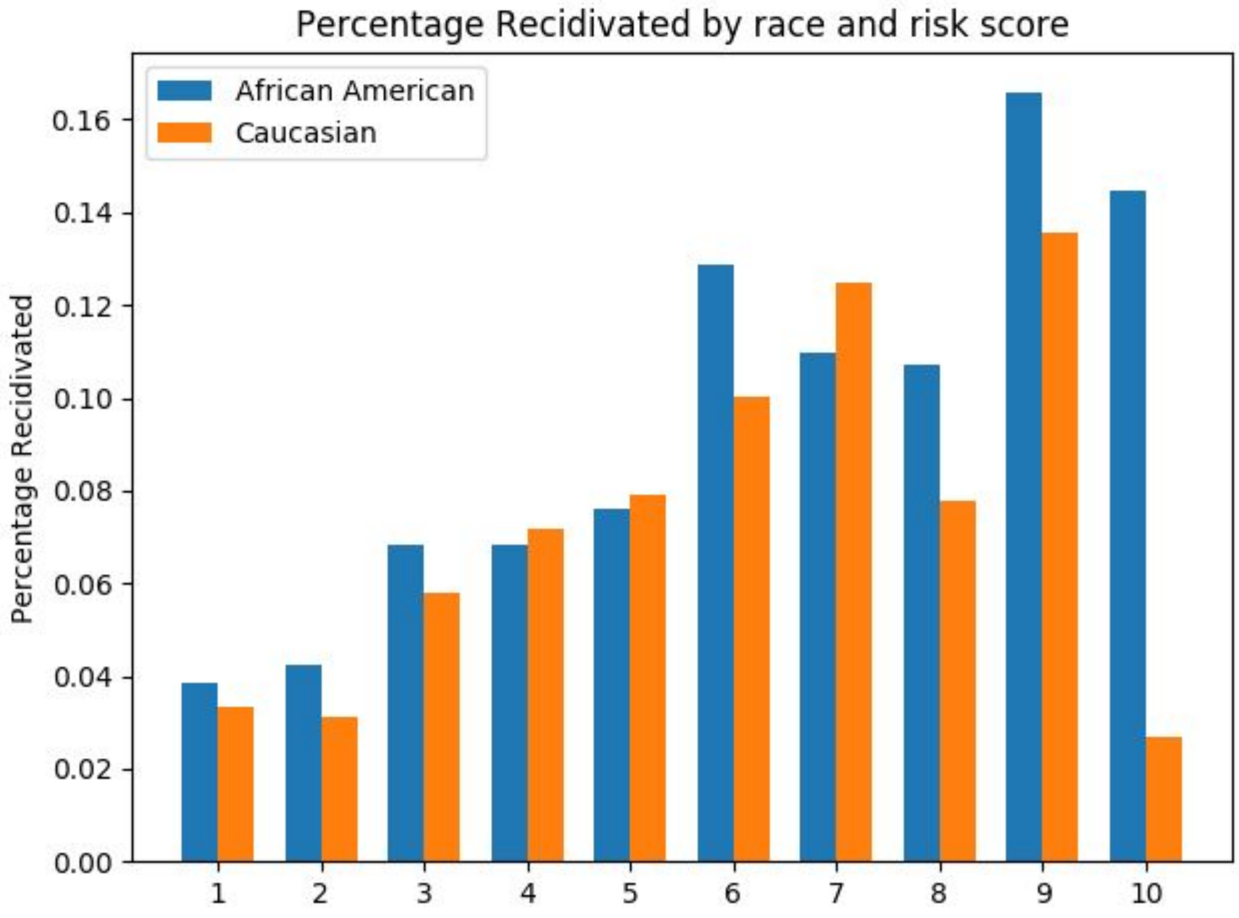


Data Pre-Processing

- Used only age, race, number of prior convictions, charge degree
- Filtered by African American and Caucasian
- Filtered by pre-trial stage
- Analyzed risk of violence and recidivism
- Used both decile score and score category as target variable



Side-by-Side Barchart for African American and Caucasians
African American
Caucasian
Count
Risk Score
0
250
500
750
1000
1250
1500
1750
2000
1
2
3
4
5
6
7
8
9
10


Percentage Recidivated by race and risk score
African American
Caucasian
Percentage Recidivated
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3
4
5
6
7
8
9
10

Methodology

Step 1

Evaluate fairness in COMPAS with respect to different notions of fairness.

Step 2

Evaluate which fairness notion are essential to a fair criminal risk score assessment

Step 3

Enforce the selected fairness notions on a classifier trained on our dataset (conjunction of false positive error rate balance and false negative error rate balance)

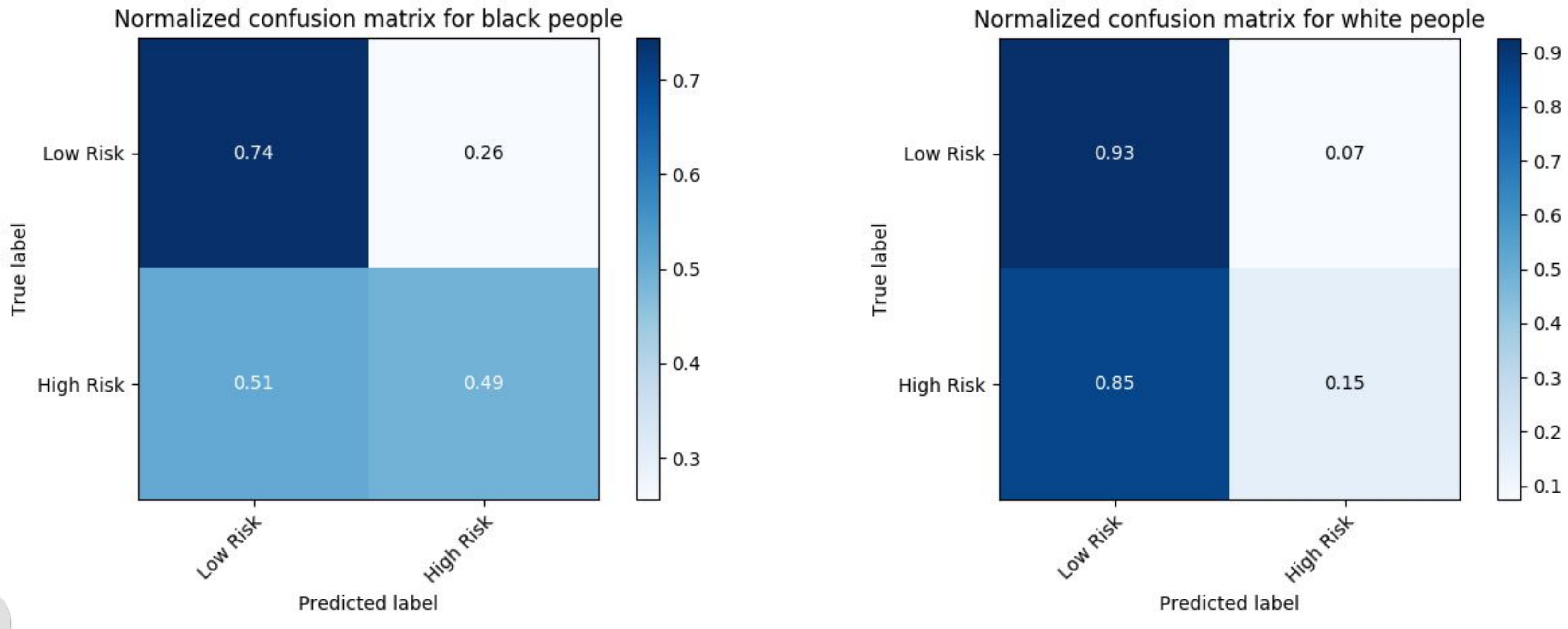
Step 4

Reassign the output results based on the classification results.

Results - Evaluation of Fairness Notions

Fairness Notion	Satisfied or Not
Statistical Parity	No
Predictive Parity	Yes
Predictive Equality (FP Error Rate Balance)	No
Equal Opportunity (FN Error Rate Balance)	No
Conditional Use Accuracy Equality	Yes
Overall Accuracy Equality	No
Treatment Equality	No
Calibration	Yes
Balance for Positive Class	No
Balance for Negative Class	No
Causal Discrimination	Yes

Results - FNR and FPR



Results - Decision Boundary-Based Classifier

	Unconstrained	*Equal Opportunity*	*Predictive equality*	*Equalized odds*
FNR, black	*0.24*	*0.2*	*0.22*	*0.16*
FNR, white	*0.12*	*0.16*	*0.15*	*0.14*
FPR, black	*0.43*	*0.52*	*0.46*	*0.55*
FPR, white	*0.77*	*0.67*	*0.70*	*0.71*
Accuracy	*66%*	*65%*	*65%*	*65%*

Summary of Contributions

- We evaluate fairness in COMPAS with respect to different notions of fairness.

- Next we choose which fairness notion are essential to a fair criminal risk score assessment and enforce the selected fairness notions on a classifier trained on our dataset.

- The results of this study showed that we are able to adjust the decision boundary according to the constraints and can enforce the mentioned fairness notions on our classifier.

Limitations and Future Work

- Lacking access to the black box COMPAS algorithm
- Did not achieve perfect equalized odds due to limited size of dataset.
- Global optimum solution not guaranteed.
- Possibility to enforce other notions of fairness that COMPAS violated.
- Possibility to evaluate COMPAS w.r.t similarity based notions of fairness.

Conclusions

- The results of this study provide insight into valuable techniques that can be used to evaluate and improve fairness in COMPAS.
- This study provides a framework to enforce the mentioned fairness notions on our classifier in improving algorithm fairness in COMPAS.