May 27, 2021

Technical Solutions to Evaluation Fairness in Algorithms

Team 7: Kristy Guo, Barbara Liang, Jingjing Lu, Chiebuka Onwuzurike, Qiqi Tang

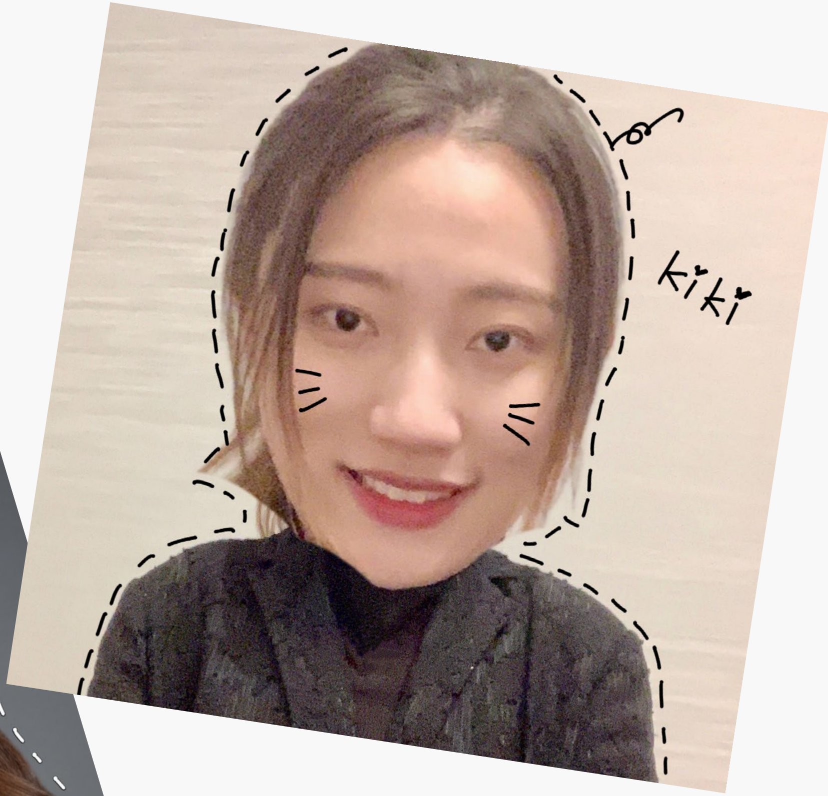
Team

kk

chichi

bb

cc



kiki

Introduction

53%

- “only 53% of organizations have a leader who is responsible for the ethics of AI systems”

52%

- “majority of consumers are more frightened (52%) about the future impact of AI on society than excited”

- Create a safe environment for users
- Increase acceptance

Is it ethical to use algorithms to predict crimes?

Are you ready?

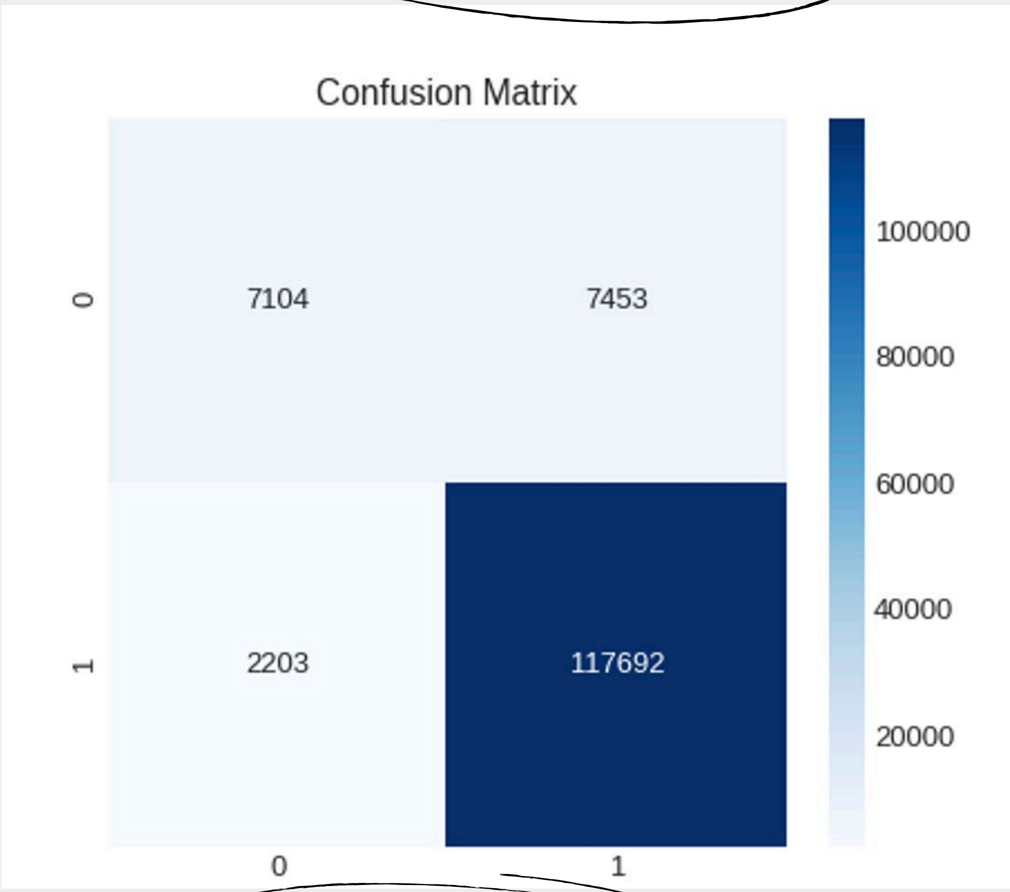
Case

- Between 1980 -2014 there were 190,282 unsolved crimes
- A model that could predict the race and sex of a perpetrator based on the victim description and case facts would be extremely useful
- Given the purpose of the model it is import the model is accurate and unbiased

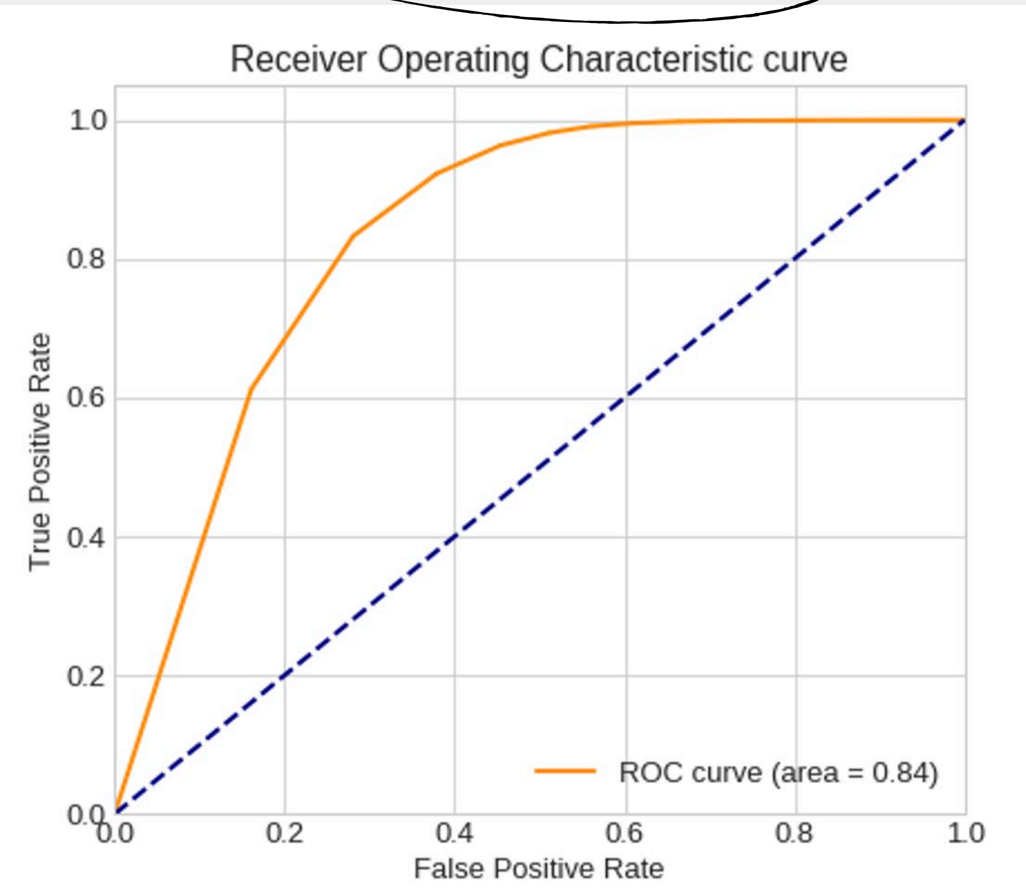
<u>Variables</u>

'Agency Name', 'Agency Type', 'City', 'State', 'Year', 'Month', 'Crime Type', 'Victim Sex', 'Victim Age', 'Victim Race', 'Victim Ethnicity', 'Relationship', 'Weapon', 'Victim Count', 'Perpetrator Count', 'Record Source'

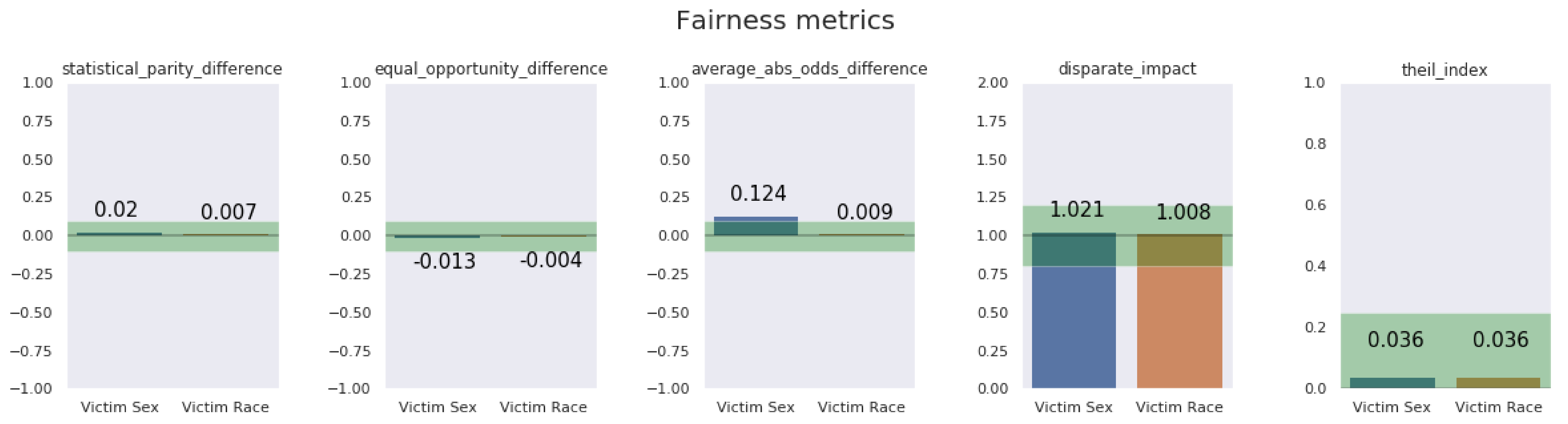
Accuracy: 92.81%



F1 Score: 96.05%



Is the model Unbiased (Fair)?



Confusion Matrix

	Predicted: positive (privileged)	Predicted: negative (unprivileged)
Actual: positive (privileged)	True Positive (TPR privileged)	False Negative (FPR privileged)
Actual: negative (unprivileged)	False Positive (FPR unprivileged)	True Negative (TPR unprivileged)

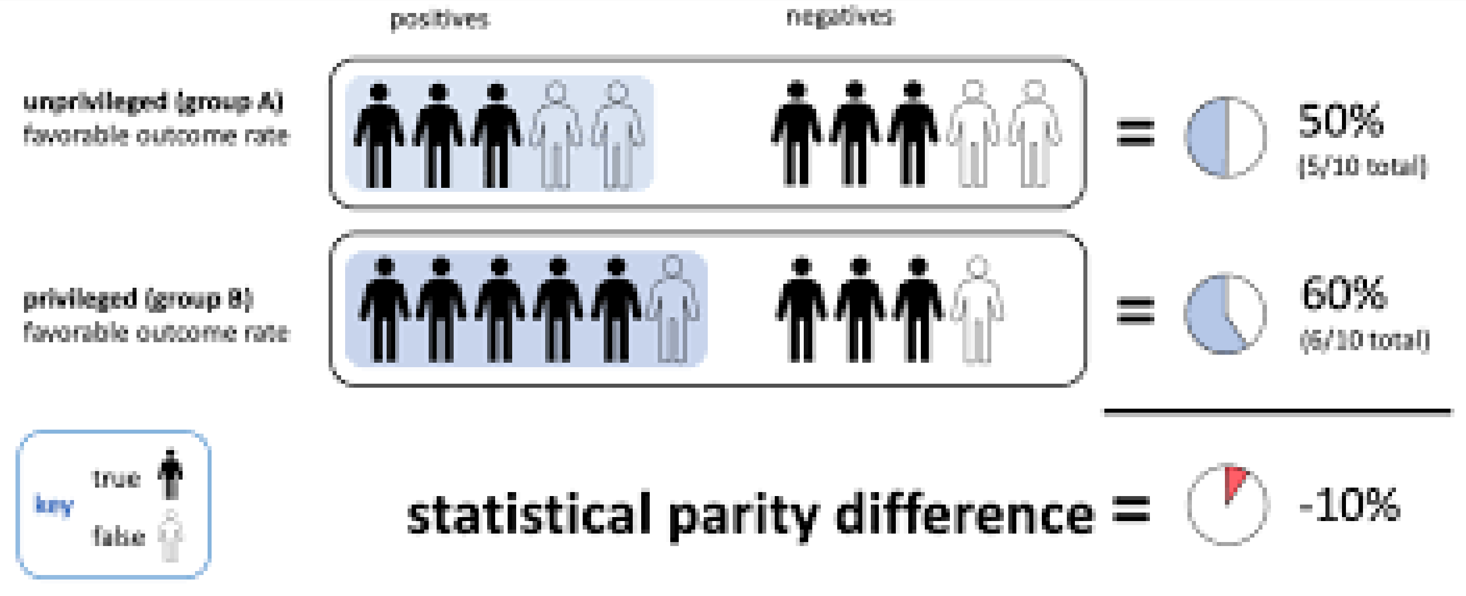
Statistical Parity Difference

Method 1

Statistical Parity Difference suggest a predictor is unbiased (bias(X,S,D)) or fair if the absolute difference between the prediction (Y) of privileged (D) and unprivileged group is lower than a certain threshold (ε)

$$bias_h(X,S,D) = Pr(Y=1 \mid D = unprivileged) - Pr(Y=1 \mid D = privileged)$$

$$|bias_h(X,S,D)| < \varepsilon$$



Is a good metric when statistical power is large and when there aren't that many underlying cofounders

Equal Opportunity Difference

	Predicted: positive (privileged)	Predicted: negative (unprivileged)
Actual: positive (privileged)	True Positive (TPR privileged)	False Negative (FPR privileged)
Actual: negative (unprivileged)	False Positive (FPR unprivileged)	True Negative (TPR unprivileged)

Method 2

$$TPR_{D=unprivileged} - TPR_{D=privileged}$$

Situation when it is a good metric

- Ideal value: 0
- Fairness: between -0.1 and 0.1

Situation when it is a bad metric

- EOD < 0 $\longrightarrow$ privileged
- EOD > 0 $\longrightarrow$ unprivileged

Average Absolute Odds Difference

- Is concerned with the whole confusion matrix.
- Average odds difference = 0
  - No bias
- Advantages : comprehensive
- Disadvantages : cumbersome

Method 3

$$\frac{1}{2}\left[\left|FPR_{D=\text{unprivileged}} - FPR_{D=\text{privileged}}\right| + \left|TPR_{D=\text{unprivileged}} - TPR_{D=\text{privileged}}\right|\right]$$

	Predicted: positive (privileged)	**Predicted: negative (unprivileged)**
Actual: positive (privileged)	True Positive (TPR privileged)	False Negative (FPR privileged)
Actual: negative (unprivileged)	False Positive (FPR unprivileged)	True Negative (TPR unprivileged)

Disparate Impact

Method 4

Formula: $$\frac{Pr(Y = 1|D = unprivileged)}{Pr(Y = 1|D = privileged)}$$

Disparate impact checks discrimination that is unintentional.

Situation when the metric is good:

- Employment: reaction test
- Possible bias against older applicants (the protected class)

Situation when the metric is bad:

- Employment: employers have business reasons to justify reaction test.
- Trade-off between costs and fairness

Theil Index

Method 5

The Theil index measures an entropic "distance" the population is away from the "ideal" egalitarian state of everyone having some defined standard.

- Value of 0 represents perfect equality
- Calculate Theil Index for gender
- Widely used to measure economic inequalities
- US Census Bureau uses to measure Income inequality

Conclusion

1. How to evaluate if our model/dataset are bias?
   - ☆ Introducing 5 metrics
   - ☆ Could see more standards for fairness
2. How to mitigate bias?

Post-Processing	In-Processing	Pre-Processing	Data Collection
• Change thresholds • Trade off accuracy for fairness	• Adversarial training • Regularize for fairness • Constrain to be fair	• Modify labels • Modify input data • Modify label/data pairs • Weight label/data pairs	• Identify lack of examples or variates and collect

May 27, 2021

Thank you!

Happy weekend!