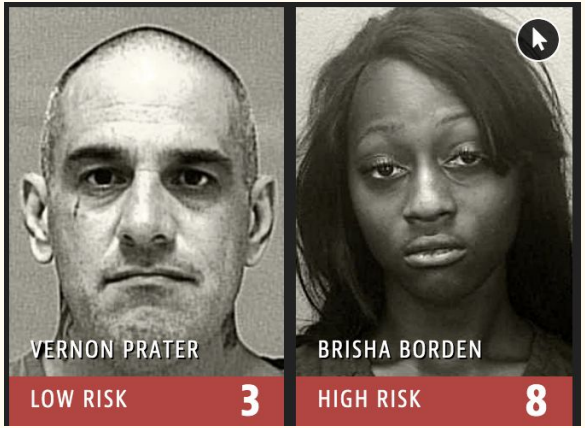
Criminal Justice Bias Investigation

—

Naomi Vaid + Inspirit AI

How Do We Use AI in the Justice System?

- ★ **COMPAS** is a risk -assessment tool that attempts to predict whether a defendant is likely to reoffend (recidivate).
- ★ It is used in the pre-trial detention phase.
- ★ Based on particular characteristics about a defendant, the model would predict whether or not someone is likely to commit another crime in the next two years if released.
- ★ This software has been used in many U.S. jurisdictions, including New York, Wisconsin, California, and Broward County, FL.



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

The Data

Columns We Used:

sex, age, age_category, race, juveline_felony_count, juveline_misdemeanor_count, juveline_other_count, prior_convictions, current_charge, charge_description, *recidivated_last_two_years*

Most common charges:

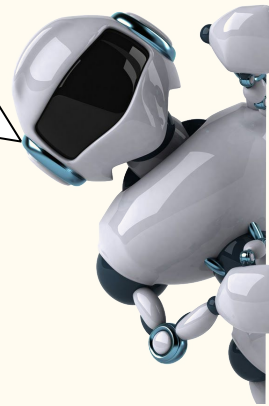
- Battery
- Arrest case no charge
- Possession of Cocaine

There are 7,214 data points

- Predominantly Caucasian and African-American defendants
- Most of the defendants here are relatively young as well (20s to 30s)
- Predominantly men

Machine Learning

By using Machine Learning, we are able to train our data to have a less biased approach to deciding sentences. The goal is to help the judge make the sentence unbiased and as fair as possible for the person that is being accused.

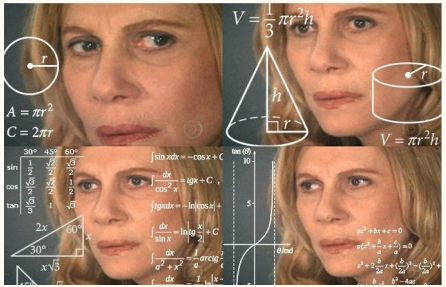


Ways to Measure Our Model

Accuracy using Logistic Regression: 0.65 - 0.70 → pretty bad

Accuracy using RandomForest: 0.702 → not any better

The accuracy is not that great, but at least the accuracy is equal among races.



Is that fair?

Fairness

TRUE POSITIVE= Predicted *to reoffend*, **reoffend**

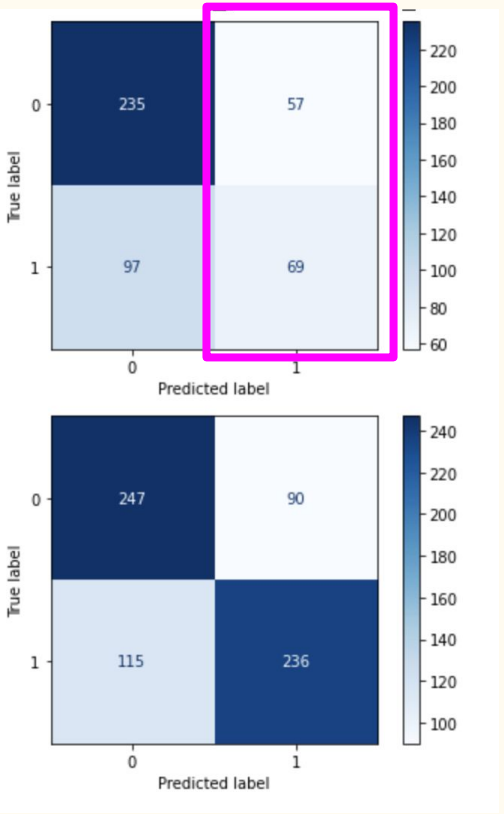
FALSE POSITIVE= Predicted *to reoffend*, **don't reoffend**

TRUE NEGATIVE= Predicted *to not reoffend*, **don't reoffend**

FALSE NEGATIVE= Predicted *to not reoffend*, **reoffend**

Using a confusion matrix:

- ➔ *Caucasian Group*: 0.153 False Positive Rate, 0.661 False Negative Rate
- ➔ *African American Group*: 0.333 False Positive Rate, 0.350 False Negative Rate
- ★ **The model is unfair**. The African American group is experiencing a *False Positive Rate* over 2 times that of the Caucasian group.



Testing Fairness

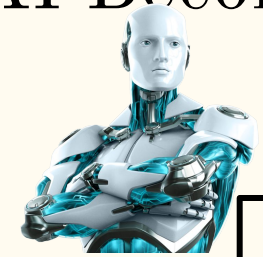
We defined fairness as - equal standards applied to everyone *independent of any personal or identifying characteristics*. It means speedy and thorough process for all people and presumption of innocence.

Group Fairness

Calibration

Is accuracy or fairness most important?

How Does AI Become Biased



Data is biased

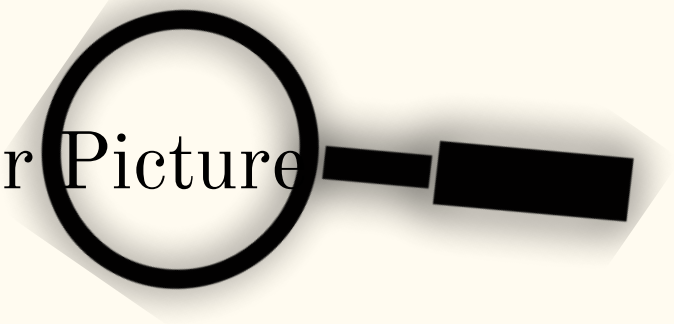
Bias in who is convicted

Bias in what charges people get

Bias in who gets arrested

Bias in policing - neighborhoods, training, etc.

The Bigger Picture



There is bias in other aspects of AI →

- Skin disease detection is only trained on Caucasian skin - that's a problem.
- https://metro.co.uk/2020/07/09/black-medic-taught-how-diagnose-white-patients-creates-handbook-show-how-conditions-look-darker-skin-12966052/
- https://www.naacp.org/criminal-justice-fact-sheet/
- gendershades.org

KAHOOT!

Go to Kahoot.it and join using the code!