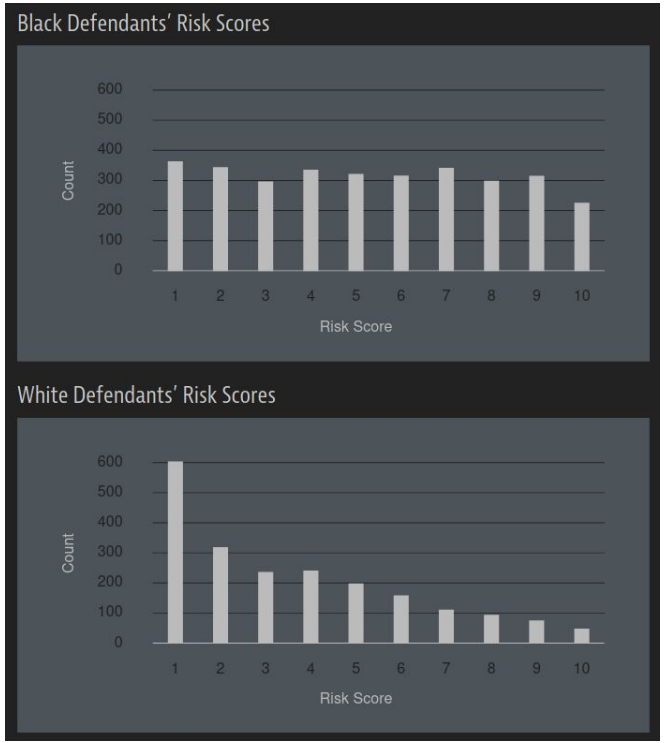
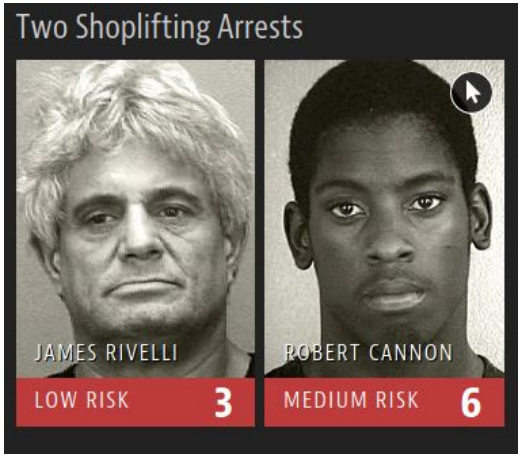
AI + Criminal Justice

Ananya Maddali, Vinay Goyal, Dani del Rosal,
Kensay Sato, Lawrence Kang, Andrew Cho

Do you think AI is currently used in the American Criminal Justice system?

COMPAS: A Case Study





Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Our dataset

- Our dataset is made up of **7214 arrests** from the public records of **Broward County, Florida**
- Original Dataset = **53 columns**; shortened Dataset = **11 columns**:

	sex	age	age_category	race	juveline_felony_count	juveline_misdemeanor_count	juveline_other_count	prior_convictions	current_charge	charge_description	recidivated_last_two_years
0	Male	69	Greater than 45	Other	0	0	0	0	F	Aggravated Assault w/Firearm	0
1	Male	34	25 - 45	African-American	0	0	0	0	F	Felony Battery w/Prior Convict	1
2	Male	24	Less than 25	African-American	0	0	1	4	F	Possession of Cocaine	1
3	Male	23	Less than 25	African-American	0	1	0	1	F	Possession of Cannabis	0
4	Male	43	25 - 45	Other	0	0	0	2	F	arrest case no charge	0

- Race categories: **8 groups**, but we're looking specifically at **African American** and **Caucasian** because they appear the most
- Looking for bias in the model; on which groups do the models make the most errors in predicting if they'll recidivate

Simple model performance

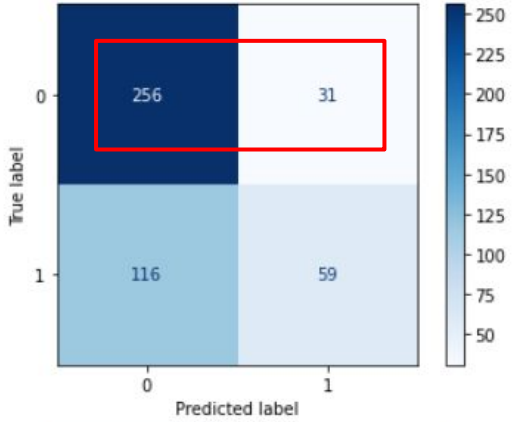
- Using Logistic Regression from scikit-learn we build, fit, and test our model which gives us accuracy between **0.65** and **0.70.**
- The COMPAS has accuracy around ~ **0.63.**
- The accuracy we get using Random Forest Classifier → **0.70.**

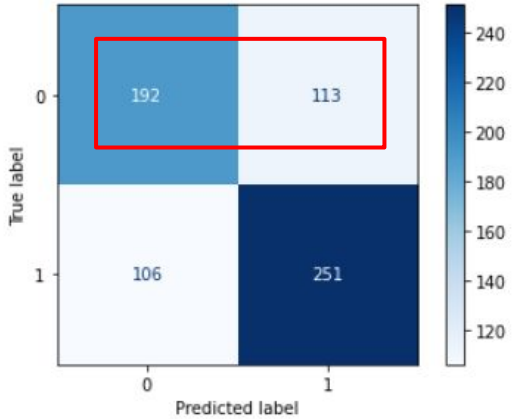
Is the Accuracy enough?

Racial bias

- Measured initial bias between races using false positive recidivation rates given by confusion matrices
- False Positive Rate = FP/(FP + TN)
- False Positive Recidivation: predicting the defendant will commit a crime after being released, but they did not

caucasians





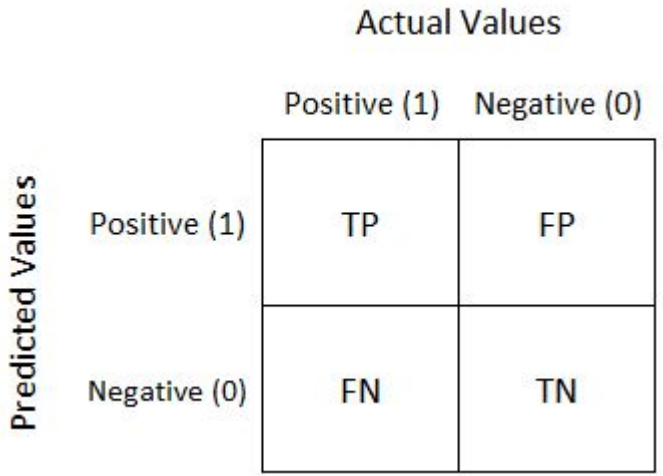
African-Americans

Fairness

*FPR: predicted to reoffend, did not reoffend; FNR: predicted to not reoffend, did reoffend

- Caucasian group: **0.108 FPR; 0.663 FNR**
- African American group: **0.370 FPR; 0.297 FNR**

How can we define fairness?



A mathematical definition of fairness

> *We chose to define fairness as equal standards being applied to everyone, regardless of personal or identifying characteristics. It means speedy and thorough due process for all people and presumption of innocence.*

Group fairness: $\mathbf{P(\hat{y}=1|R=a) = P(\hat{y}=1|R=c)}$

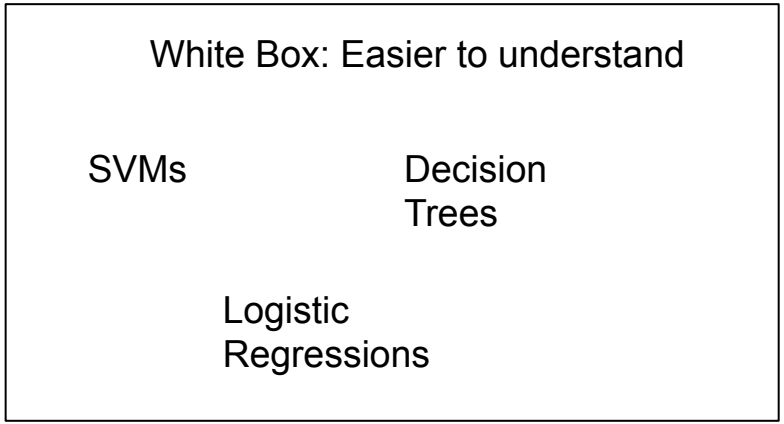
- P(A|B): the probability of event A given event B
- $\hat{y}$: model's predicted label
- R(a, c): protected features such as race

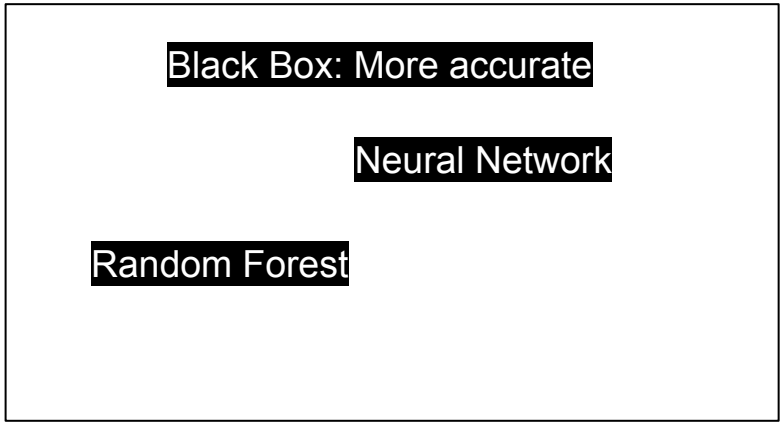
Caucasian acceptance rate = **0.195**; African american acceptance rate = **0.550**

Margin: **25%**

Model interpretability

- Interpretability: the extent in which a human understands the model's decision process (what, why, how)
- Accuracy (what) vs Interpretability
- Interpretability allows to assess if the model is **fair** or not



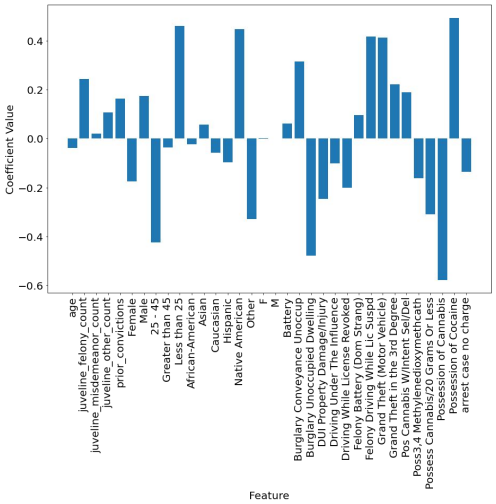


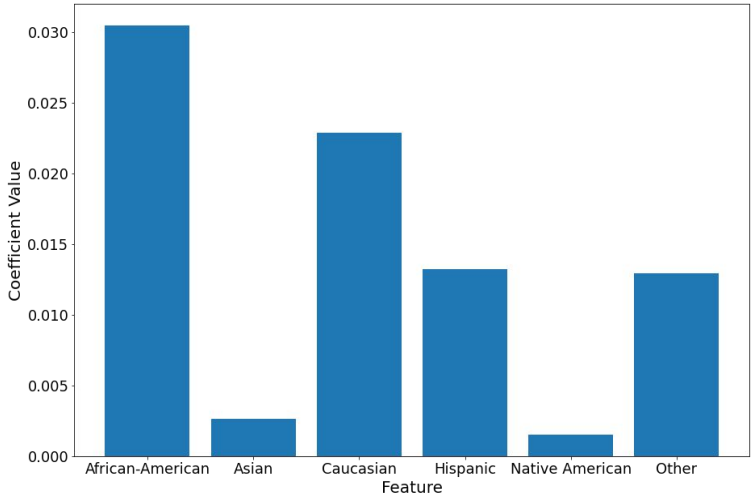
Importance of protected features in different models

#1: Linear SVMs:
Training accuracy: *0.68*
Testing accuracy: *0.67*

#2: Random Forest:
Training accuracy: *0.78*
Testing accuracy: *0.69*

#3: Neural Networks:
Training accuracy: *0.72*
Testing accuracy: *0.67*





- Researchers are finding ways to understand NN
- interpretability -accuracy trade-off

- Protected features: **race**, gender, age
- Different impact in different models

Building fairer models

- Our simple models still were biased and inaccurate
- Getting rid of race:
  - Slightly helped model in terms of fairness
  - But still wasn't perfectly accurate; we wanted to look for something more
- Separating by race:
  - The model was more fair & accurate
  - Not significantly so, but this can be fine-tuned

Ethical dilemma

In the models separated by race:

- Is it ok to separate people based on race?
- Makes the assumption that all people from a certain race have the same likelihood of committing/not committing a crime

Conclusion

- Initial models displayed racially unfair results
- Random forest classifiers produced the most accurate results, but is hard to interpret
- There is still work to be done on improving accuracy and fairness in AI criminal justice applications
- Was our dataset already biased?

What questions do you have?