Predicting Recidivism Rates

Jessie, LA, Stephanie

Problem Statement

In this project, we are conducting analysis on behalf of the Department of Justice to investigate the validity of recidivism prediction models. Specifically, we are interested in pinpointing where racial bias occurs. To this end, we will make recommendations to the Department of Justice, based on our findings, on whether or not it's advisable for states to continue using recidivism models such as COMPAS.

Data Preparation, EDA

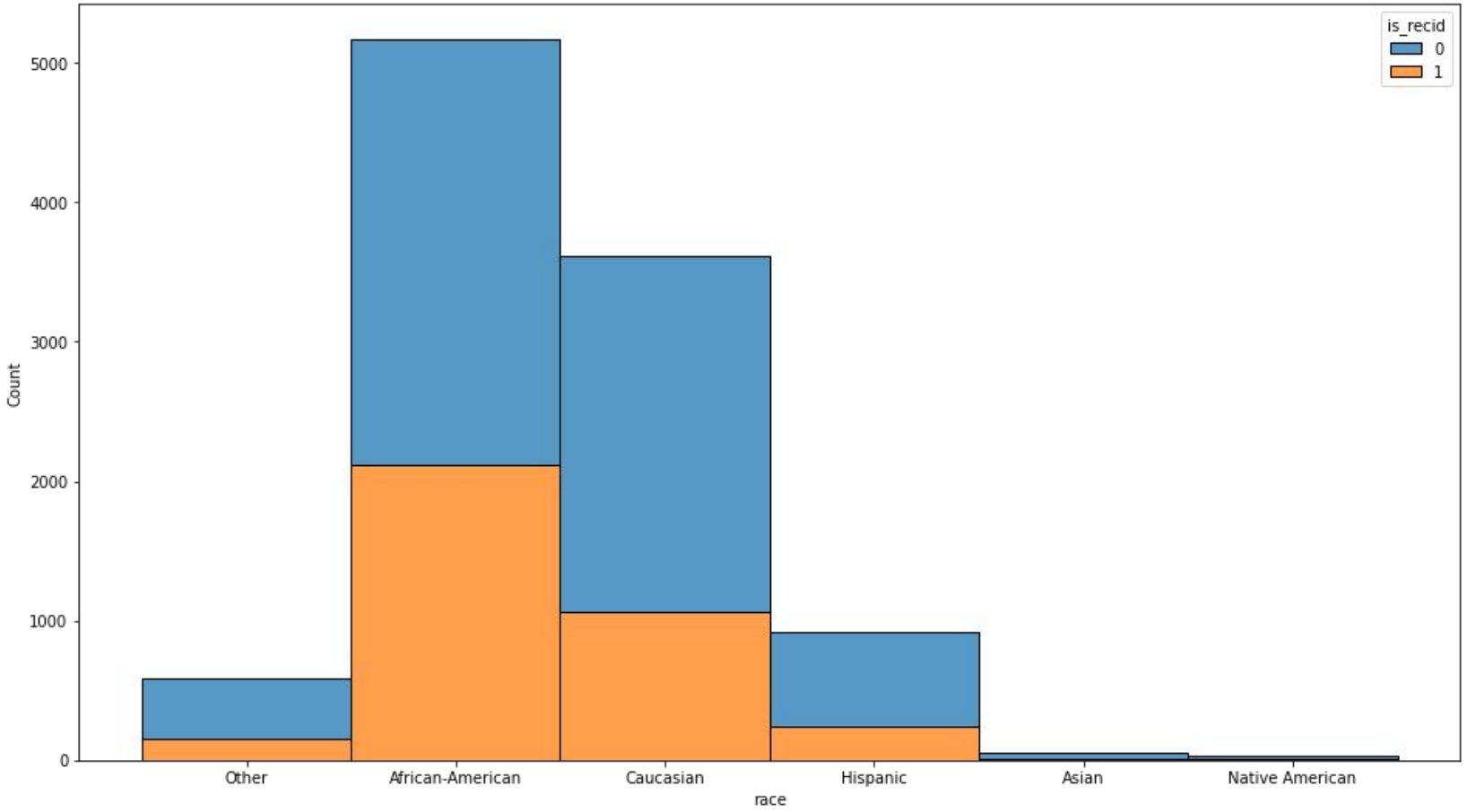
Data Collection

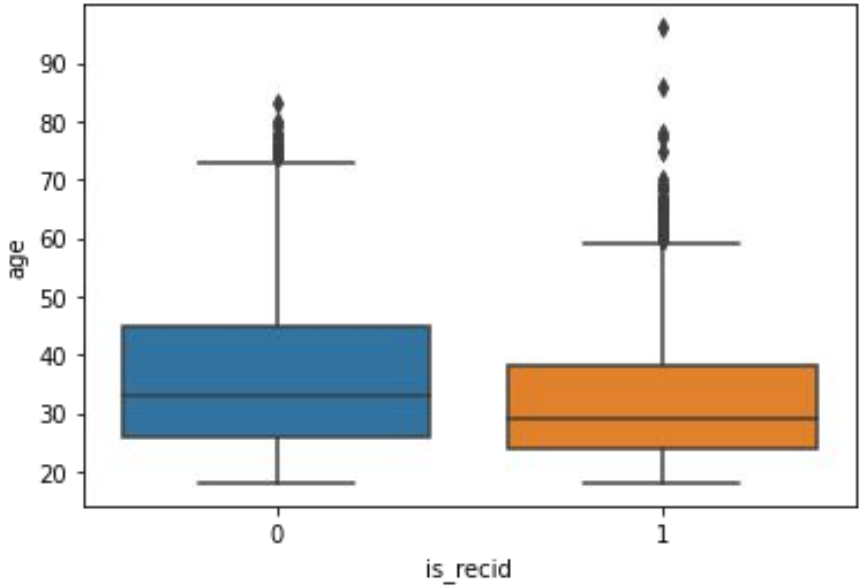
- SQLite Database from ProPublica
- Defendants from Broward County, 2013-2014

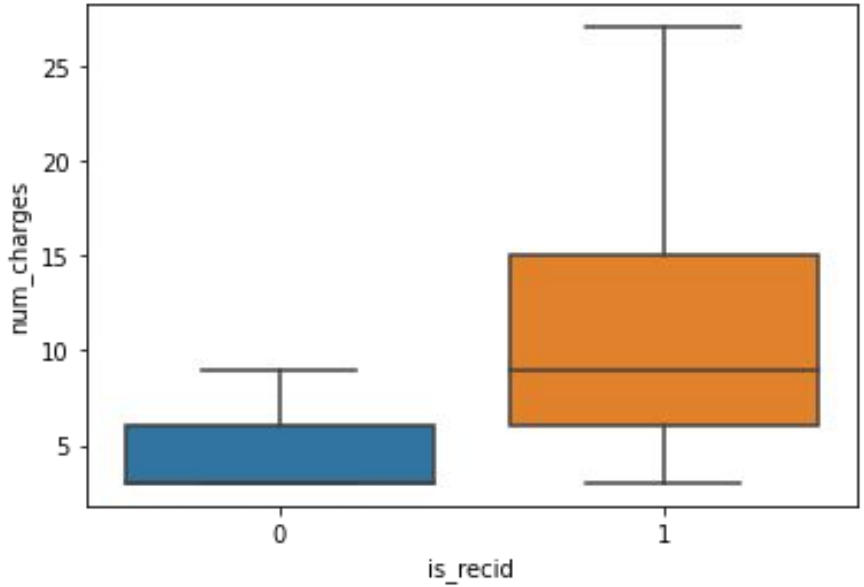
Columns:

- Demographics: sex, age, race, marital status
- Prior offenses, juvenile offenses
- Charge degree (felony vs misdemeanor), times in/out of custody (jail & prison)

Challenge: missing data dictionary


is_recid
0
1
5000
4000
3000
2000
1000
0
Count
Other
African-American
Caucasian
Hispanic
Asian
Native American
race


age
90
80
70
60
50
40
30
20
0
1
is_recid


num_charges
25
20
15
10
5
0
1
is_recid

Modelling

model	train	test
Bernoulli NB	0.7550	0.7564
Gaussian NB	0.6972	0.6955
Decision Tree	0.7733	0.7722
KNN	0.8445	0.7205
Logistic Regression	0.7313	0.7409
Random Forest	0.6913	0.6985
AdaBoost	0.7689	0.7691
Gradient Boosting	0.7747	0.7679
Multilayer Perceptron	0.7702	0.7718
Quadradic Discrimination	0.6524	0.6491
XGBoost	0.8045	0.7656

	Train	Test
Decision Tree	0.7648	0.7652
Random Forest	**0.7727**	**0.7656**
XGBoost	0.8045	0.7655

Discussion

Addressing Representation in Workflow

- Frequent Discussions
- Didn’t have control of data collection
- African-Americans over-represented
- Caucasians under-represented
- Data only included Florida corrections

Proxies for race

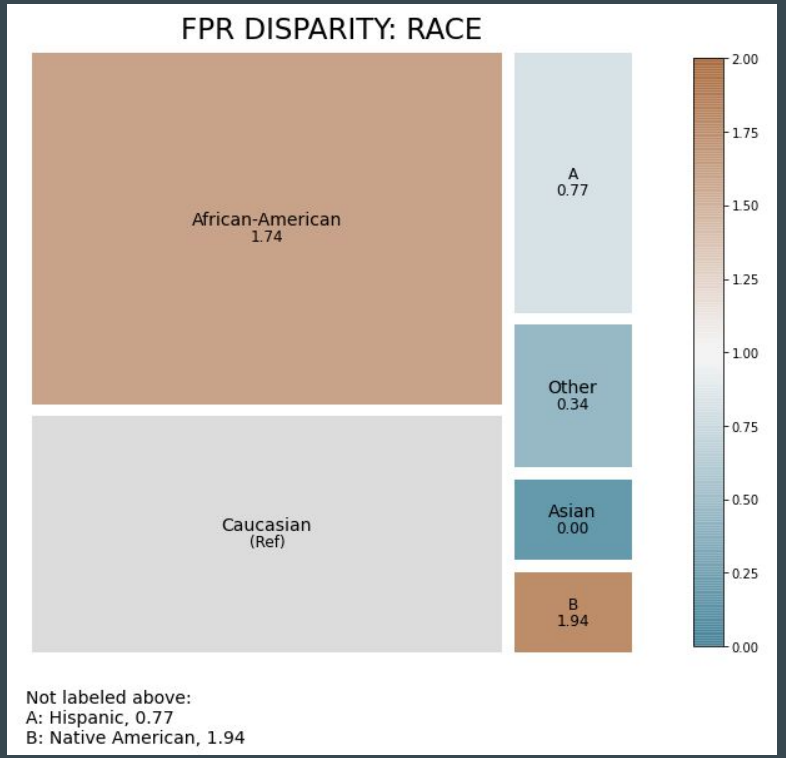
- Systemic Bias
- Priors count, juvenile counts

Analysis of findings

- 22% FPR - African-Americans
- 9% FPR - Hispanic
- 13 % FPR - Caucasian

- 32% FNR - African-Americans
- 49% FNR - Hispanic
- 44 % FNR - Caucasian

Fairness Threshold = 1.25



Next steps/Conclusions

- Collect better data
- More data for smaller populations
- Try under/over sampling
- Account for proxies of race