Data Ethics: Criminal Recidivism Risk Analysis for Parole Decisions

June 2019

Presented by:

Apoorva Srinivas
Harsha Wardhan Reddy Duvvrv
Pratik Khandelwal
Prerana Patil
Ritumbhra Sagar

Agenda

- ➔ Problem Statement
- ➔ Data Exploration
- ➔ Analysis & Findings
- ➔ Conclusion

Problem Statement

Problem Statement

To test whether machine learning model is biased against certain groups of people (Gender and Race).

1. Build a machine learning model to predict crime re-offenders
2. Identify discrimination (bias)
3. Remove discrimination (bias)

Image Source: https://californiainnocenceproject.org/issues-we-face/recidivism-rates/

Dataset & Variables

About Dataset

Data Source: https://github.com/propublica/compas-analysis/

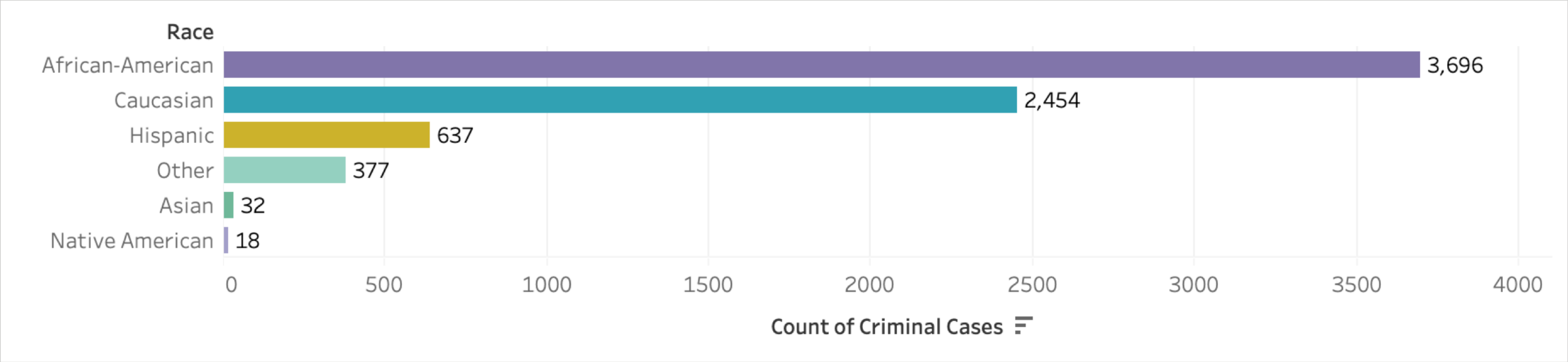
Data Collection Period: January 1, 2013 to September 9, 2014

Key Variables of Interest

Y Variable (Label)	is_recid
X Variables (Features)	sex, age, race, juv_fel_count, juv_misd_count, juv_other_count, priors_count, days_b_screening_arrest, c_charge_degree
Bias Variables (Features)	sex, race

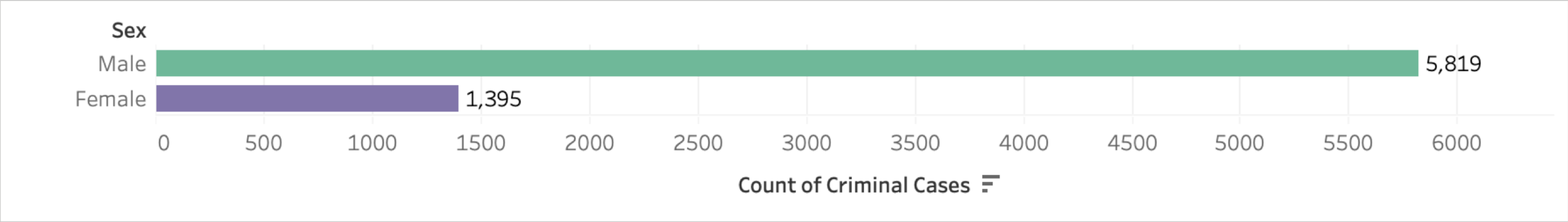
Data Exploration

Exploratory Analysis



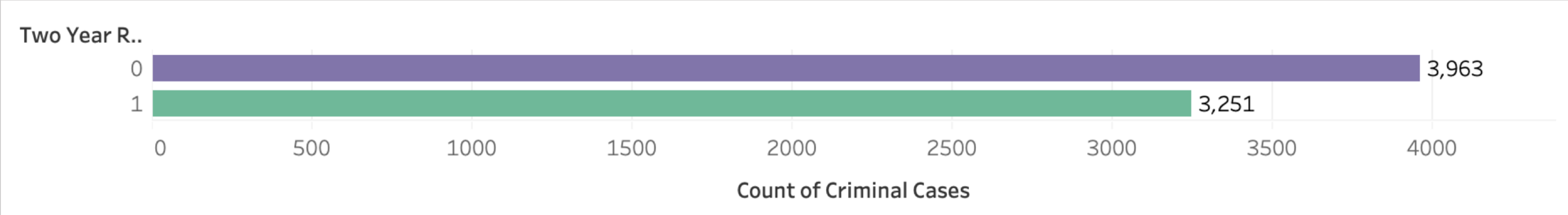
	Number of Records	Percentage
Total number of records	7214	-
African-American	3696	51%
Caucasians	2454	34%
Hispanic, Asian, Native American and Others	1064	15%

Exploratory Analysis



	Number of Records	Percentage
Total number of records	7214	-
Male	5819	80%
Female	1395	20%

Exploratory Analysis



	Number of Records	Percentage
Total number of records	7214	-
Will Not Reoffend	3693	55%
Will Reoffend	3251	45%

Model & Analysis

Measuring Potential Discrimination

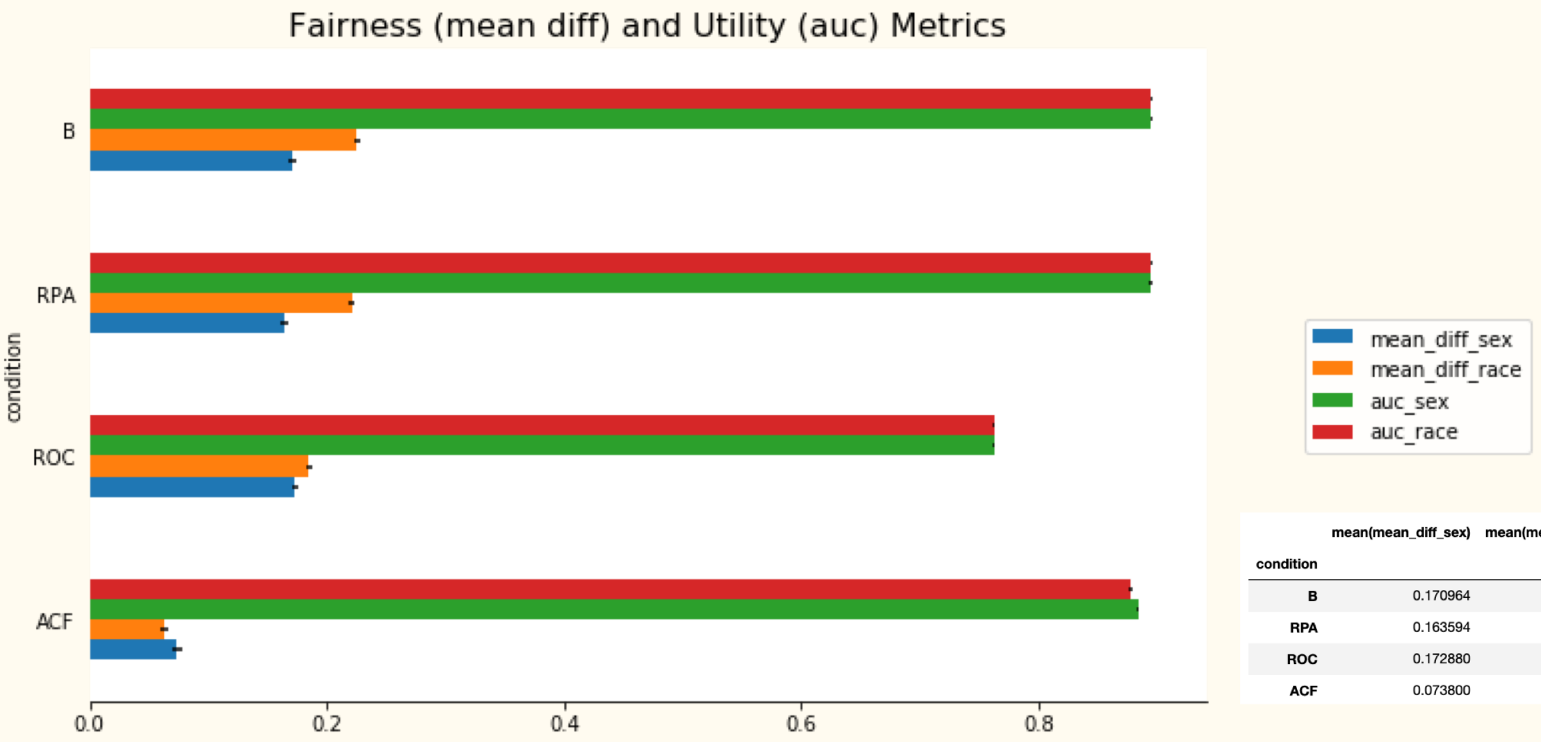
- Mean difference scores:
  - protected class = sex: 0.139, 95% CI [0.107-0.170]
  - protected class = race: 0.139, 95% CI [0.113-0.165]
- The mean differences above suggest that Men and African-Americans are more likely to reoffend compared to Women and Caucasians respectively.
- This suggests that there is a bias against men and African-Americans.

Classifiers Used

- **Baseline (BB):** Train a model on all input variables, including protected attributes.
- **Remove Protected Attribute (RPA):** Train a model on input variables without protected attributes. This is the naive fairness-aware approach.
- **Reject-Option Classification (ROC):** Train a model using the Reject-option Classification method.
- **Additive Counterfactually Fair Model (ACF):** Train a model using the Additive Counterfactually Fair method.

Conclusion & Limitations

Classifier Outputs: Fairness and Utility Tradeoff



condition	mean(mean_diff_sex)	mean(mean_diff_race)	mean(auc_sex)	mean(auc_race)
B	0.170964	0.225361	0.894819	0.894819
RPA	0.163594	0.220592	0.894391	0.895070
ROC	0.172880	0.184840	0.762372	0.762372
ACF	0.073800	0.062998	0.883807	0.877298

Classifier Outputs - Table

condition	mean(mean_diff_sex)	mean(mean_diff_race)	mean(auc_sex)	mean(auc_race)
B	0.170964	0.225361	0.894819	0.894819
RPA	0.163594	0.220592	0.894391	0.895070
ROC	0.172880	0.184840	0.762372	0.762372
ACF	0.073800	0.062998	0.883807	0.877298

Limitations

- Data is for two years only and is not recent
- Features available are limited
- Data has been pre-processed (Not sure of prior assumptions)

Thank You.