Analyzing the Relationship Between Race and Gender and Predicting the Rate of Crime: A DC Case Study

Kholeigh Foster
Data 303: Applied Machine Learning, Bias, and Ethics
February 28, 2024

Justification

- This research aims to unpack the intersection between the criminal justice system, race, gender, and crime within the US.
    - Focusing on the disproportionate impact on minority communities
- Criticism of traditional criminological theories due to lack of consideration for race and gender
    - Blasdell, R. (2015) proposed a theoretical model that incorporates intersection of race and gender with *Routine Activity Theory*
        - An accepted theory of criminal offending and victimization
- Race and Gender are structuring forces that impact:
    - how people behave, how others classify and react to those behaviors, who has the power to classify and label behaviors, and how the legal system is organized and attentive to control behavior (Blasdell, 2015)
- Understanding the influences of race and gender on arrest outcomes is a crucial step in addressing disparities in the criminal justice system
    - Lofstrom and Raphael(2016) emphasize the importance of a nuanced approach

Insights to Race, Gender, Crime, and Prevention Strategies

Lofstrom, M., & Raphael, S. (2016)

- Explores theories behind declining crime rates, changes to criminal justice policies, and demographic shifts
- Highlights the disproportionate impact of crime and incarceration rates on lower income and minority communities

Steward, D., & Stolzenberg, L (2003)

- Discussed Race and the Probability of Arrest
- Used multivariate logistic regression model
- Findings suggest disproportionately high arrest rate for black citizen is mostly attributed to involvement in reported crime

Narayan, P.K., Nielsen, I., & Smyth, R. (2010)

- Discusses the concept of the "natural rate of crime"
- Explores the relationship between crime rates and sanctions
- Suggests addressing root causes of criminal motivation as an approach

According to District Crime Data Comparison, about 26% of all crimes committed in DC are Violent Crimes.

Abstract

Investigate the impact of race and gender on the likelihood of individuals being arrested

- Utilized the *Adult Arrests* dataset from Open Data DC
- Employed machine learning techniques by using a logistic regression model
  - Analyzed patterns between demographic variables and arrest rates with a specific focus on Simple Assault
- Contribute to understanding patterns and correlations in arrest data.
- Inform efforts and assist in implementation techniques for circumventing crime

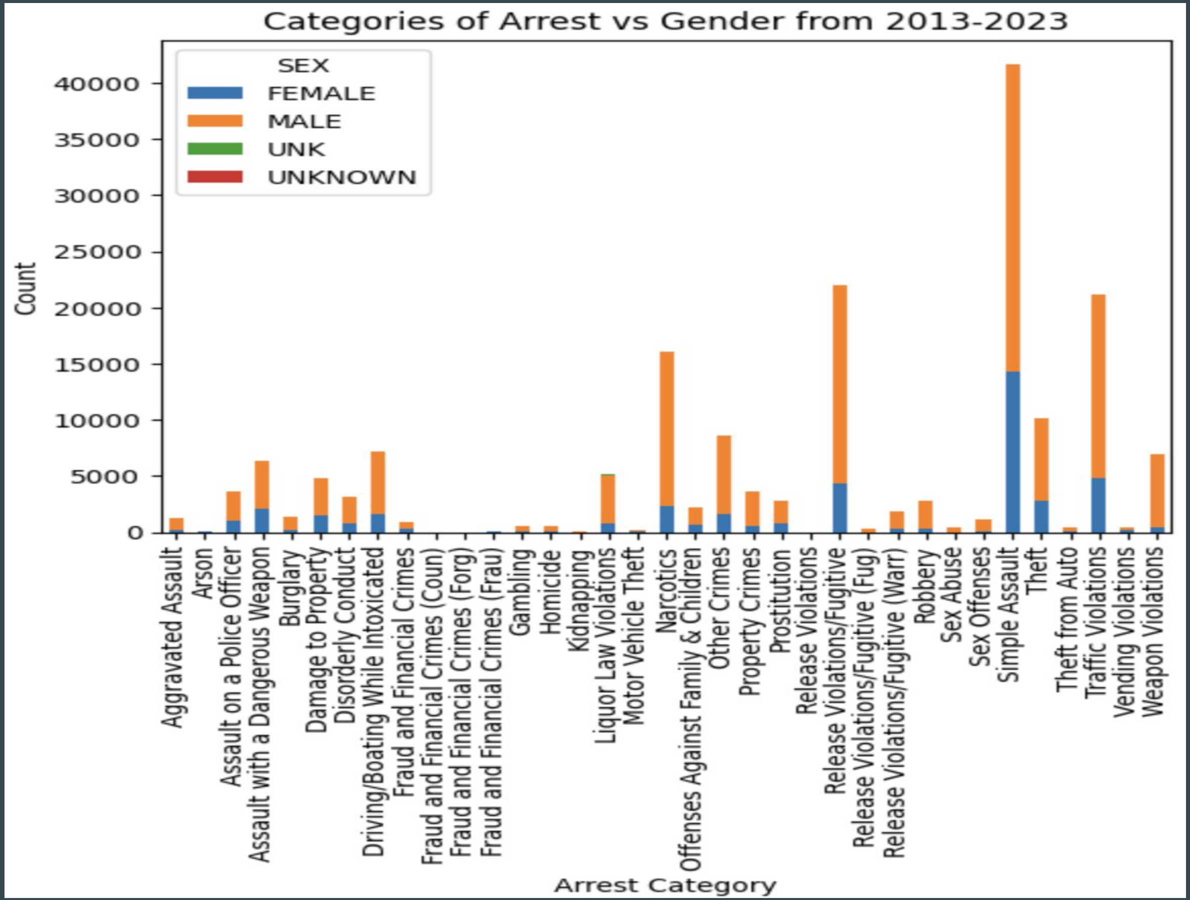
The Dataset: Adult Arrests

Adult arrests in DC from 2013-2023

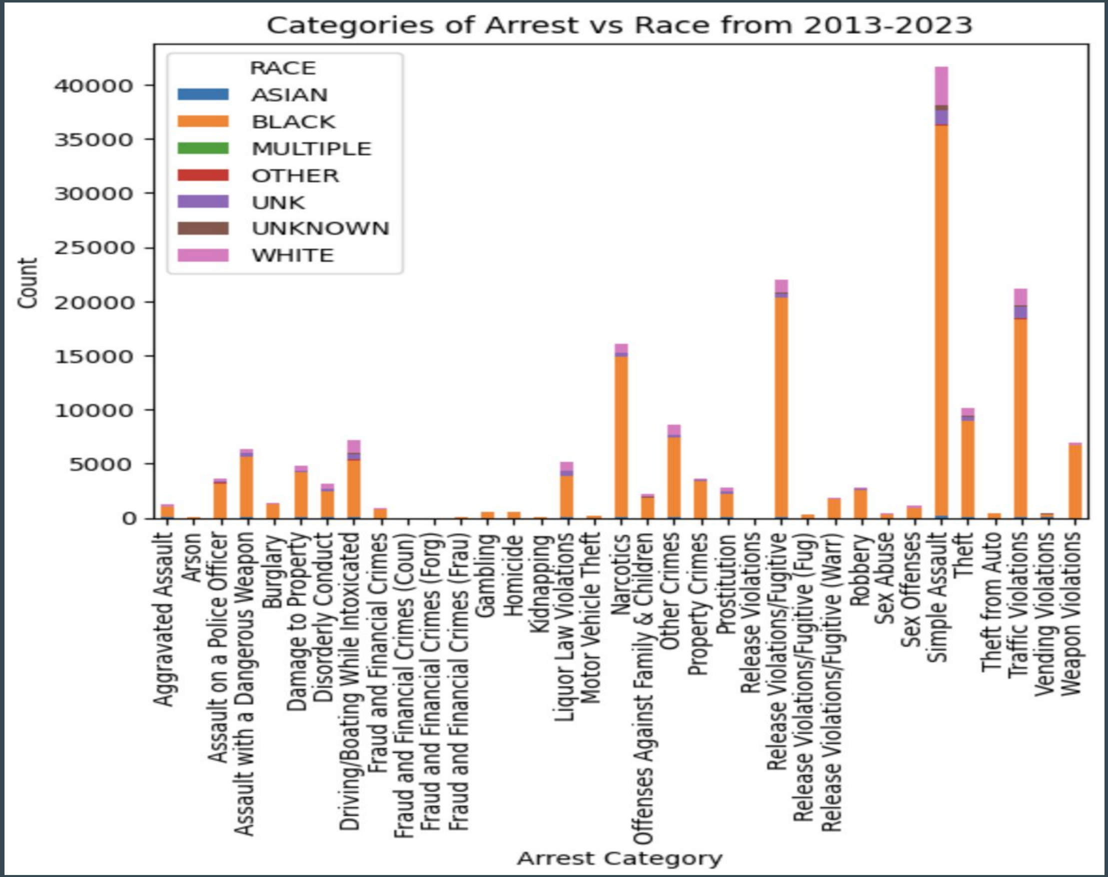
- This dataset included arrest made by the Metropolitan Police Department(MPD).
- Noted Variables:
  - Arrest Year [int]
  - Arrest Date [str]
  - Arrest Hour [int]
  - Arrestee's Age [int]
  - Defendant Race [str]
  - Defendant's Ethnicity [str]
  - Arrest Category [str]
  - Arrest Location District [str]
  - Offense District [str]
- 259,875 rows

Methodologies

- Data Preparation:
  - Dropped columns that were not essential to my analysis
    - Focused on ‘Race’ and ‘Sex” as independent variables
  - Checked for missing values and dropped rows with NaN values
  - Created a dictionary to correct misspelled words in the ‘CATEGORY’ column
- Data Analysis:
  - Subset data to focus on desired category of arrest ‘Simple Assault’
    - Simple assault was labeled as the largest category of arrests within the dataset
  - Encoded categorical variables using one-hot encoding
  - Split the adult.arrests dataset into training and testing sets
- Model Training and Evaluation:
  - Trained a logistic regression model using training dataset
  - Calculate model’s accuracy on testing data, calculated the ROC curve and AUC to assess model performance
  - Regularization using Lasso and Ridge techniques
  - Evaluate statistical models



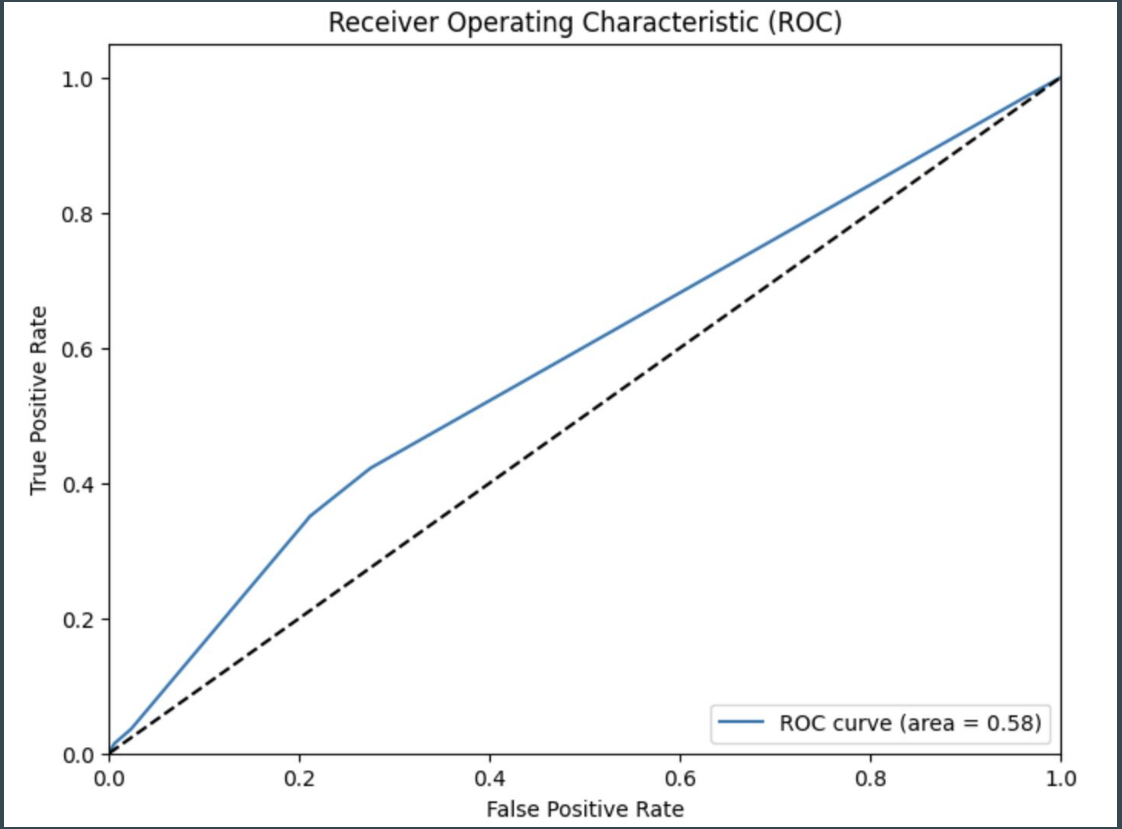
Independent Variable: Gender



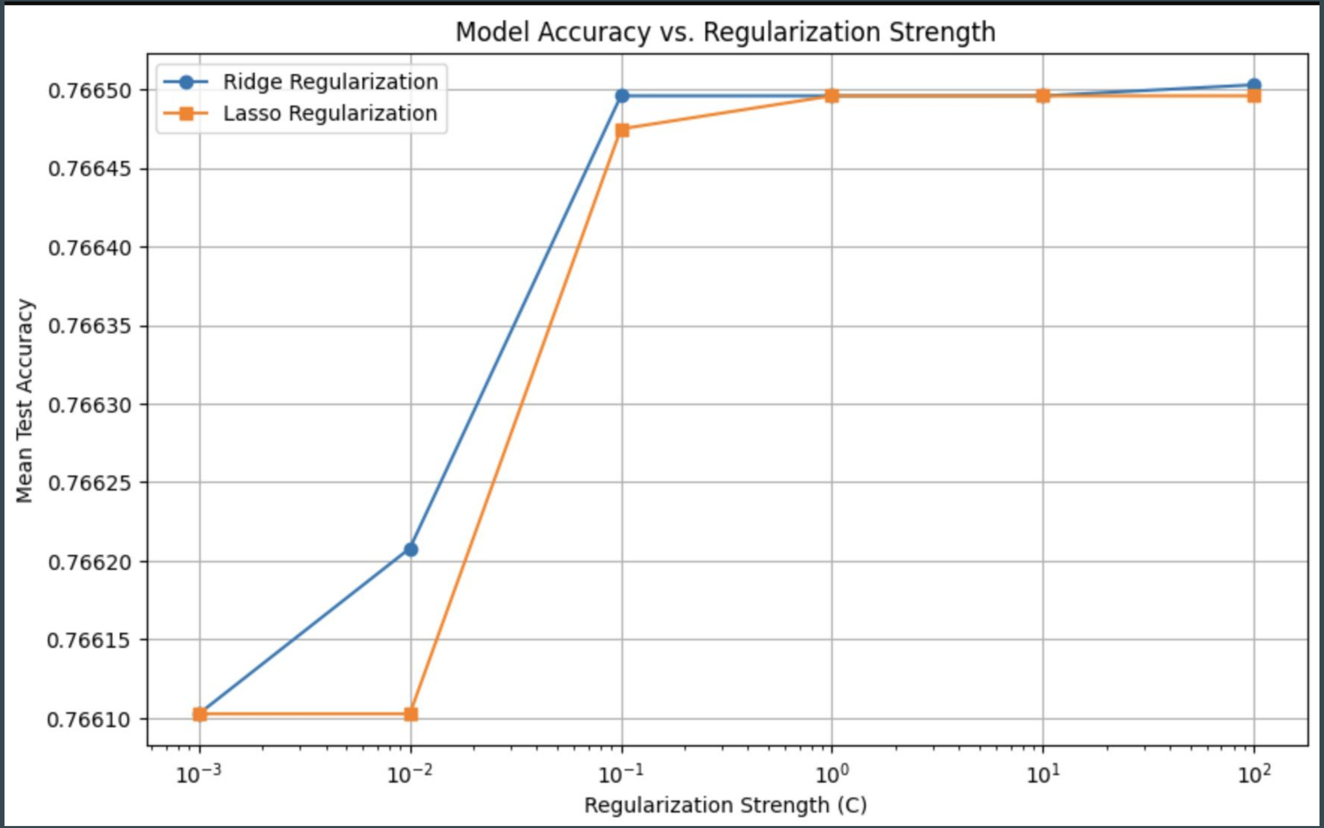
Independent Variable Race:

Results: Logistic Regression

- Approximately 76.3% of the model's predictions were correct
- Area Under the Receiver Operating Characteristic Curve was 0.58.
- True Negatives: 27,154
  - Correctly classified as not being Simple Assault
- False Positive: 27
  - Incorrectly classified as Simple Assault
- False Negatives: 8,378
  - Instances were incorrectly classified as not being Simple Assault when they were
- True Positive: 40
  - Instances correctly classified as being Simple Assault



Model Accuracy: 0.7738
ROC AUC: 0.5773



Model Accuracy vs. Regularization

```python
# Create a DataFrame for predictions
predictions_df = pd.DataFrame({
    'Predicted_Probability': y_prob,
    'Actual_Target': y
})

# Display the predictions DataFrame
print(predictions_df)
```

```
        Predicted_Probability  Actual_Target
13                   0.198710              0
15                   0.198710              0
17                   0.336711              0
18                   0.198710              0
19                   0.198710              1
...                       ...            ...
259870               0.198710              0
259871               0.198710              0
259872               0.336711              1
259873               0.198710              0
259874               0.198710              0

[177942 rows x 2 columns]
```

Predicted Probability

```
Optimization terminated successfully.
         Current function value: 0.534065
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:               142353
Model:                          Logit   Df Residuals:                   142343
Method:                           MLE   Df Model:                            9
Date:                Fri, 01 Mar 2024   Pseudo R-squ.:                 0.01816
Time:                        16:58:40   Log-Likelihood:                -76026.
converged:                       True   LL-Null:                       -77432.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const            -0.6577      0.091     -7.202      0.000      -0.837      -0.479
RACE_BLACK       -0.0200      0.091     -0.220      0.826      -0.199       0.159
RACE_MULTIPLE     0.5367      0.549      0.977      0.328      -0.539       1.613
RACE_OTHER        0.6680      0.251      2.663      0.008       0.176       1.160
RACE_UNK         -0.0867      0.097     -0.891      0.373      -0.277       0.104
RACE_UNKNOWN      0.8764      0.112      7.793      0.000       0.656       1.097
RACE_WHITE        0.0886      0.094      0.947      0.344      -0.095       0.272
SEX_MALE         -0.7167      0.014    -51.903      0.000      -0.744      -0.690
SEX_UNK          -0.3657      0.274     -1.335      0.182      -0.903       0.171
SEX_UNKNOWN      -0.9336      0.401     -2.328      0.020      -1.720      -0.147
=================================================================================
```

Summary of Logistic Regression Results

Implications and Improvements

- The AUC of 0.58 suggests that the model's ability to correctly discriminate between positive and negative cases is limited.
- Although the accuracy is relatively high, the AUC is not as strong.
    - It is not as effective at distinguishing between the specific classes (Simple Assault and other arrest types).
- Assist in creating and informing community based crime prevention strategies based on specific crime types
    - May provide a starting point for understanding a frequency of specific types of crime
    - Creating tailored partnerships with community stakeholders, government, and the community
- Provide insight on the influence of race and gender on one's likelihood to be arrested for a specific crime
- In the future:
    - Explore additional features or variables that may have more predictive power:
        - Factors such as socioeconomic status, and neighborhood characteristics
    - Check the classes of my dataset
        - Ex: compare and contrast arrest types
        - Consider oversampling or undersampling techniques to ensure balanced dataset
    - Experiment with various Machine Learning algorithms

Works Cited

Blasdell, R. (2015). The intersection of race, gender, and class in routine activities: A proposed criminological model of victimization and offending. *Race, Gender & Class*, *22*(3–4), 260–273. https://www.jstor.org/stable/26505360

D'Alessio, S. J., & Stolzenberg, L. (2003). Race and the Probability of Arrest. *Social Forces*, *81*(4), 1381–1397. http://www.jstor.org/stable/3598118

Lofstrom, M., & Raphael, S. (2016). Crime, the Criminal Justice System, and Socioeconomic Inequality. *The Journal of Economic Perspectives*, *30*(2), 103–126. http://www.jstor.org/stable/43783709

KLINGELE, C. (2019). MEASURING CHANGE: FROM RATES OF RECIDIVISM TO MARKERS OF DESISTANCE. *The Journal of Criminal Law and Criminology (1973-)*, *109*(4), 769–817. https://www.jstor.org/stable/48572943

Narayan, P. K., Nielsen, I., & Smyth, R. (2010). Is There a Natural Rate of Crime? *The American Journal of Economics and Sociology*, *69*(2), 759–782. http://www.jstor.org/stable/40607844

DER, R. H., & KITCHEN, T. (2013). Putting Crime Prevention Through Environmental Design into Practice via Planning Systems: A Comparison of Experience in the US and UK. *Built Environment (1978-)*, *39*(1), 9–30. http://www.jstor.org/stable/43296830