Analysis of Arrests

By Dylan Garsee

The Terry Stop

In the 1968 Supreme Court case "Terry v. Ohio", the court found that a police officer was not in violation of the "unreasonable search and seizure" clause of the Fourth Amendment after he stopped and frisked suspects only because their behavior was suspicious. Thus the phrase "Terry Stops" are in reference to stops made of suspicious drivers.

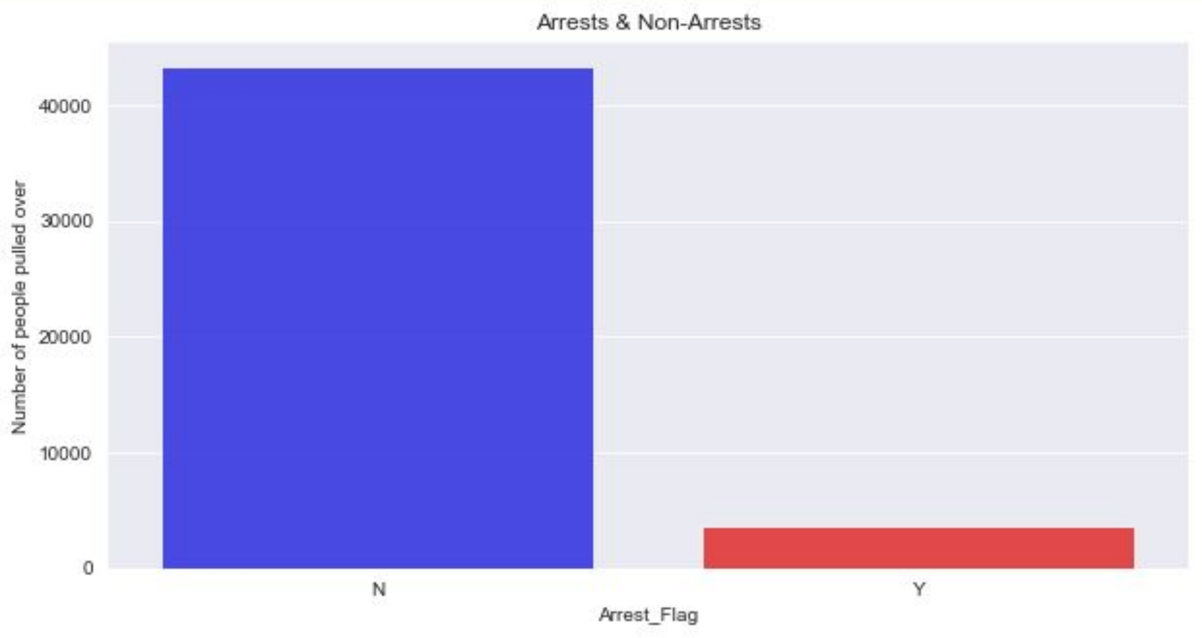
Goals of Analysis

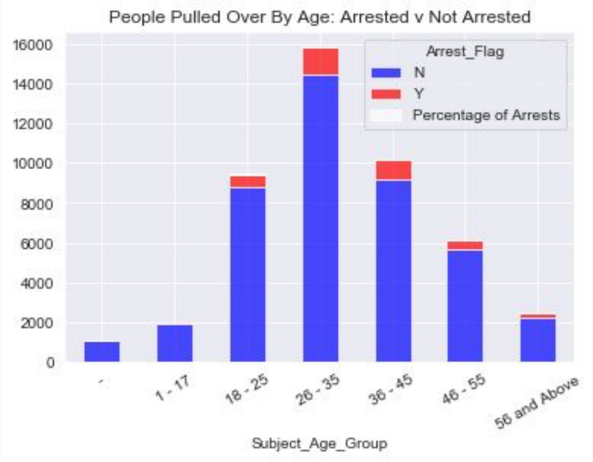
This is an analysis of over 48,000 Terry Stops, with a goal of predicting if an arrest will be made based off time of day, whether a suspect was frisked, and racial & gender demographics of both the suspects and officers.

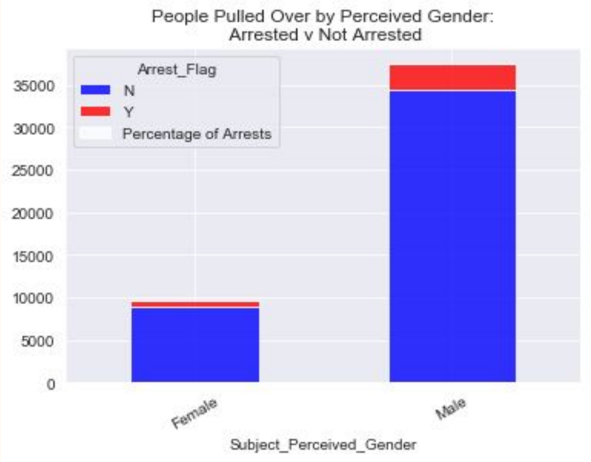
The overall goal of the analysis is to have the highest possible recall, to minimize false positives, of accidentally classifying subject who were not arrested as arrested.

Initial Analysis

The dataset used in this analysis had over 48,000 entries of Terry Stop data. Of the 48094 entries, 3701 of the stops ended in with an arrest of the subject, or a little over 8% of total stops.







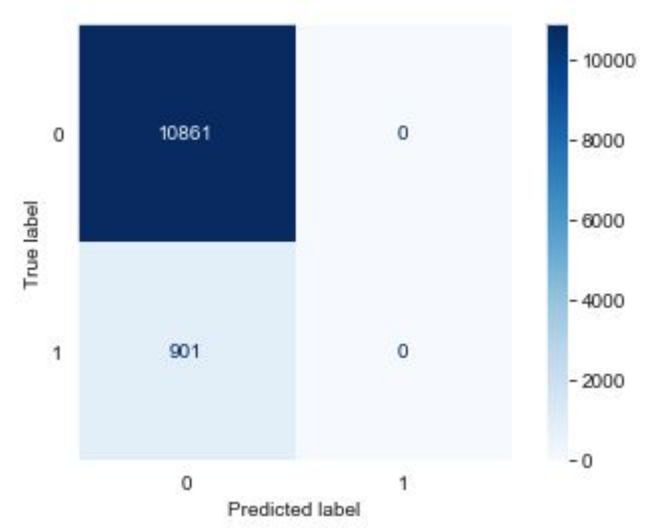
Initial Analysis

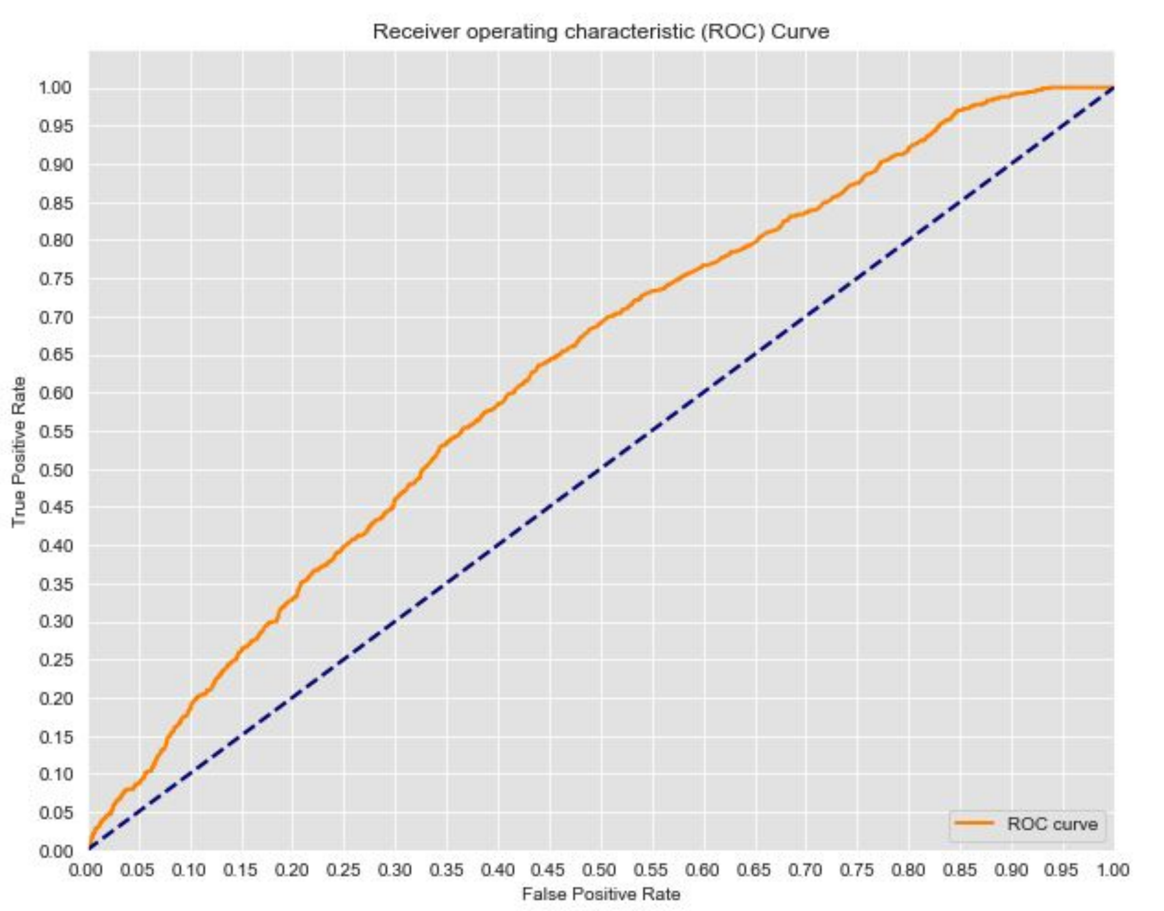
Of the age groups of subjects pulled over, 25-35 were the most pulled over. However, only 8.4% of those stops resulted in an arrest, compared to the 9.3% arrest rate in the 35-45 age range.

In terms of gender divide, those perceived to be male were stopped nearly 4 times as often as those presenting female. Those presenting male were arrested 8% of the time while female presenting were arrested at a rate of 6.8%.

Modeling

A challenge with modeling this data is that our target variable, arrests, are extremely weighted to one side. Of the over 48,000 entries in the dataset, 92% reflect a case when a Terry Stop did not result in an arrest. With a disparity like this, the model, will just predict "no arrest" 100% of the time, because at worst it'll be right 92% of the time.





False positives

The graph to the left shows the ROC curve, which illustrates the "true positive" rate vs. the "false positive" rate. The closer the orange line is to the top left corner, the better. If the line is along the blue dotted line. This shows that this model is slightly better than a coin flip.

Final Attempts at Normalization

There were additional steps taken at normalizing data including SMOTE, which is an algorithm used to address drastic discrepancies in data such as this data set. Also Lasso and Ridge were used, which are other forms of logistic regression, however to little success.

Recommandations

1. While 90% accuracy sounds nice, it does not reflect the actual data and is actually just a reflection of the discrepancy in arrest v non-arrest data. Do not take that number as a sign of good predictions

2. There is a race/gender disparity in officer representation that came up in initial analysis and should be addressed.

Future Work

Extract more data from the initial dataset. It's always better to have more data than less data.

THANK
YOU