Midterm

Sarah, Maggie, Jayden, & Maggie

November 13, 2024

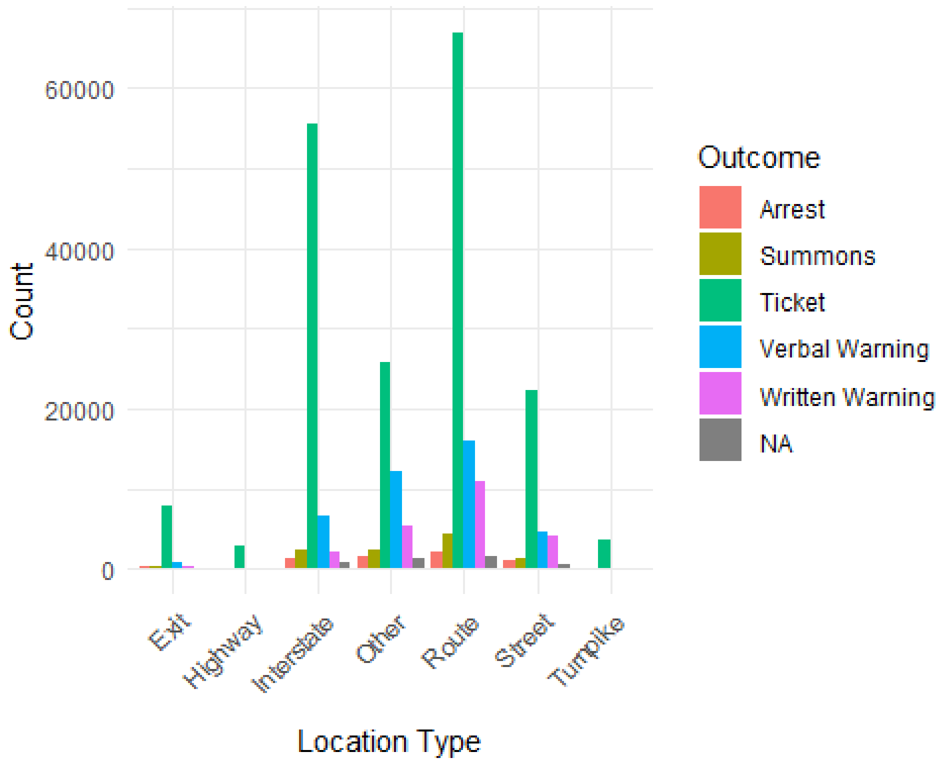
Part One:

Introduction: The data includes information about traffic stops from October 2013 to March 2015 in the state of Connecticut. There are 268,669 traffic stops and each stop is detailed with the driver's age, gender, race, location, reason for stop, along with quite a few more variables. A key objective of this project is to predict the outcome of a traffic stop based on various features in the dataset. To achieve this, we will conduct thorough data exploration and run logistical regression.

Exploring Data: Each of the nearly 270,000 traffic stops included additional information and demographics, including driver's age, race, and gender, where the stop occurred, the outcome of the stop, and whether or not a search was conducted. We focused most of our efforts on exploring four predictor variables: age, race, gender, and stop location. We created a bar graph for each variable against stop outcome or whether or not a search was performed. For the age variable, we cut the driver age into 4 different buckets and defined age ranges

- Young: Drivers aged 15-24.
- Middle: Drivers aged 25-40.
- Older: Drivers aged 41-60.
- Senior: Drivers 61 and older.

The chosen age buckets were based on life stages and their potential influence on driving behavior and interaction with law enforcement. For instance, drivers in the 15-24 age range are often younger, less experienced, and may have a higher propensity for risky driving behaviors. The 25-40 range represents a broader, more stable age group that encompasses many working adults. The 41-60 range often includes drivers with significant driving experience, potentially leading to fewer traffic infractions. The 61-and-older group, while experienced, may face different challenges or perceptions, such as age-related driving limitations or biases There are some potential downsides to age categorization. While it provides simplicity, some of the granularity can get lost by obscuring age-specific trends within the groups. For example, someone who is 41 might act completely different than a 60 year old, but they are both classified under "Older". Upon performing a Chi Square test on the new age_group category and stop_outcome, we saw a p-value of < 2.2e-16, indicating a statistically significant association between the variables.



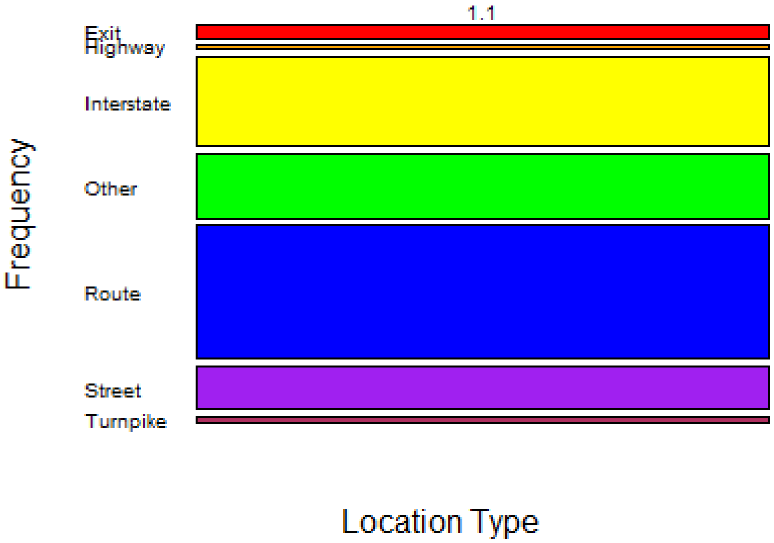


Figure 1 and 2: Comparison of stop outcome among the seven location types. Each location type more or less follows the same distribution of stop outcomes, with tickets being the most common for each. We chose this predictor because different types of streets may be patrolled more than others and for different reasons. For instant streets like highways and the interstate have a faster speed limits and a larger volume of cars driving on them at a time than neighborhood streets. Therefore, there may be a greater motive to patrol highways because they are busier than neighborhood streets. Police may focus more on speeding violations due to the faster speed limit or reckless driving that could lead to a potential fatal accident. In neighborhoods, they may focus more on safety-related stops in order to protect the residents of that area.

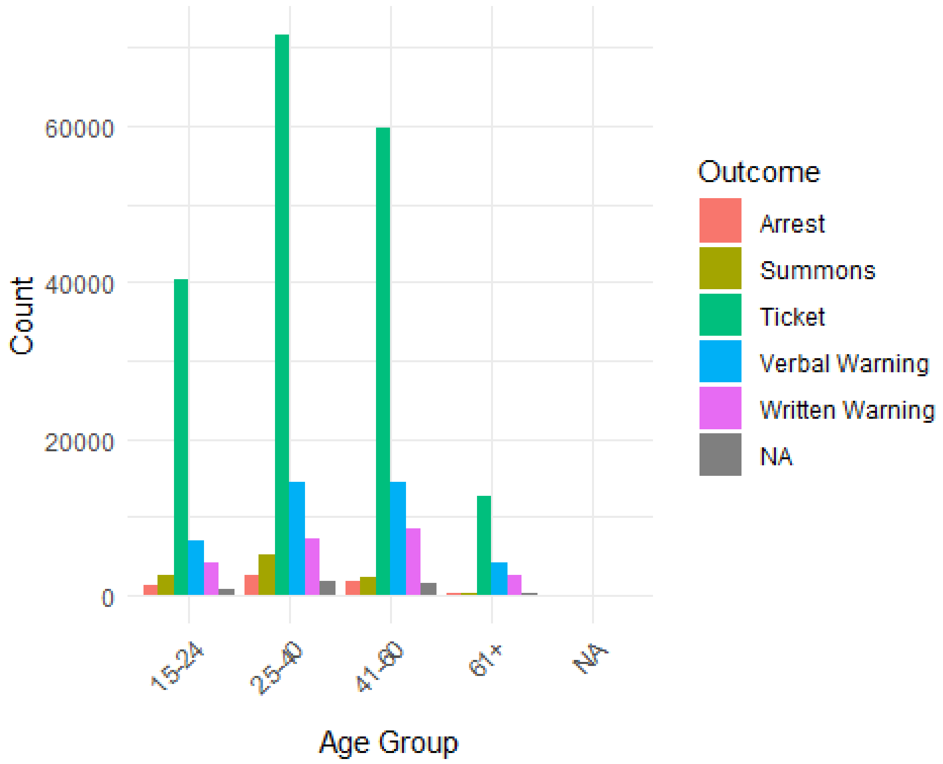


Figure 3: Comparison of stop outcome counts among different age groups. Each age group more or less follows the same distribution of stop outcomes, with tickets being the most common for each. There were virtually no cases of arrests or summons in the 61+ age group and way fewer cases of summons between the 25-40 and the 41-60 age groups. Verbal warnings was the second most stop outcome and written warnings were the third in all age groups.

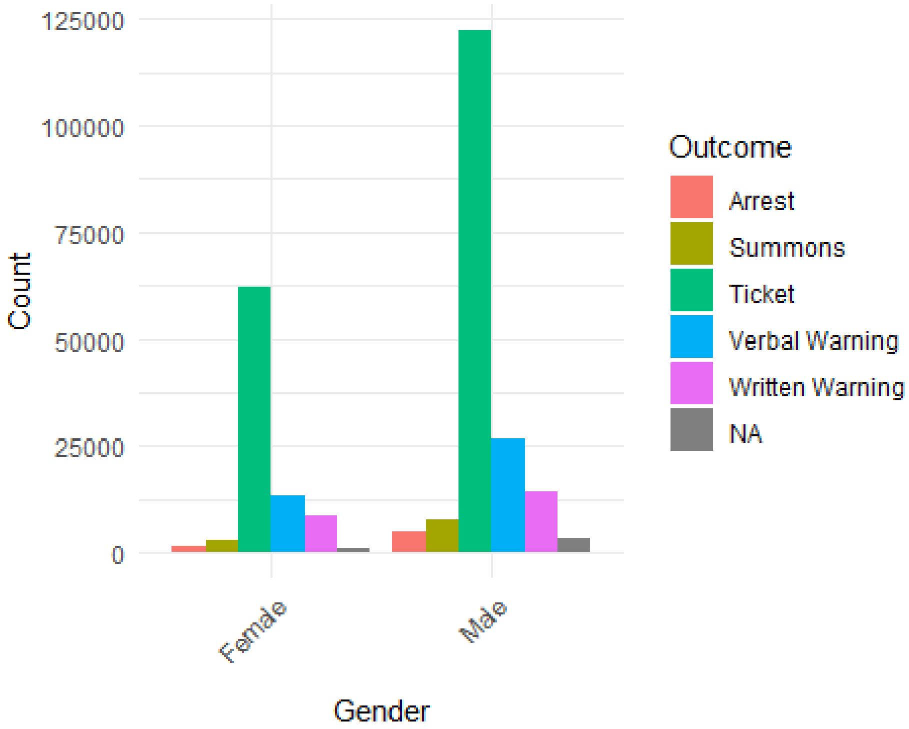


Figure 4: Comparison of stop outcome count among gender. A significant more number of males were pulled over than females, but both genders follow the same distribution of stop outcomes. Receiving a ticket was the most common for each, verbal warning being the second most common outcome, and written warnings being third. Being arrested was the least common outcome for both genders.

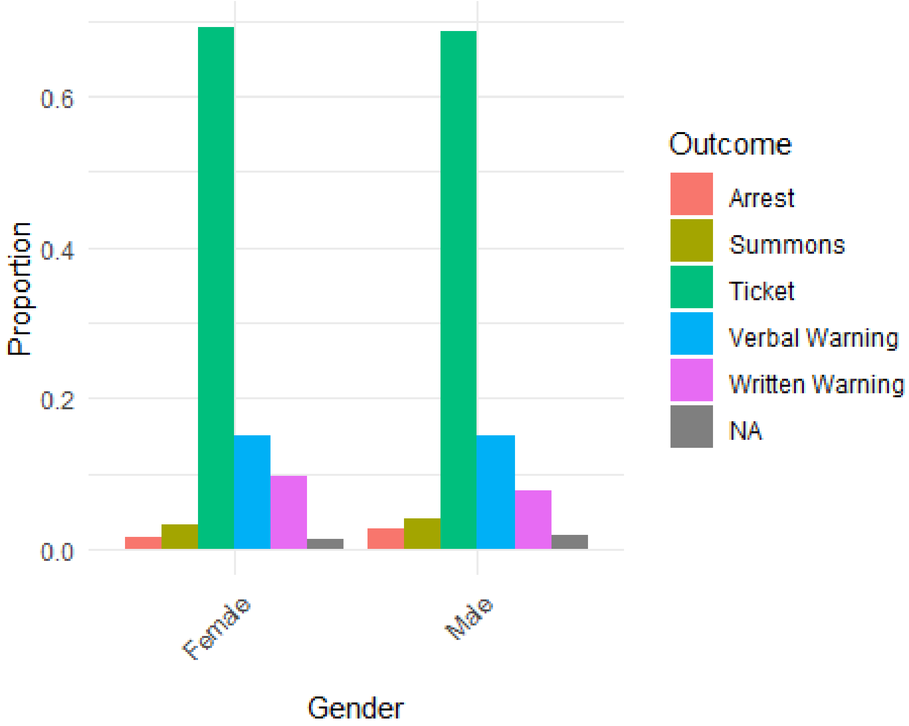


Figure 5: Comparison of stop outcome proportions among gender. Both males and females follow the same proportion distribution of stop outcomes, with tickets being the most common for each, verbal warning being the second most common outcome, and written warnings being third. The proportion of females who received a ticket looks almost identical to the proportion of males who also received a ticket. The proportion of arrests was the least common outcome for both genders.

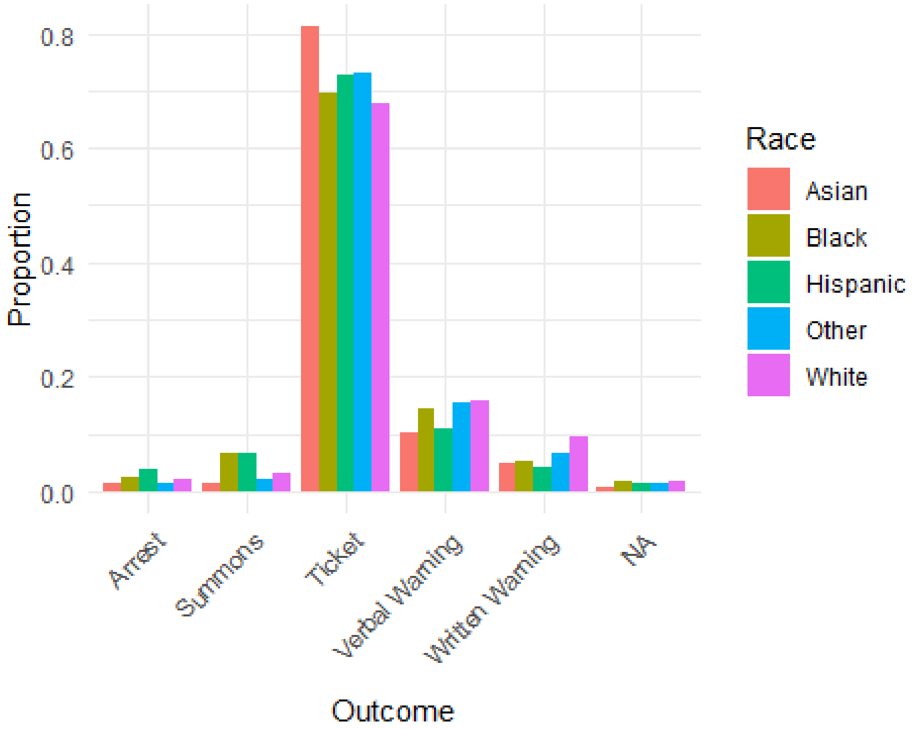


Figure 6: Comparison of stop outcome proportions among the different races. Each race follows similar distributions. Arrest and Summons have similar distributions and verbal warning and written warning follow similar distributions. Ticket is different as Asian individuals have the highest proportions of receiving a ticket among all races. For the two different warnings, White individuals have the highest proportion of receiving some type of warning. For Arrest and Summons, Hispanic individuals proportionally received the most of this type of outcome.

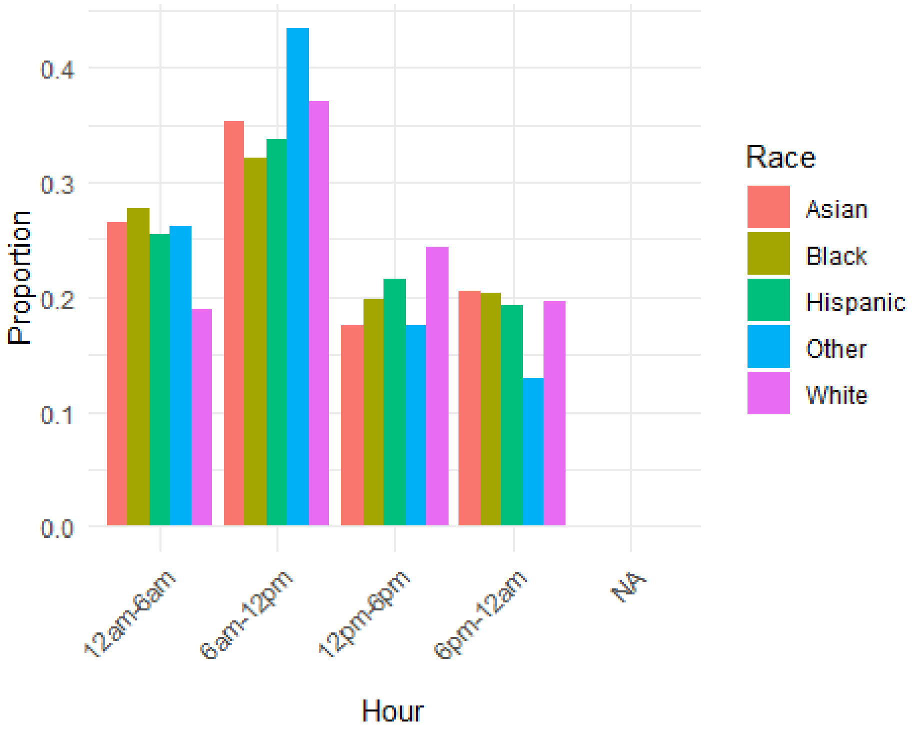


Figure 7: Comparison of the proportion of stops at different times of day across races. During 12pm-6pm, White individuals were pulled over proportionally more. During 6pm-12am, proportionally each race is pulled over about the same with the Other category being pulled over less often. During 6am-12pm the Other category is pulled over the most proportionally. During 12am-6am, White individuals are pulled over the least proportionally.

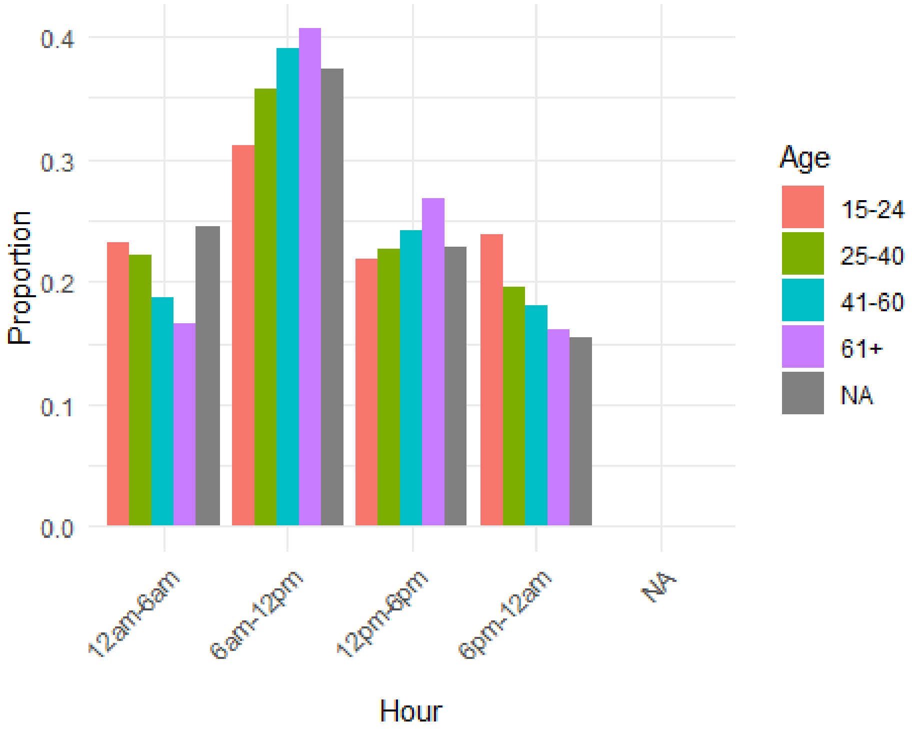


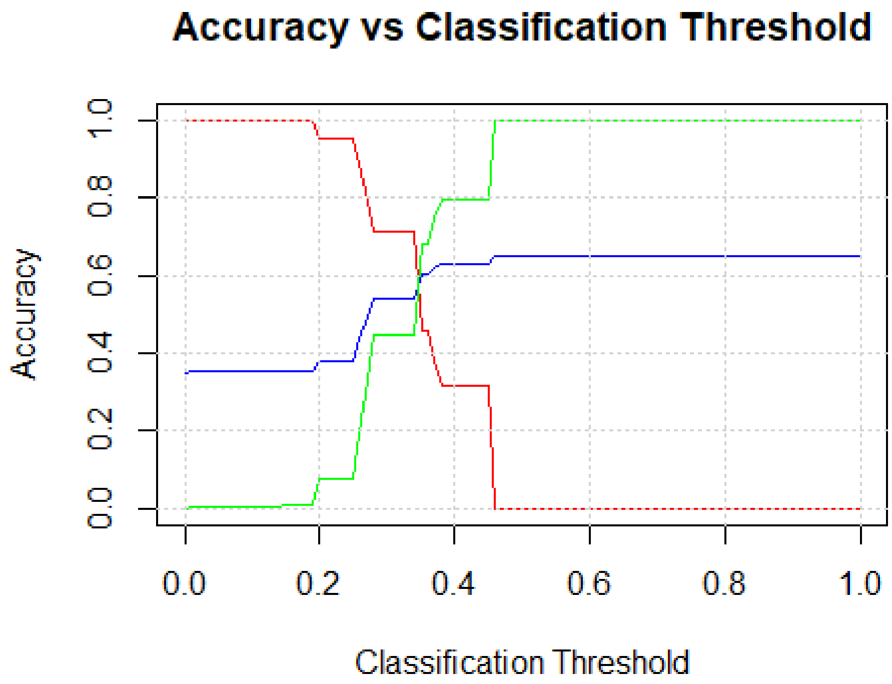
Figure 8: Comparison of the proportion of stops at different times of day across age groups. During 12pm-6pm, people 61+ were pulled over proportionally more. During 6pm-12am, as age increases, less people from that age group are pulled over proportionally. During 6am-12pm people who are 61+ are pulled over the most proportionally. During 12am-6am and 6pm-12am, the age group 15-24 are pulled over the most proportionally. Overall, we see that during the day from 6am-6pm as age increases, more people from older age groups are stopped. From 6pm-6am as age decreases, more people from younger age groups are stopped.

Part Two:

The Logistic Model For the model, we chose to do a generalized linear model. A GLM, such as logistic regression, is chosen for this type of analysis because it is appropriate for binary outcomes, interpretable, allows for statistical inference, and effectively handles various covariates. Given the nature of this analysis, we are predicting whether contraband was found (yes/no). These characteristics make it ideal for examining predictive relationships and drawing meaningful conclusions from traffic stop data, especially when investigating potential biases in search outcomes.We first started by filtering our data to include traffic stops where a search was conducted and filtering out rows with NAs. After splitting our data, we built a model with variables race, hour category(12am-6am, 6am-12pm, 12pm-6pm, 6pm-12am), gender, age, and location. We found that the hour category and age were significant. Overall the accuracy of the model is around 56%. Our sensitivity came out to be 0.405 and our specificity is 0.732.

Model Selection: For our model selection procedures we chose forward and backward selection. We chose forward and backward selection because they are simple and easy to understand, and efficient when picking only the most important predictors for our model. These methods helped us remove variables based on how much they contribute to the model, allowing us to exclude unnecessary factors. Compared to more complex techniques, like LASSO which can be harder to interpret, forward and backward selection gave us clear results which we used when choosing our variables. Both of these resulted in only age being significant for our model. With this new model, our accuracy increased by 0.04 (new accuracy = 0.603). Our sensitivity is 0.456 and our specificity is 0.681; sensitivity increased while specificity decreased.

Optimizing the Threshold for Accuracy: As the threshold varies from 0 to 1 we found that the optimal threshold for overall accuracy was 0.46, giving an overall accuracy of 65%. However, when we considered sensitivity and specificity, we found that a lower threshold would be more accurate. Given the intersection of the graph below, we went with a threshold of .36 instead of the recommended .46. Having a lower threshold will account for a class imbalance that we may have missed due to the high threshold.



Results Summary: Our model doesn't appear to show discrimination as race is not significant in predicting whether contraband was found. A caveat of our results is that overall the majority of the data was of White individuals and this may have affected the results. Some other types of data that would be helpful to have would be color of car, race of police officer, gender of police officer, number of people in the car, number of times they have received a ticket or been arrested, and the weather (if it is raining or snowing for example).