Analysis of Traffic Stops in Connecticut

Maddy Rilling, Eva Peters, Sophie Pope, and Julia Haas

Section 1: Introduction

This dataset contains information about traffic stops conducted by the Connecticut State Police Department. This dataset ranges from October 2013 to March 2015. This data was used to develop a model to predict the outcome of a traffic stop and whether a search was conducted. We first transformed the dataset for logistic regression modeling to include fewer extreme values and to make it normal by looking at the distributions of predictor variables and identifying any skewness. We did this by analyzing relationships between different predictor variables and looking at relationships between predictor variables and stop outcomes. We used different graphs such as box plots, histograms, KDE plots, and stacked bar plots to analyze these relationships. This helped us analyze what factors influence the likelihood of various outcomes of traffic stops, which can range from a verbal warning to an arrest. A few of the predictor variables we used were search type, contraband found, and when the stop occurred (month and time). For part two, we will create logistic models to predict whether contraband was found after a search by using race and other predictors. From there, we will select a model, optimize the threshold for accuracy, summarize our results, and test the accuracy of our model on a test dataset.

Section 2: Exploring and Transforming the Data

Description of the Data

This dataset contains many variables characterizing traffic stops, in total nearly 270,000 rows, each of which represents a single stop. Each row includes basic information pertaining to the stop like the date, time, county, officer, and length of the stop. It also includes demographic information about each driver, such as gender, age, and race. Along with these predictor variables, each row also includes attributes pertaining to the outcome of each stop, such as the type of violation, if a search was conducted, type of search, if a contraband was found, the stop outcome, and if the driver was arrested.

Cleaning the data:

To start, we removed duplicate “raw” columns of the data set that had a cleaned uniform version. These columns were not necessary and were removed. In dealing with NA values in the data set, we first changed any NA values in the column search_type to be “no search”, as these were stops where a search was not conducted. The majority of other NA values came from the columns, stop outcome and whether an arrest was made. We decided to drop these rows because knowing the stop outcome is crucial for our model. We can justify the dropping of these rows because they make up 1.67% of the total rows in the data set,

which is a small fraction of the data. 340 rows remained with with NA values in either driver_age, stop_time, or county_name. These rows were dropped because we felt that imputation techniques could create skewed patterns, and because these rows make up 0.129% of the data, dropping was a reasonable solution.

We also created two new columns that grouped the time of the stop into intervals and the date of the stop into month categories for ease in analysis and visualization.

Transformations:

While exploring the quantitative variables, we found that the driver age variable was significantly right skewed with many high outliers. To fix the distribution to match a more normal distribution, we took the natural log of the age values. The other quantitative variable in the dataset was stop time. However, we did not believe a transformation was necessary as the data was not very skewed and also had no outliers.

Relationships between predictor variables and stop outcome:

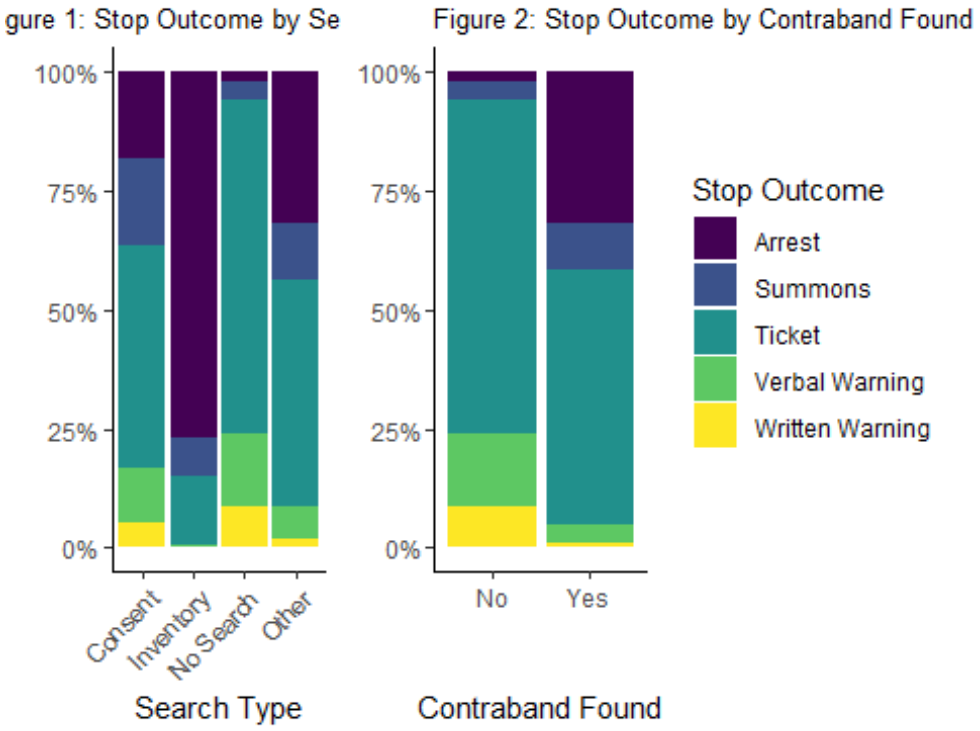
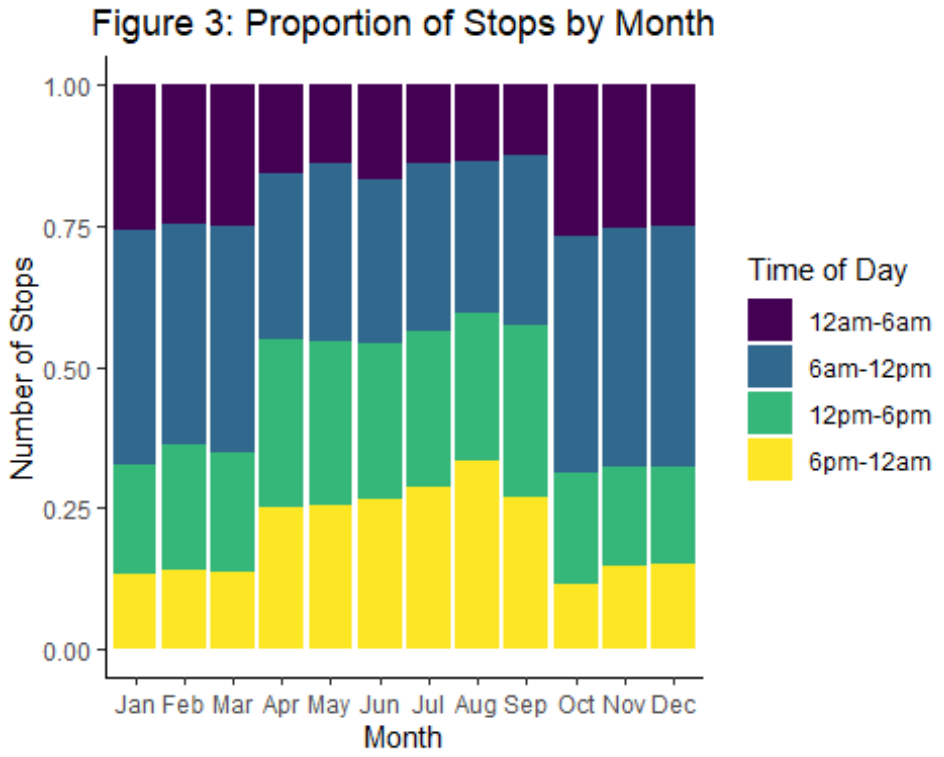


Figure 1 explores the relationship between the type of search that was conducted and the outcome of the stop. This figure shows that 77% of inventory searches resulted in an arrest compared to other search types where only a small proportion resulted in arrests. On the other hand, we can see that 70% of those who were not searched resulted in a ticket. This shows that the predictor variable, search type, has an effect on the response variable, because different search type's result in different stop outcomes.

Figure 2 highlights the impact of contraband discovery on stop outcomes. When looking at the figure, we can see that 32% of stops where contraband was found resulted in an arrest, whereas only 2% of stops where contraband was not found resulted in an arrest. Conversely, 24% of stops where contraband was not found resulted in a written or verbal warning, whereas only 5% of stops where contraband was found resulted in a written or verbal warning.

We also explored other relationships between predictor variables and different stop outcomes, however, these were the relationships that behaved most differently between outcomes.

Relationships between predictor variables:



Although now visually shown in **Figure 3,** the most stops occurred in the month of March, followed by October and then November. Figure 3 does visualize that the proportion of stops based on time of day changes from month to month. Stops increase in the 12pm-6pm and 6pm-12am intervals and decrease for the intervals 12am-6am and 6am-12pm intervals for the months April through September. This seems to be the most notable trend in difference in proportions. Overall there seems to be a common trend in stops for colder months October through March compared to warmer months April through September. The months October through March have a higher average stop count, and a difference in stop occurance throughout the day.

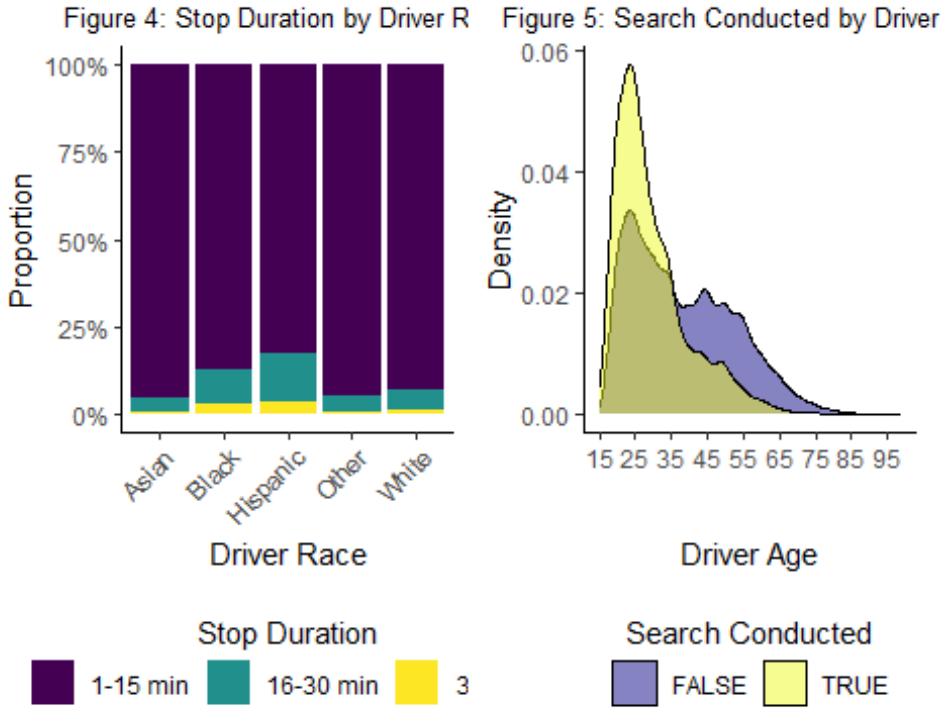


Figure 4 shows the relationship between a driver's race and the duration of the stop. Looking at the distributions, notice how Hispanic people, followed by Black people, and then followed by White, Asian and other races were generally stopped for a longer amount of time. Another important thing to note is that there were significantly more white people in the data compared to the other races, which makes this difference even more prominent.

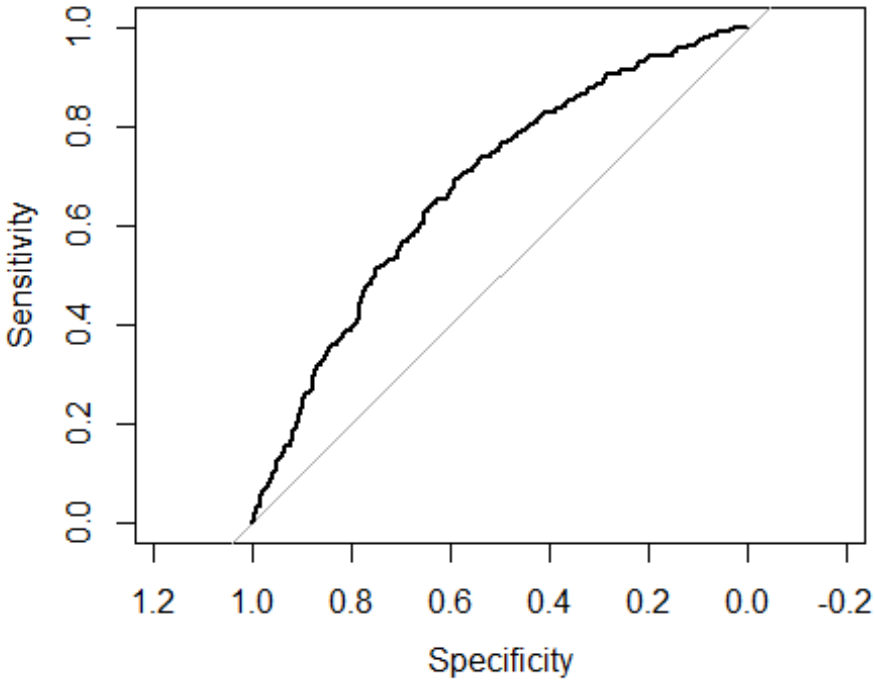
Figure 5 shows the relationship between a driver's age and if a search was conducted or not. Looking at the curves, notice that generally younger people, aged 15 to 36 were searched more compared to older people, aged 37 to 99.

Section 3: The Logistic Model

```
## Contingency Table

## Table 1: Confusion Matrix for Contraband Prediction

##         predy.opt
##          Pred=0 Pred=1
##   Obs=0    519    105
##   Obs=1    247    139
```



The first thing we did to make a model to predict contraband being found was to look only at the officer stops where they decided to search the vehicle. From there, We split the data into a training set (75%) to make our model and a testing set (25%) to test it. We used driver race, driver age (logged), driver gender, stop duration, search type, time of day, county, season, and violation as predictors in our baseline model. We did not pick use stop outcome or is arrested because those variables are not known when the officer decides to search a vehicle. Whereas even during the stop duration, the officer would know what time he is on track before deciding to search a vehicle. We used a threshold of .5 for binary classification on whether contraband would be found. If the model predicted the odds of contraband being found to be over .5, it would predict contraband being found; otherwise, it would not predict that nothing was found.

After creating the model, we examined the accuracy, specificity, and sensitivity of the model on the test set. We found an accuracy of 0.651, a sensitivity of 0.36, and a specificity of 0.832. We also made a contingency table for our results, as seen in Table 1. This model works okay with accuracy and specificity; however, the sensitivity is really low, so it is bad at predicting true positives (When contraband is found). The model also gives every variable a non-zero coefficient, so it uses every variable and does not exclude any that are ineffective in predicting, which is not ideal. Also, a threshold of .5 is not good because the model has trouble finding true positives. It implies that the threshold is too high. The probability threshold of it being true, leading to it being labeled true, should be lowered.

Section 4: Model Selection

```
## 35 x 1 sparse Matrix of class "dgCMatrix"
##                                            s1
## (Intercept)                          1.87665717
## (Intercept)                          .
## driver_raceBlack                    -0.12049884
## driver_raceHispanic                  .
## driver_raceOther                    -0.81060595
## driver_raceWhite                     0.23686989
## driver_age_new                      -1.02273794
## driver_genderM                       0.01683718
## stop_duration16-30 min               .
## stop_duration30+ min                 0.53554637
## search_typeInventory                 .
## search_typeOther                     0.88168109
## time_category12pm-6pm                0.14103962
## time_category6am-12pm                0.04280215
## time_category6pm-12am                0.01246869
## county_nameHartford County          -0.12879380
## county_nameLitchfield County         0.13456470
## county_nameMiddlesex County          0.18426396
## county_nameNew Haven County          .
## county_nameNew London County         .
## county_nameTolland County            0.07670744
## county_nameWindham County            0.36175077
## seasonSpring                         .
## seasonSummer                         .
## seasonWinter                         .
## violationEquipment violation         .
## violationLicense violation          -0.06665807
## violationLights violation            .
## violationMoving violation           -0.20567536
## violationOther violation             0.33473998
## violationRegistration violation     -0.02032292
## violationSafe movement violation     0.41180170
## violationSeat belt violation         .
## violationSpeeding violation          0.02093192
## violationStop sign/light             .
```

We chose to use Lasso regression for model selection because Lasso regression applies regularization and feature selection. Different from other model selection techniques, Lasso applies a penalty that encourages the model to shrink certain coefficients to zero. Therefore, Lasso regression removes less important predictors from the model. This makes the model simpler, more interpretable, and reduces overfitting because only relevant features are retained. Overall, Lasso regression balances between predictive accuracy and interpretability, so we decided to use this as our model selection procedure.

In the model chosen by Lasso regression, only one predictor was removed from the baseline model. This predictor was which season the stop occurred in, and by removing it,

our model becomes slightly simpler to interpret. Comparing the results of our baseline model and what we deemed to be our best model, we see that accuracy increases slightly. In our baseline model, we had an accuracy of 65.15% and for the best model we had an accuracy of 65.54%. Our sensitivity and specificity also increased slightly from our baseline model to our best model. For our baseline model, sensitivity was 36.01% and specificity was 83.17%. In our best model, sensitivity was 36.27% and specificity was 83.65%. Although these changes between the baseline model and our best model are small, they reflect gains in all metrics, without sacrificing one metric for another. This indicates that the model selected through Lasso regression provides a more balanced and effective model.

Looking at both models, we see that sensitivity is fairly low and specificity is fairly high. This is likely due to the fact that there are many more times where contraband is not found compared to contraband being found. Since there are more cases where contraband is not found, the model may learn to classify most of the cases as negative because it will more likely be correct that way. Specificity measures the true negative rate, so in a dataset where there are many more negative data points, it is much easier for the model to correctly identify negative cases. Therefore, it makes sense that our specificity is higher and sensitivity is lower for both models, because sensitivity measures the true positive rate.

Section 5: Optimizing the Threshold for Accuracy

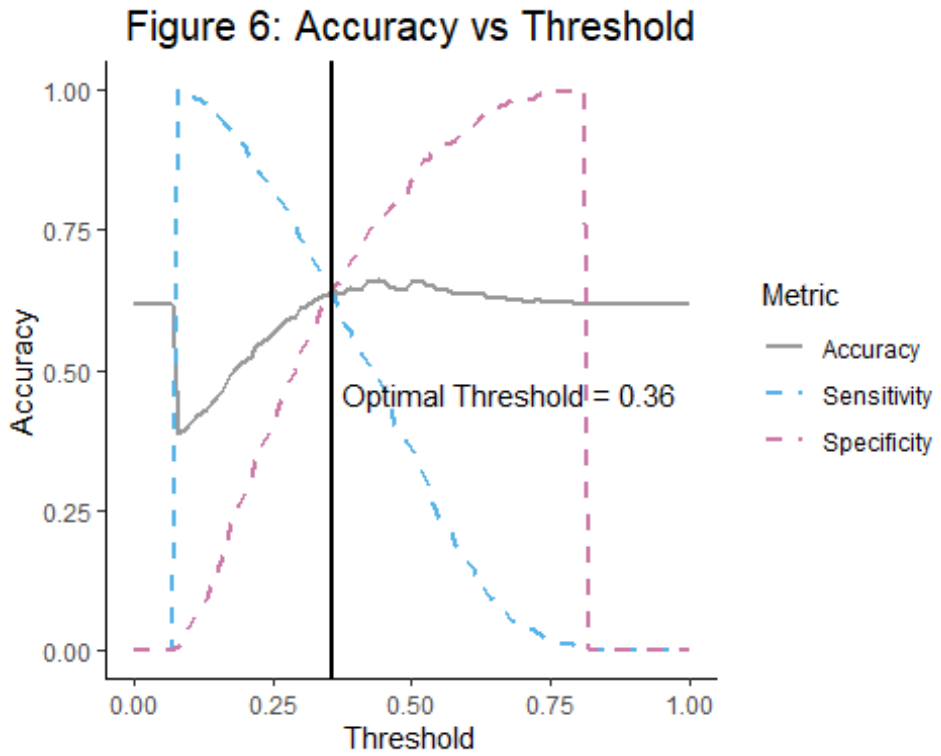


Figure 6 shows the relationship between accuracies and threshold for our model, showing that the threshold that maximizes accuracy in predicting the outcome of a contraband

being found correctly is 0.36 with a corresponding accuracy of 0.64. Notice the graph starts with a high accuracy of about 0.66 at a threshold of 0. This is because here it is predicting all cases as false, which is about 66% of the cases in our test data. Accuracy then drops to about 0.34 once the threshold increases past 0, because it is now predicting all cases as true, which is about 34% of the cases in our data. Accuracy then slowly rises to our optimal threshold of 0.36, and then eventually plateaus to return to an accuracy of 0.66, predicting all cases as false again.

An important thing to note is that the optimal threshold for accuracy is not quite at the highest accuracy point on the graph. This is due how we calculated the optimal threshold, by maximizing the sum of sensitivity and specificity. If we were to increase our optimal threshold to the highest accuracy, we would then sacrifice sensitivity and specificity. As threshold increases from 0.36 to 0.44 (where our highest accuracy point is), specificity increases, but sensitivity decreases, which is not ideal since our goal is to be most accurate in predicting correct true cases of contraband found.

Section 6: Results Summary

Through data exploration, transformation, and model selection, we have found our best logistic model for predicting if contraband was found after a search was conducted during a traffic stop. This model includes data on driver race, driver age, driver gender, stop duration, search type, time category, county, and violation. Our final model removes season completely when compared to the baseline by forcing all coefficients to be 0. Certain violations and stop durations, counties, search types, and driver races were also forced to have 0 coefficients. Our model also changes the coefficients of each predictor, and overall our model found through lasso regression performs slightly better than the chosen baseline in terms of overall accuracy, specificity, and sensitivity. Our best threshold that maximizes sensitivity and specificity is .36, and when performed on the testing data the accuracy was 65.54%, sensitivity was 36.27% and specificity was 83.65%, which were all better than our baseline model.

Our final model suggests that race does influence the likelihood of contraband being found during a traffic stop. Our baseline race classification was ‘Asian’, and ‘Hispanic’ race was the only classification that wasn’t seen to make a notable difference in our model. ‘Black’, ‘White’, and ‘Other’ all had changing coefficients from the baseline.

Our final model also suggests that if a search took place for 30 more more minutes there is a higher chance of contraband being found. Our baseline for this predictor was 1-15 minutes, and the range 15-30 was not found to be different enough in from the baseline. Because of this, the coefficient was driven to 0.

The county a stop took place in also has an impact on the likelihood of contraband being found as an outcome of a traffic stop. Our baseline county in the model is Fairfield County. New Haven and New London counties were not found be significantly different enough from our baseline to warrant coefficients in our model, but all other counties were. Interestingly enough Hartford County presents a negative coefficient in our model,

suggesting that if the stop occurs in this county, the likelihood of contraband being found is lower than all other counties. Middlesex County has the highest coefficient compared to other counties in our model, suggesting that if the stop is occurring in this county, the probability of contraband being found is higher compared to others.

Considering violation types, our baseline violation in our model is a 'cell phone violation'. Violations that were driven to 0 in our model include 'equipment violation', 'lights violation', 'seat belt violation', and 'stop sign violation'. Violations 'license', 'moving', and 'registration' all had small negative coefficients, indicating that if a traffic stop occurred for any of these reasons, the likelihood of a contraband being found if a search was conducted is smaller compared to baseline violations. 'Other' and 'safe movement' violations both had high positive coefficients. If a stop and search occurred as a result of these violations, the chance of contraband being found is higher compared to baseline violations.

Having more detailed descriptions on what 'other' entails in multiple predictor variables could be useful in increasing accuracy in our model. Features like race, stop violation, and search type all have 'other' describing data that doesn't fit into their other groupings. Understanding these variables may be helpful and could increase accuracy.

Section 7: Accuracy Test

Note: In the final predictions of CT_prediction_set, 82 predictions come out as NA. This is due to missing values in the prediction data set, specifically in row search_type. To maintain the order of the rows in the data set we chose not to remove these predictions. Because we did not use imputing methods for NA values in search_type to build the model, we felt it best not to impute values in CT_prediction_set. **