Mini-Project Part II – by Olamide Oladeji & Abuzar Royesh

Our project focuses on predictive modeling and inference using data of all police arrests in New Orleans, Louisina.

Part I:

Regression

In part I of the project, our best performing model for the regression task was an elastic-net model fitted on all covariates and interactions. Our RMSE on the validation dataset had been **12.91732**, and we had argued that it served as an unbiased estimator of the RMSE on the test dataset. We used this fitted model to generate predictions for the held-out test dataset. We conducted the same clean-up we had done for the training dataset1. The resulting RMSE on the test dataset was **12.76803**, slightly better than our estimate from the validation data. This was not surprising given that we had a large number of observations in each of the training, validation, and test datasets, and that our validation data essentially served as a test dataset. The slight difference, we believe, is down to uncontrollable variations in the data.

Classification

For the classification task, we had deemed the base logistic regression with no interactions as the best model, using accuracy, sensitivity, and specificity as our criteria. Our best estimate for these metrics had come from the validation dataset, yielding accuracy of **0.9242**, sensitivity of **0.9778**, and specificity of **0.6318**. We had argued that these were unbiased estimates for the performance on the test dataset. When running the model on test dataset, we found our argument to be true: our model had an accuracy of **0.9219**, sensitivity of **0.9761**, and specificity of **0.6254**. The numbers were marginally lower for the test dataset, but we believe this is all due to random noise in the dataset.

Part II:

II.a.

When we examined the significance of the coefficients in the regression outputs (as obtained using logistic regression on the training set), we identified statistical significance on a number of covariates. The logistic regression model had **16** covariates (of which **15** were categorical), yielding **49** covariates + intercept in total after dummy-encoding. Of these, the coefficients for **29** had significance while there were **20** non-significant coefficients. However, it should be noted that these were each level from the original **16** categorical covariates. If we analyze these significance results with respect to the original **16** non-dummified covariates, we find that **ALL 16** covariates were significant in the regressed model.

We recognize that due to the fact that these significances were obtained after multiple hypothesis testing on **49** coefficients, and the fact that we favorably selected the original 16 covariates, the significance are likely over-estimated. That said, looking at the significance of the covariate levels from the current analysis and our subjective understanding of the covariates, we find the significance in certain covariates such as *contrabound_found*, *frisk_performed* credible. Considering the post-selection bias and other issues we raised above, we recommend that the significance results from our model be compared to results from fitting the model on an unbiased data, as well as from other entirely different modeling projects, before concluding these coefficients are really significant.

For our trained model, after calculating the odds from the odd logs, we find the most important factors in predicting whether an arrest was made were whether the person was searched and whether contraband was found. The odds of arrest for people who were searched were 6.03 times (603%) higher than for those who were not. Similarly, the fact that a contraband was found made it 7.86 times more likely for a person to be arrested. With respect to race, we find that Blacks were 2.58 times more likely to be arrested compared to Asian/Pacific Islanders (significant at the 0.001 level). This

[1] One of the clean-up steps on the training dataset had been to drop all observations containing missing values. For cleaning the test dataset, initially we decided not to drop those observations, but realized that not dropping missing values created additional covariates and changed the scale when we dummified and rescaled our data to prepare it for the elastic-net model. Therefore, we decided to also drop the missing values from the test dataset.

is while Whites and Hispanics were 1.83 and 1.81 times more likely to be arrested, respectively (also significant at the 0.001 level). As for the time of the day, compared to the afternoons, the odds of being arrested was 13% higher for the night-time and 14% lower for the morning (both significant at the 0.001 level). Evenings were not statistically different than afternoons. These results were all in line with what we had expected. The only surprising result was that being a man was associated with lower likelihood of getting arrested (0.79 compared to 1 for females). The full regression output is attached as Appendix I.

II.b

The matrix to the right summarizes the significance of the variables when running the models on training and test datasets. We see from the comparison matrix above that 25 coefficients were significant in both the fit on the training data as well as the fit on the test data. The main differences between the two models were that one variable (*officer_assignment_2nd_District*) that was not significant in the train model was significant for the test model, and four variables (*district4*, *officer_assignment_4th_District*, *reason_for_stop_suspect_vehicle*, and *month10*) that were significant in the train model were not significant for the test model.

	Significance on Test				
Significance on Train		*	**	***	p > 0.05
	*	0	2	0	2
	**	0	0	1	1
	***	3	4	15	1
	p > 0.05	1	0	0	19

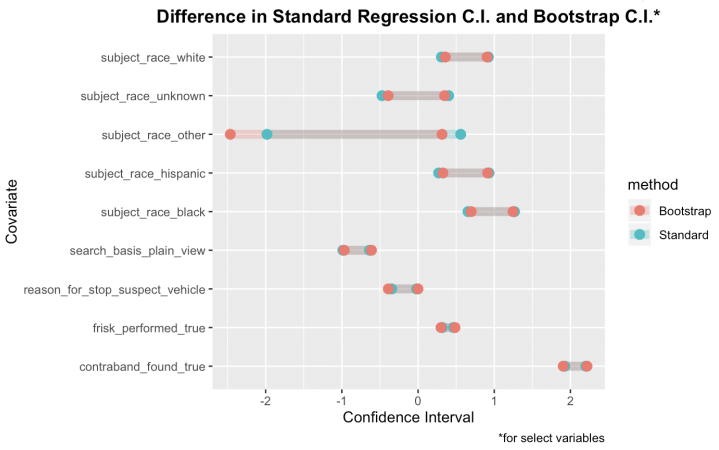
We believe the differences in coefficients significance between the training data and test data may have been due to a number of factors.

1. First, in selecting the covariates for our original training model fitting, our feature selection process had favorably biased to those covariates, given that we examined model performance and p-value significance before deciding to include all of them.
2. The test data is an unbiased data source hence the disparity. The test data also has a different (smaller) sample size n to the training test, which may affect sampling distribution and their associated hypothesis testing.
3. Finally, the test procedure for significance involved multiple hypothesis testing on 49 covariates at 0.05 significance level, raising the very high possibility that we have false positive significance when testing on both the training data and test data fits. The non-overlap of false positive significance may be what is reflected in the above table.

Finally, 15 covariates were significant for both models at the 0.001 level and 19 covariates were not significant at the 0.05 level in either of the models.

II.c. Confidence Intervals of Bootstrapped coefficients vs. Fitted Coefficients

We used the boot::boot function to run bootstrap on our data with 1,000 replications[2] and number of indices equal to the number of observations in the dataset (163,7 87). After running the bootstrap, we explored several types of confidence interval estimation methods, including normal, quantile, basic, and bias corrected (BCA)[3].



[2] We were unable to increase the number of replications because of computational limitations.

[3] We were unable to generate results for this as we received error messages indicating insufficient vector memory (despite setting the maximum RStudio Virtual Memory = 100GB, and using a MacPro with 16GB RAM, i7). We assume this is due to the fact that R boot library's underlying approach involves generating several large 2D 'importance arrays' of (n of observations x no of repetitions) each. Note that we had >200,000 observations.

We store the confidence interval for all 49 coefficients, and using the 3 methods above in a dataframe attached in Appendix II. We compare these with the confidence intervals for the 49 coefficients obtained from our logistic regression fit on the data. For the most part, the lower and/or upper bounds of the confidence intervals change slightly between the two methods. The figure on the previous page shows the confidence intervals for the nine covariates that had the greatest discrepancy. Overall, we find that bootstrapping changed the confidence intervals most significantly for race. We believe that these discrepancies are due to the fact that the standard regression makes assumptions about the distribution of the population, while bootstrapping makes no such assumptions. In the case of race, for instance, it is possible that the coefficient has a non-normal distribution that is captured more accurately through the bootstrap methods. However, we cannot be sure about the unbiasedness of our bootstrap method since we were unable to run bias corrected bootstrap due to the matrix memory issues mentioned earlier.

II.D. Model Comparisons with and without All Covariates.

Our feature selection consisted of first starting with the full features (left after cleaning and dropping some perfectly collinear variables) and afterwards examining ANOVA tests and the coefficient p-values for non-significant variables to drop. Since our logistic regression's predictive error did not decrease from dropping these non-significant variables or adding interaction terms, we kept all post-cleaning covariates. As a result, we are unable to compare with a model that does not include all post-cleaning covariates.

	Gender-Interaction Model				
Final Model		*	**	***	p > 0.05
	*	0	1	1	2
	**	0	0	0	2
	***	3	2	14	4
	p > 0.05	1	1	1	17

In the feature selection process, we had considered gender interactions, but based on the predicted error, we had decided not to choose this model as our best. Out of curiosity, we then compared the significance of our best against a model that also included gender interactions (Figure to the left). We find that eight covariates that were significant to our best model, were not significant for the gender interaction model. These covariates were *district5*, *subject_sex_male*, *search_basis_plain_view* (originally significant at the 0.001 level), *month10* and *month12* (originally significant at the 0.01 level), and *reasons_for_stop_suspect_vehicle*, and *weekday_weekend* (originally significant at the 0.05 level). We thought of two possible reasons for this: 1) the interaction is an overfitted model that might introduce unnecessary multicollinearity in the data, 2) in reality these eight variables are not significant but because of favorable selection bias, our model has picked those covariates as significant in the final model.

II.E. Potential Problems with our Analysis

We analyze potential problems with our analysis with respect to three issues:

Multicollinearity: One evidence that we may have had some multicollinearity is the fact that, even though we had a very large training dataset (n>200,000), validation and test sets (n>80,000), the estimates and significance of some coefficients changed as we moved from the training set to the test set. This is perhaps due to multicollinearity, among other things, in which case the model cannot accurately calculate the magnitude of effect of the collinear variables on the outcome variable. Also, another evidence of multicollinearity is that while the p-value and thus the significance from our Pearson chi-squared test between the response binary variable *arrest_made* versus *month*, *reason_for_stop*, and *district* covariates showed significant relationships (in addition to the fact that removing it did not improve the predictive power) which informed us inputting all 16 covariates into the final training regression, these covariates had reduced significance in the final regression model on the unbiased test set. For *month,* we found that all but the *month_1* level were deemed non-significant in the unbiased test set fit. The *reason_for_stop* covariate was entirely non-significant in the test set despite it having showed significant relationship to the response in the feature selection stage. Many of the *district* levels in the test fit did not show significance and the levels which did, did not have as strong significance as the feature-selection testing made us to believe. In addition to the above, we explored literature for other ways to detect evidence of

multicollinearity in our fitted model but the other methods we found online such as the Variance Influence Factor (VIF) are apparently only suitable for continuous response and covariates.

Hypothesis testing: The original 16 covariates became expanded to 49 covariates due to almost all of them being categorical. Using a 0.05 alpha, we were almost certain (with a probability of >0.9) that we get at least one significance column by chance, even if all were truly not significant. Thus, it is very likely, with that many covariates, that we have a number of falsely significant coefficients. This, as we know, is one of the problems with multiple hypothesis testing. The **Bonferroni** correction corrects for this by dividing the test p-values by the number of covariates before comparing it with the significance level alpha = 0.05. When we applied the Bonferroni correction, we realized that four variables *district_2, district_4, reason_for_stop_suspect_vehicle, weekday_weekend* variables were now deemed non-significant under alpha = 0.05, despite being significant prior to correction. We also applied the **Benjamini-Hochberg (BH)** correction, and found only two formerly significant coefficients, *reason_for_stop_suspect_vehicle* and *district* changed to non-significant.

Post-Selection Inference: As described previously, our feature selection consisted of first starting with the full features (left after cleaning incomplete data, and dropping some perfectly collinear or subjectively determined as useless variables) and afterwards examining results of p-values of pair-wise Pearson chi square tests (for categorical response and categorical variables) and pairwise one-way ANOVA (for continuous response vs categorical covariates), and then examining the predictive error in the event a deemed non-significant variable was dropped to determine whether to drop it or not.

Since our logistic regression's predictive error did not worsen from dropping the one non-significant variable we obtained, we eventually kept all post-cleaning covariates. Interaction terms did not improve the prediction error and so were not added in the final model.

In carrying out the above, by examining the predictive error on the validation set, as well as looking at the p-values, we had favorably biased the model to those coefficients, even though we eventually kept all post-cleaning coefficients (removing *search_vehicle* the non-significant variable from the feature selection process did not reduce our predictive error so we kept it in the final model). This is because, for feature selection we had used p-values from hypothesis testing with alpha = 0.05 to check significance (in the pair-wise ANOVA and chi square tests) and the multiple hypothesis testing may mean that we have a reasonable chance at getting false positive significance in some covariates and thus including them in the model when they otherwise wouldn't be significant. Also, our favorable selection bias means that the eventual p-values obtained after fitting of these coefficients could be smaller than they really are, making us have the many significant variables we have in our final model.

II.F

Our findings for each of our models merely test for associations between the covariates, and we cannot interpret any of our findings from the previous sections as causal. On the most basic level, we cannot plausibly contend that demographic data have a causal effect on arrests by the police, without the presence of intermediate variables. There is some debate in the statistics and social science community whether demographic variables such as "race" and "gender" can be seen as causal variables, given that they cannot be "controlled" as in a Random Controlled experiments, since researchers cannot assign them to participants in RCTs[4] while others oppose this counterfactual view of causal inference. We decided to take a counterfactual view to causal inference and interpret only variables which we can "control" in an RCT as potential causal variables if they are significant. If causal inference is to be undertaken, two of the variables that could be candidates for further analysis on causality would be contraband_found and search_person, since those were most associated with arrests and from our subjective reasoning they seem so. To test for causality around these variables, we would propose a random controlled experiment in which we assign groups different values of these and examine the Average Treatment

[4] See: P, Holland, *Causation and Race. ETC Jan.* 2003 [Online] https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2003.tb01895.x and Q. Zhao, *Topics in Causal and High Dimensional Inference,* Stanford. 2017. https://web.stanford.edu/~hastie/THESES/Qingyuan_Zhao_thesis_augmented.pdf

Effects. Some covariates that might confound our ability to infer causal relationship in the case of *search_person* and *contraband_found* could be police racial bias, laws regarding search, contraband and arrest, and legal jurisdictions. For instance, if we analyze data from two states on frisk and arrest, we would need to account for the laws of the two states. Similarly, the level of bias on the part of the police can affect both conducting searches and arresting individuals.

PART III: Discussion

III.a.

In reality, giving the type of data we have and the covariates, we expect that decision makers / users of the model will be more inclined to use it for inference rather than prediction. However, given the issues we raised in Part **II. D-F** on potential pitfalls of conducting inference with our model as-is, we would be wary of such use. On the prediction side, our model performs well on overall accuracy and sensitivity and so this is a more confident aspect of the current model for us, although the specificity value obtained means the user has to pay attention to the likelihood of false-positives.

III.b.

A number of covariates in our model can change depending on several factors such as federal, state, or local policies, mandatory law enforcement training or procedural change over time. The demographics across districts may also change over time. If there are sudden changes to these, we would therefore recommend that the model be refit after a year. Without any drastic changes in these, a longer timeframe may be better, considering that the current model was fit to over 10 years of data and a very large amount of data points.

III.c.

For model users, we would focus on pitfalls around using this model for inference, since this will be their likely use-case.

1. For example, we would mention that our feature selection process favorably biased the coefficients to significance and that they take our model's p-values and significance with a grain of salt. To make the model more robust, they can fit the model on completely new, unbiased data and interpret the p-values and significance for those new data points instead.
2. We would also warn them that even for new data, to still be careful about interpreting significance because using the 49 covariates (after categorical dummification), the chance of having at least one false positive under significance level 0.05 purely by chance is very high. We would encourage them to consider applying methods such as Benjamin-Hochberg correction to their new analysis.
3. Finally, we suggest not to interpret causal relationships on significant variables without examining results from analysis by other researchers, running RCTs, controlling for confounding variables, and qualitative research.

III.D&E

Overall, for the data collection process, we would have changed the open-ended questions filled by police officers from which certain covariates were derived to categorical questions. These would lead to more data integrity and reduce the amount of useful but missing data. Also, we recommend that demographic information is extracted from government records (subject to privacy laws and ethics) instead of being filled by police officers, which can be influenced by police biases. The process of data collection should be streamlined in all the states and police districts, which would allow for easier data analysis. We would also appreciate additional covariates around state and local laws, the violation committed, level of education, descriptions of the incident, rationale for conducting searches, and other confounding variables.

If we were to analyze the dataset again, we would have wished to delve even deeper into the issue of multicollinearity, identify which sets of covariates were collinear, and think about how to improve the accuracy of our estimates. For inference, we would have used a hierarchical model for adding covariates, whereby we could see the impact of adding a variable on all the preceding variables step by step. We would also analyze the dataset for autocorrelation, especially when there is a time component to our analysis. Finally, we would have experimented with more models, especially those that are more computationally intensive.

Appendix

Appendix A

Variable	Train				Test			
	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)
(intercept)	1.33	0.2	6.73	0	1.04	0.35	2.98	0
time_of_dayevening	-0.03	0.03	-0.99	0.32	0.02	0.05	0.38	0.7
time_of_daymorning	-0.15	0.03	-4.87	0	-0.12	0.05	-2.19	0.03
time_of_daynight	0.12	0.03	4.32	0	0.15	0.05	3.06	0
district2	0.2	0.09	2.22	0.03	0.47	0.16	2.94	0
district3	-0.11	0.09	-1.35	0.18	-0.22	0.15	-1.46	0.14
district4	0.26	0.11	2.34	0.02	0.2	0.19	1.05	0.29
district5	-0.43	0.09	-4.56	0	-0.49	0.17	-2.96	0
district6	0.13	0.08	1.57	0.12	0.12	0.15	0.78	0.44
district7	-0.62	0.1	-6.22	0	-0.5	0.18	-2.83	0
district8	0.17	0.09	1.81	0.07	0.08	0.17	0.49	0.62
subject_age	-0.01	0	-9.33	< 2e-16	-0.01	0	-6.35	0
subject_raceblack	0.95	0.16	6.09	0	0.96	0.28	3.49	0
subject_racehispanic	0.59	0.17	3.5	0	0.67	0.3	2.25	0.02
subject_raceother	-0.57	0.64	-0.89	0.37	-0.3	0.94	-0.32	0.75
subject_raceunknown	-0.04	0.22	-0.16	0.87	0.36	0.37	0.98	0.33
subject_racewhite	0.6	0.16	3.84	0	0.65	0.28	2.34	0.02
subject_sexmale	-0.24	0.02	-10.88	< 2e-16	-0.22	0.04	-5.65	0
officer_assignment2nddistrict	-0.17	0.09	-1.76	0.08	-0.4	0.17	-2.4	0.02
officer_assignment3rd district	-0.05	0.09	-0.56	0.58	0.13	0.16	0.79	0.43
officer_assignment4th district	-0.43	0.12	-3.68	0	-0.29	0.2	-1.44	0.15
officer_assignment5th district	-0.01	0.1	-0.07	0.95	0.06	0.18	0.32	0.75
officer_assignment6th district	-0.59	0.09	-6.31	0	-0.81	0.17	-4.85	0
officer_assignment7th district	0.75	0.1	7.13	0	0.66	0.18	3.6	0
officer_assignment8th district	-0.18	0.1	-1.86	0.06	-0.12	0.18	-0.7	0.49
officer_assignmentother	-0.05	0.08	-0.61	0.54	0.11	0.14	0.75	0.45
officer_assignmenttraffic	-0.57	0.08	-7.53	0	-0.49	0.14	-3.58	0
typevehicular	-0.71	0.03	-26.71	< 2e-16	-0.7	0.05	-14.81	< 2e-16
contraband_foundnot searched	-3.82	0.11	-35.45	< 2e-16	-3.57	0.19	-18.99	< 2e-16
contraband_foundtrue	2.06	0.07	29.17	< 2e-16	2.19	0.12	17.99	< 2e-16
frisk_performedtrue	0.39	0.04	10.36	< 2e-16	0.34	0.07	5.1	0

search_persontrue	1.8	0.07	25.03	< 2e-16	1.62	0.13	12.9	< 2e-16
search_vehicletrue	-3.07	0.08	-37.87	< 2e-16	-2.83	0.13	-21.18	< 2e-16
search_basisother	0.97	0.06	15.3	< 2e-16	1.17	0.11	10.19	< 2e-16
search_basisplain view	-0.81	0.09	-9.1	< 2e-16	-0.86	0.16	-5.55	0
search_basisprobable cause	-3.14	0.1	-30.73	< 2e-16	-2.83	0.17	-16.17	< 2e-16
reason_for_stopsuspect vehicle	-0.18	0.08	-2.18	0.03	-0.11	0.14	-0.79	0.43
month2	0.06	0.05	1.25	0.21	0	0.09	-0.03	0.97
month3	-0.04	0.05	-0.94	0.35	-0.11	0.08	-1.39	0.16
month4	-0.04	0.05	-0.91	0.36	0.04	0.08	0.54	0.59
month5	-0.01	0.05	-0.15	0.88	-0.02	0.08	-0.23	0.82
month6	0.08	0.05	1.69	0.09	-0.04	0.08	-0.44	0.66
month7	0	0.05	0	1	0.06	0.08	0.72	0.47
month8	0.02	0.05	0.46	0.65	0.15	0.08	1.78	0.07
month9	0.04	0.05	0.78	0.43	0	0.08	-0.05	0.96
month10	0.14	0.05	2.94	0	0.06	0.09	0.73	0.46
month11	0.05	0.05	0.96	0.34	0.07	0.09	0.83	0.41
month12	0.14	0.05	2.85	0	0.28	0.09	3.32	0
weekdayWeekend	0.06	0.02	2.41	0.02	0.11	0.04	2.73	0.01

Appendix B

Variable	Standard Logistic Regression		Bootstrap					
			Percentile		Normal		Basic	
	lower	upper	lower	upper	lower	upper	lower	upper
(intercept)	0.93	1.71	0.98	1.69	0.99	1.68	0.96	1.67
time_of_dayevening	-0.08	0.02	-0.08	0.03	-0.08	0.03	-0.08	0.03
time_of_daymorning	-0.21	-0.09	-0.21	-0.09	-0.20	-0.09	-0.21	-0.09
time_of_daynight	0.07	0.18	0.06	0.18	0.07	0.18	0.07	0.18
district2	0.02	0.38	0.03	0.38	0.02	0.37	0.02	0.37
district3	-0.28	0.05	-0.27	0.04	-0.26	0.04	-0.27	0.04
district4	0.04	0.47	0.02	0.48	0.03	0.47	0.03	0.49
district5	-0.61	-0.24	-0.62	-0.23	-0.62	-0.23	-0.63	-0.24
district6	-0.03	0.30	-0.03	0.30	-0.04	0.30	-0.03	0.29
district7	-0.82	-0.43	-0.82	-0.40	-0.83	-0.42	-0.84	-0.43
district8	-0.01	0.35	0.00	0.35	0.00	0.34	-0.01	0.34
subject_age	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
subject_raceblack	0.65	1.27	0.70	1.25	0.68	1.20	0.65	1.20
subject_racehispanic	0.27	0.93	0.33	0.91	0.30	0.87	0.27	0.86

subject_raceother	-1.98	0.56	-2.46	0.31	-3.06	2.47	-1.46	1.32
subject_raceunknown	-0.47	0.40	-0.39	0.35	-0.41	0.33	-0.42	0.32
subject_racewhite	0.31	0.92	0.36	0.91	0.33	0.86	0.30	0.85
subject_sexmale	-0.28	-0.20	-0.28	-0.20	-0.28	-0.20	-0.28	-0.19
officer_assignment2nd district	-0.35	0.02	-0.35	0.02	-0.36	0.02	-0.36	0.01
officer_assignment3rd district	-0.23	0.13	-0.21	0.11	-0.21	0.11	-0.21	0.11
officer_assignment4th district	-0.66	-0.20	-0.65	-0.19	-0.66	-0.19	-0.67	-0.20
officer_assignment5th district	-0.21	0.19	-0.22	0.20	-0.22	0.20	-0.22	0.21
officer_assignment6th district	-0.77	-0.41	-0.78	-0.40	-0.77	-0.40	-0.77	-0.40
officer_assignment7th district	0.54	0.95	0.53	0.95	0.54	0.95	0.54	0.96
officer_assignment8th district	-0.37	0.01	-0.38	0.00	-0.37	0.00	-0.37	0.01
officer_assignmentother	-0.21	0.11	-0.22	0.11	-0.21	0.12	-0.21	0.12
officer_assignmenttraffic	-0.72	-0.43	-0.73	-0.42	-0.72	-0.43	-0.73	-0.42
typevehicular	-0.76	-0.66	-0.77	-0.65	-0.77	-0.65	-0.77	-0.65
contraband_foundnot searched	-4.03	-3.61	-4.02	-3.62	-4.01	-3.62	-4.01	-3.62
contraband_foundtrue	1.92	2.20	1.90	2.22	1.90	2.22	1.91	2.22
frisk_performedtrue	0.32	0.46	0.30	0.48	0.30	0.48	0.30	0.48
search_persontrue	1.66	1.94	1.65	1.94	1.66	1.94	1.65	1.94
search_vehicletrue	-3.23	-2.92	-3.25	-2.91	-3.24	-2.89	-3.24	-2.90
search_basisother	0.85	1.10	0.84	1.11	0.84	1.10	0.83	1.10
search_basisplain view	-0.99	-0.64	-0.97	-0.61	-1.00	-0.63	-1.01	-0.65
search_basisprobable cause	-3.34	-2.94	-3.36	-2.93	-3.34	-2.92	-3.35	-2.91
reason_for_stopsuspect vehicle	-0.34	-0.02	-0.39	0.00	-0.37	0.01	-0.36	0.03
month2	-0.03	0.15	-0.05	0.16	-0.04	0.16	-0.04	0.17
month3	-0.13	0.05	-0.13	0.05	-0.14	0.05	-0.14	0.04
month4	-0.13	0.05	-0.14	0.06	-0.14	0.05	-0.14	0.06
month5	-0.10	0.08	-0.10	0.09	-0.10	0.09	-0.10	0.08
month6	-0.01	0.17	-0.02	0.17	-0.02	0.17	-0.02	0.17
month7	-0.09	0.09	-0.09	0.10	-0.10	0.09	-0.10	0.09
month8	-0.07	0.11	-0.07	0.11	-0.07	0.11	-0.07	0.12
month9	-0.06	0.13	-0.06	0.13	-0.06	0.13	-0.06	0.13
month10	0.05	0.23	0.05	0.23	0.04	0.23	0.05	0.23
month11	-0.05	0.14	-0.06	0.15	-0.06	0.15	-0.06	0.16
month12	0.04	0.24	0.03	0.24	0.04	0.24	0.04	0.25
weekdayweekend	0.01	0.10	0.01	0.10	0.01	0.10	0.01	0.10

Appendix C—Bonferroni Correction

	X	p_train	p_test	p_train_bonferroni	p_train_BH
1	(Intercept)¬†	0	0	0	0
2	time_of_dayevening¬†	0.32	0.7	1	0.42378378
3	time_of_daymorning¬†	0	0.03	0	0
4	time_of_daynight¬†	0	0	0	0
5	district2¬†	0.03	0	1	0.05068966
6	district3¬†	0.18	0.14	1	0.252
7	district4¬†	0.02	0.29	0.98	0.0362963
8	district5¬†	0	0	0	0
9	district6¬†	0.12	0.44	1	0.17294118
10	district7¬†	0	0	0	0
11	district8¬†	0.07	0.62	1	0.11064516
12	subject_age¬†	2.00E-16	0	9.80E-15	3.92E-16
13	subject_raceblack¬†	0	0	0	0
14	subject_racehispanic¬†	0	0.02	0	0
15	subject_raceother¬†	0.37	0.75	1	0.44219512
16	subject_raceunknown¬†	0.87	0.33	1	0.91744681
17	subject_racewhite¬†	0	0.02	0	0
18	subject_sexmale¬†	2.00E-16	0	9.80E-15	3.92E-16
19	officer_assignment2ndDistrict¬†	0.08	0.02	1	0.1225
20	officer_assignment3rd District	0.58	0.43	1	0.64590909
21	officer_assignment4th District	0	0.15	0	0
22	officer_assignment5th District	0.95	0.75	1	0.96979167
23	officer_assignment6th District	0	0	0	0
24	officer_assignment7th District	0	0	0	0
25	officer_assignment8th District	0.06	0.49	1	0.098
26	officer_assignmentOther	0.54	0.45	1	0.61534884
27	officer_assignmentTraffic	0	0	0	0
28	typevehicular¬†	2.00E-16	2.00E-16	9.80E-15	3.92E-16
29	contraband_foundNot searched	2.00E-16	2.00E-16	9.80E-15	3.92E-16
30	contraband_foundTRUE	2.00E-16	2.00E-16	9.80E-15	3.92E-16
31	frisk_performedTRUE	2.00E-16	0	9.80E-15	3.92E-16
32	search_personTRUE¬†	2.00E-16	2.00E-16	9.80E-15	3.92E-16
33	search_vehicleTRUE¬†	2.00E-16	2.00E-16	9.80E-15	3.92E-16
34	search_basisother¬†	2.00E-16	2.00E-16	9.80E-15	3.92E-16
35	search_basisplain view¬†	2.00E-16	0	9.80E-15	3.92E-16
36	search_basisprobable cause	2.00E-16	2.00E-16	9.80E-15	3.92E-16

37	reason_for_stopSUSPECT VEHICLE	0.03	0.43	1	0.05068966
38	month2¬†	0.21	0.97	1	0.28583333
39	month3¬†	0.35	0.16	1	0.43974359
40	month4¬†	0.36	0.59	1	0.441
41	month5¬†	0.88	0.82	1	0.91744681
42	month6¬†	0.09	0.66	1	0.13363636
43	month7¬†	1	0.47	1	1
44	month8¬†	0.65	0.07	1	0.70777778
45	month9¬†	0.43	0.96	1	0.50166667
46	month10¬†	0	0.46	0	0
47	month11¬†	0.34	0.41	1	0.43842105
48	month12¬†	0	0	0	0
49	weekdayWeekend¬†	0.02	0.01	0.98	0.0362963

Appendix D—Code

```
#Importing data and setting aside Test
```{r, message=FALSE}
library(tidyverse)
library(lubridate)
library(caret)
library(ROCR)
library(boot)

new_orleans_file <- "hp256wp2687_la_new_orleans_2019_08_13.csv.zip"

new_orleans <-
  read_csv(new_orleans_file)

set.seed(1)
categories <- sample(1:2, size = nrow(new_orleans), replace = TRUE,
prob = c(0.8, 0.2))
train_uncleaned <- new_orleans[categories == 1,]
test_uncleaned <- new_orleans[categories == 2,]
```


#Data cleaning for Train set
```{r}
train_cleaned <-
  train_uncleaned %>%
  filter(year(date) >= 2012) %>%
  mutate(
    time_of_day =
      case_when(
        hour(time) < 6 ~                "night",
        hour(time) >= 6 & hour(time) < 12 ~   "morning",
```
```

```
        hour(time) >= 12 & hour(time) < 18 ~  "afternoon",
        hour(time) >= 18 ~              "evening"
      )
  ) %>%
  select(
    arrest_made, date, time_of_day, district, subject_age,
subject_race, subject_sex,
    officer_assignment, type, contraband_found, frisk_performed,
search_person,
    search_vehicle, search_basis, reason_for_stop
  ) %>%
  group_by(reason_for_stop) %>% #cleaning out entries with more than
one reason for stop
  filter(n() > 100) %>%
  ungroup() %>%
  group_by(officer_assignment) %>% #cleaning out entries with more
than one officer assignment
  filter(n() > 5) %>%
  ungroup() %>%
  mutate(
    officer_assignment =
      recode(
        officer_assignment,
        "FOB" = "Other",
        "ISB" = "Other",
        "MSB" = "Other",
        "NCIC" = "Other",
        "PIB" = "Other",
        "Reserve" = "Other",
        "SOD" = "Other",
        "Superintendent" = "Other"
      ),
    district = as.factor(district),
    contraband_found = as.character(contraband_found),
    contraband_found = if_else(is.na(contraband_found), "Not
searched", contraband_found),
    search_basis = if_else(is.na(search_basis), "Not searched",
search_basis),
    arrest_made = if_else(arrest_made == TRUE, 1, 0),
    month = as.factor(month(date)),
    weekday = wday(date, label = TRUE),
    weekday = if_else(weekday %in% c("Sun", "Sat"), "Weekend",
"Weekday")
  ) %>%
  filter_all(~ !is.na(.)) %>%
  select(-date)

set.seed(1)
categories_2 <- sample(1:2, size = nrow(train_cleaned), replace =
TRUE, prob = c(0.8, 0.2))
orleans_train <- train_cleaned[categories_2 == 1,]
orleans_valid <- train_cleaned[categories_2 == 2,]
```

###Best Regression Model: Elastic Net with All Interactions

```{r}
variables_for_dummy <-
```

```
  c("time_of_day", "district", "subject_race", "subject_sex",
"officer_assignment",
    "type", "contraband_found", "frisk_performed", "search_person",
"search_vehicle",
    "search_basis", "reason_for_stop", "month", "weekday",
"arrest_made"
  )

orleans_train_dummified <-
  fastDummies::dummy_cols(
    orleans_train,
    select_columns = variables_for_dummy,
    remove_most_frequent_dummy = TRUE
  ) %>%
  select(-variables_for_dummy) %>%
  mutate_at(vars(-subject_age), ~ scale(.))

ctrl <- trainControl(method = "none", number = 10, savePredictions =
TRUE)

train.elastic <-
    train(
      subject_age ~ . + .:.,
      data = orleans_train_dummified,
      method = "glmnet",
      trControl = ctrl
)
```

###Best Classification Model: Base Logistic

```{r}
logit_model <-
  glm(arrest_made ~ . , family = binomial(link = 'logit'), data =
orleans_train)

confint(logit_model)

```

###Base logisitic with gender interactions
```{r}
logit_model_inter <-
  glm(arrest_made ~ . + subject_sex * . , family = binomial(link =
'logit'), data = orleans_train)

summary(logit_model_inter)
```

###Data cleaning for test set
```{r}
test_cleaned <-
  test_uncleaned %>%
  filter(year(date) >= 2012) %>%
  mutate(
```

```
    time_of_day =
      case_when(
        hour(time) < 6 ~                "night",
        hour(time) >= 6 & hour(time) < 12 ~   "morning",
        hour(time) >= 12 & hour(time) < 18 ~  "afternoon",
        hour(time) >= 18 ~              "evening"
      )
  ) %>%
  select(
    arrest_made, date, time_of_day, district, subject_age,
subject_race, subject_sex,
    officer_assignment, type, contraband_found, frisk_performed,
search_person,
    search_vehicle, search_basis, reason_for_stop
  ) %>%
  group_by(reason_for_stop) %>% #cleaning out entries with more than
one reason for stop
  filter(n() > 100) %>%
  ungroup() %>%
  group_by(officer_assignment) %>% #cleaning out entries with more
than one officer assignment
  filter(n() > 5) %>%
  ungroup() %>%
  mutate(
    officer_assignment =
      recode(
        officer_assignment,
        "FOB" = "Other",
        "ISB" = "Other",
        "MSB" = "Other",
        "NCIC" = "Other",
        "PIB" = "Other",
        "Reserve" = "Other",
        "SOD" = "Other",
        "Superintendent" = "Other"
      ),
    district = as.factor(district),
    contraband_found = as.character(contraband_found),
    contraband_found = if_else(is.na(contraband_found), "Not
searched", contraband_found),
    search_basis = if_else(is.na(search_basis), "Not searched",
search_basis),
    arrest_made = if_else(arrest_made == TRUE, 1, 0),
    month = as.factor(month(date)),
    weekday = wday(date, label = TRUE),
    weekday = if_else(weekday %in% c("Sun", "Sat"), "Weekend",
"Weekday")
  ) %>%
  filter_all(~ !is.na(.)) %>%
  select(-date)
```


###Running the regression model on the test set
```{r}
orleans_test_dummified <-
  fastDummies::dummy_cols(
    test_cleaned,
```

```
    select_columns = variables_for_dummy,
    remove_most_frequent_dummy = FALSE
  ) %>%
  select(-variables_for_dummy) %>%
  mutate_at(vars(-subject_age), ~ scale(.)) %>%
  select(
    -c(
      time_of_day_evening, district_3, subject_race_black,
subject_sex_male,
      `officer_assignment_3rd  District`, type_vehicular,
      `contraband_found_Not searched`, frisk_performed_FALSE,
search_person_FALSE,
      search_vehicle_FALSE, `search_basis_Not searched`,
      `reason_for_stop_TRAFFIC VIOLATION`, month_3, weekday_Weekday,
      arrest_made_0
    )
  )

# Predict
pred_cv_elastic <- predict(train.elastic, orleans_test_dummified)


orleans_test_dummified %>%
  cbind(pred_cv_elastic) %>%
  mutate(error = (subject_age - pred_cv_elastic) ^ 2) %>%
  summarize(rmse = sqrt(mean(error, na.rm = TRUE)))
```


###Running the classification model on the test set
```{r}
pred <-
  test_cleaned %>%
  mutate(
    pred = predict(logit_model, .),
    pred_clean = as.factor(if_else(pred > 0, 1, 0)), arrest_made =
as.factor(arrest_made)
  ) %>%
  pull(pred_clean)


Y <-
  test_cleaned %>%
  mutate(arrest_made = as.factor(arrest_made)) %>%
  pull(arrest_made)

confusionMatrix(pred, Y)

pred_train <-
  orleans_train %>%
  mutate(
    pred = predict(logit_model, .),
    pred_clean = as.factor(if_else(pred > 0, 1, 0)), arrest_made =
as.factor(arrest_made)
  ) %>%
  pull(pred_clean)

Y_train <-
```

```
  orleans_train %>%
  mutate(arrest_made = as.factor(arrest_made)) %>%
  pull(arrest_made)

confusionMatrix(pred_train, Y_train)

pred_valid <-
  orleans_valid %>%
  mutate(
    pred = predict(logit_model, .),
    pred_clean = as.factor(if_else(pred > 0, 1, 0)), arrest_made =
as.factor(arrest_made)
  ) %>%
  pull(pred_clean)

Y_valid <-
  orleans_valid %>%
  mutate(arrest_made = as.factor(arrest_made)) %>%
  pull(arrest_made)

confusionMatrix(pred_valid, Y_valid)
```


#PART TWO

```{r}
summary(logit_model)
```

###Running Model on the test set
```{r}
logit_model_test <-
  glm(arrest_made ~ ., family = binomial(link = 'logit'), data =
test_cleaned)

summary(logit_model_test)


```


###Bootstrapping

```{r}
##Manual Attempt
# variables <- names(sample_logit$coefficients)
#
# variables <-
#   variables[!variables %in% c("search_basisNot searched",
"reason_for_stopTRAFFIC VIOLATION")]
#
# bootstrap_df <- as_tibble(matrix(ncol = 49, nrow = 100), row.names =
variables)
#
# colnames(bootstrap_df) = variables
```

```
#
# ptm <- proc.time()
#
# set.seed(1)
# for (i in 1:100) {
#   sample_data <- orleans_train %>% sample_n(50000, replace = TRUE)
#
#   sample_logit <- glm(arrest_made ~ ., family = binomial(link =
'logit'), data = sample_data)
#
#   bootstrap_df[i, ] <- summary(sample_logit)$coefficients[, 1]
# }
#
# proc.time() - ptm
#
# bootstrap_df %>%
#   summarize_all(sd) %>%
#   gather(key = key, value = se)
#
#
# bootstrap_df %>%
#   summarize_all(quantile(0.025, 0.975))
```

```{r}
##Using boot library
library(boot)

# Function for boot
boot_coef <- function(formula, data, indices) {
  d <- data[indices,]
  fit <- glm(formula, family = binomial(link = 'logit'), data = d)
  return(summary(fit)$coefficients[, 1])
}

ptm <- proc.time()

# Bootstrap with 1000 replications and estimate CI's
results <-
  boot(
    data = orleans_train,
    statistic = boot_coef,
    R = 1000,
    formula = arrest_made ~ .
  )

proc.time() - ptm

boot_ci <- tibble(
  conf_level = rep(NA, 49),
  perc_lower = rep(NA, 49),
  perc_upper = rep(NA, 49),
  norm_lower = rep(NA, 49),
  norm_upper = rep(NA, 49),
  basic_lower = rep(NA, 49),
  basic_upper = rep(NA, 49)
)
```

```
for (i in 1:49) {
  conf_int <- boot.ci(results, type = c("perc", "norm", "basic"),
index = i)

  boot_ci$conf_level[i] <- conf_int$percent[,1]
  boot_ci$perc_lower[i] <- conf_int$percent[,4]
  boot_ci$perc_upper[i] <- conf_int$percent[,5]
  boot_ci$norm_lower[i] <- conf_int$normal[,2]
  boot_ci$norm_upper[i] <- conf_int$normal[,3]
  boot_ci$basic_lower[i] <- conf_int$basic[,4]
  boot_ci$basic_upper[i] <- conf_int$basic[,5]
}
```

```{r}
readxl::read_xlsx("boot_ci_calculations.xlsx", sheet = "Sheet2") %>%
  ggplot() +
  geom_segment(
    aes(
      x = lower,
      xend = upper,
      y = variable,
      yend = variable,
      color = method
    ),
    size = 3,
    alpha = 0.3
  ) +
  geom_point(
    data = . %>% gather(key = key, value = value, -variable, -method),
    aes(value, variable, color = method),
    size = 3
  ) +
  labs(
    x = "Confidence Interval",
    y = "Covariate",
    title = "Difference in Standard Regression C.I. and Bootstrap
C.I.*",
    caption = "*for select variables"
  ) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5)
  )
```

```{r}
#p-value correction with Bonferroni and BH
data = csv_read("boot_ci_calculations.corrected")
```

```
data$p_train_bonferroni = p.adjust(data[,2], "bonferroni")
data$p_train_BH = p.adjust(data[,2], "BH")

indx_sign_bonf = which(data$p_train_bonferroni<0.05)
indx_sign_BH = which(data$p_train_BH<0.05)

df_sign_bonf = data[indx_sign_bonf,]
df_sign_BH = data[indx_sign_BH,]
df_nonsign_bonf = data[-indx_sign_bonf,]
df_nonsign_BH = data[-indx_sign_BH,]
```