ST309 Group Project Report:

Investigating the Nature of Severe Crimes in L.A

Candidate number	Contribution
[redacted]	50%
[redacted]	50%

Table of contents

1.0 Introduction

Predictive policing is a method of forecasting crime before it happens. One of the very first adopters of predictive policing was the Los Angeles Police Department (LAPD) in 2008. Crime has always been an existing factor in every city but is an even bigger issue in larger cities such as Los Angeles.

As of 2020, Los Angeles has a police department of 10,000 staff along with a crime rate of 387 per 100,000 residents (Magnus and Brandon, 2022). It is extremely hard for a department of 10,000 to deal with this crime rate without any information beforehand. Therefore, the LAPD has developed some programs such as LASER and PredPol to predict areas with higher chances of gun and property crimes respectively. However, LASER was shut down in 2019 due to inconsistencies in how people were selected and the public's distaste of the program.

Data analytics plays a huge role in predictive policing. To create a program that identifies hot spots of crime, we would first need to find out which factors contribute to severe crimes. For instance, the PredPol program uses data from past crimes, and prints out jurisdiction maps with areas that should be more heavily patrolled. The analysis of past crime data reveals significant insights on crime factors that help the LAPD with policing the city. As such, predictive policing and crime forecasting can be seen as a data analytics problem.

In this project, we are aiming to carry out some data analysis on different factors that could possibly affect the severity of a crime such as the age, sex and race of the victim as well as the general area and premise of the crime. We will be using supervised learning models to understand the impact that each of these factors will have on the severity of a crime. We will be using the Crime Cd variable in our dataset for our analysis with lower crime codes signifying more severe crimes. These Crime Codes represent the UCR crime codes shown in the LAPD UCR vs COMPSTAT document (2018) in our bibliography.

We have structured our analysis as a classification problem by creating a response variable with two possible values: Severe and Non-Severe. Crimes with codes lower than 300 have been assigned as severe while those with codes higher than 300 have been assigned as Non-Severe. On the basis of our analysis being a classification problem, the goals of our project are to:

1. Create several supervised learning models to identify variables that have a strong link with severe crimes
2. Explain the results obtained from our models and assess how our analysis could be improved
3. Demonstrate how crime analysis can further improve predictive policing programs

2.0 Data Preparation

Description of dataset

We had available two sets of datasets on crime in Los Angeles from Kaggle (Version 3) (Referenced in Bibliography as well). One from 2010-2019 and another from 2020 - Present. The Los Angeles dataset includes variables like victim ages, race, locations etc. These variables are commonly present in predictive policing strategies, which is why we decided that this dataset would be suitable for our goals mentioned above.

To decide which dataset to use, or if we should merge them, we compared the distributions of crime severity in both sets and found that they were roughly similar.

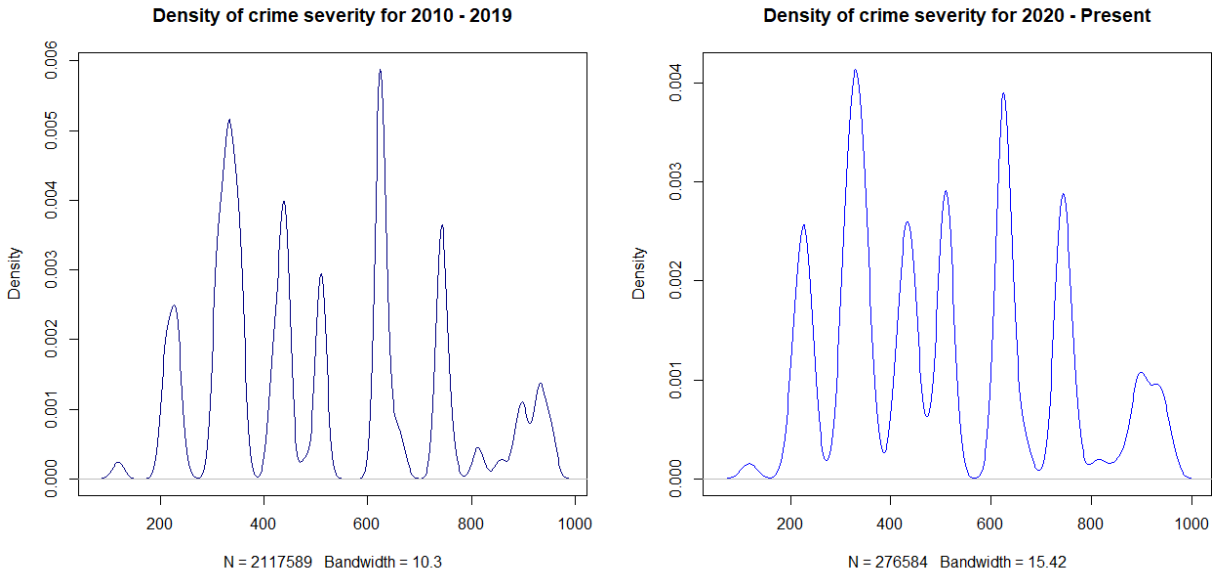


Figure 2.1.1: Distribution of severe crimes in both datasets

Upon looking at the distribution above, we decided that using the 2020 - Present data would be more relevant to predicting crime in the future as compared to data from 2010 - 2019 as it was more recent. Furthermore, this dataset included 28 attributes and 276584 records which we felt was sufficient for conducting our analysis.

Some important variables in the dataset that we use for our analysis are:

CrimeCd: A number that indicates the primary and most severe (if several) crime that was committed. CrimeCd2, CrimeCd3, and CrimeCd4 were not considered as many records did not have a secondary crime.
TimeOCC: The time that the crime occurred in 24 hour time.
Area: This variable refers to the 21 areas in Los Angeles that are managed by a unique police station. This is a categorical variable that takes values between 1 and 21.
VictAge: The age of the victim of the crime.
VictSex: The sex of the victim taking values M for male, F for female, and X for unknown.
VictDescent: The race of the victim takes several different values for different races. This description of each value was taken from the Kaggle page.
Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian.
PremisDesc: This tells us the type of location where the crime took place and whether any important structures or vehicles were nearby.
WeaponUsedCd: The type of weapon used for the crime. Null values represent no weapons used.

Data cleaning

The first thing we did before starting to work with the data was to relabel all the columns. The original dataset didn't have the most meaningful column names. We relabelled them so it was easier to identify which variable we were working with at any point in time.

The first few predictors we decided to remove were weapon descriptions, crime descriptions and area names, because these variables all had a corresponding quantitative pair to them which we found more helpful for our analysis. For example, we did not need a crime description to tell us whether a crime was severe because we had a matching crime code associated with each record. The crime code alone would be enough to tell us if the crime was a severe one as the crime codes were ordered based on the severity of the crime with the most severe crimes having the lowest crime codes. The crime description column, however, could be used as an extra description label for the crime code if needed.

As a whole, the predictors we decided on were the victim's age, sex and race as well as the area and premise of the crime, the time it occurred, and whether a weapon was involved. All other columns were removed as they wouldn't be helpful in our analysis.

Since the dataset contained more than 270,000 records, we used the dplyr package to extract out distinct records only. After making sure we had no duplicate records, we checked for null entries. For the WeaponCd column, we replaced null values with a 0 to indicate no weapon was used. For the other columns with null entries, we removed the records. In total, around 60,000 records were removed because we took out both null entries and unknown entries. Unknown entries corresponded to VictSex and VictRace. These records were removed because they would not have contributed any useful information to our analysis. We also removed records that had a VictAge of 0 as a large proportion of these crimes were shoplifting, vandalism, and vehicle robberies where there was no relevant victim. Our final dataset after cleaning had 203585 records and 28 variables.

Exploratory data analysis

We conducted exploratory data analysis on our predictors and created bar plots to gain a preliminary understanding of their relationship with severe crimes. Later in the report, we will conduct modelling to understand which of these relationships are most significant. This exploratory analysis was done for severe crimes only as that is what we are mainly trying to predict. As mentioned in our introduction, we have denoted severe crimes as those with crime codes lower than 300.

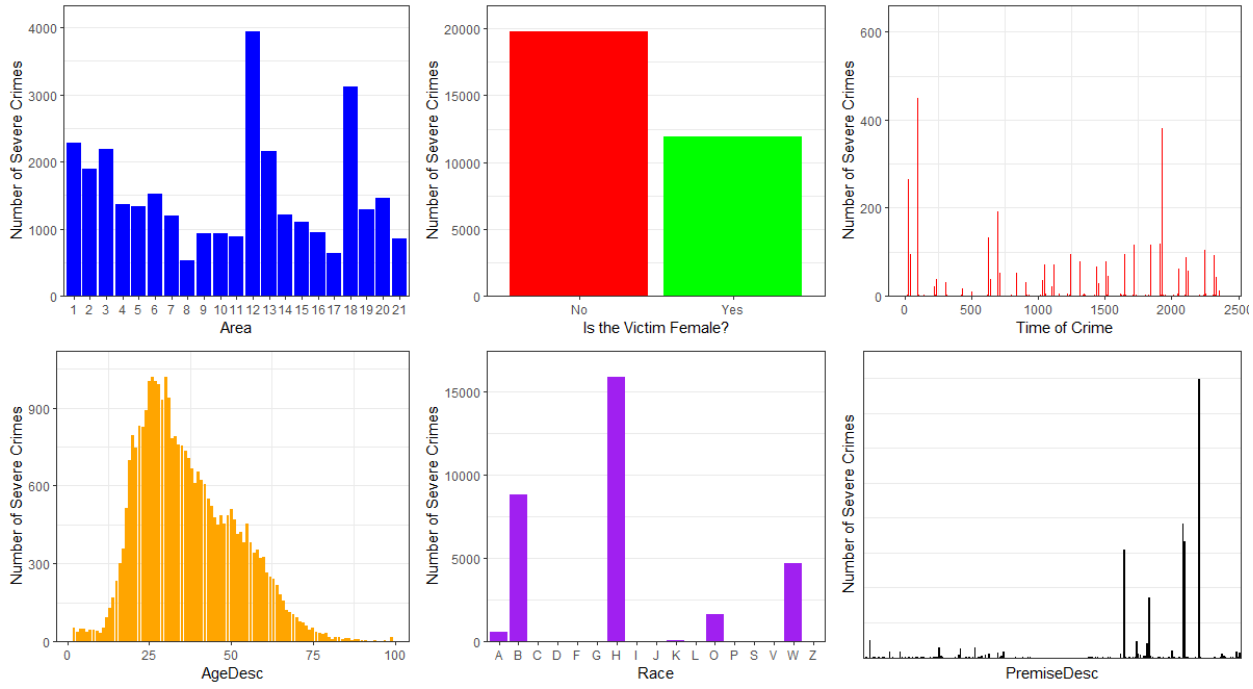


Figure 2.3.1: Exploring the relationship of each predictor with severe crimes

We discuss each figure below:

Area: We can see that the most severe crimes take place in Area Codes 12 and 18 with over 3000 severe crimes. This corresponds to the 77th Street and Southeast Areas. This is interesting as both these areas border each other. (See Map in Appendix Section 1). Our analysis gives us a good estimate of the main location in LA where severe crimes are likely to take place.

Sex: Our data suggests males are targeted for severe crimes more than females in LA.

Time Of Crime: The most severe crimes seem to happen after 8 PM and before 3 AM which corresponds to what one would typically expect.

Age: Severe crimes seem to happen most to people around the ages of 25-35.

Race: Our data suggests that people of Black and Hispanic descent seem to be targeted most for severe crimes.

PremiseDesc: There were many different premise descriptions and it was very difficult to make a plot with a readable x-axis. The top 5 Premise Descriptions shown in the plot with their respective severe crime frequencies are:

- Street with 9974 crimes.
- Sidewalk with 4792 crimes
- Single Family Dwelling with 4157 crimes
- Multi-Unit Dwelling with 3857 crimes
- Parking Lot with 2162 crimes

Area, Race, and PremiseDesc are all categorical variables having a very large number of unique values. We will hence transform these variables for our analysis. We will also transform VictSex as it is categorical.

Weapon: As one would expect, almost all severe crimes in our dataset had a weapon involved. We will look at this again when we begin the modelling process.

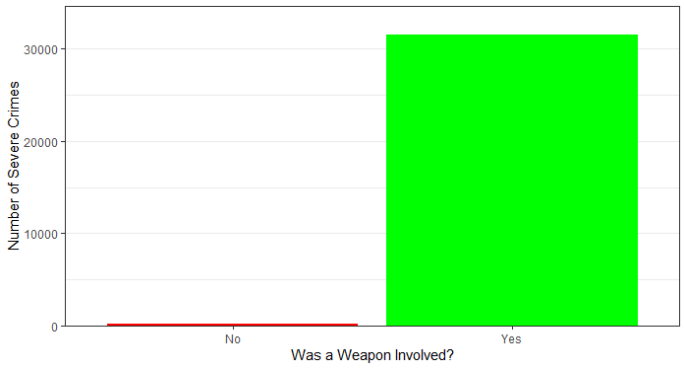


Figure 2.3.2: The number of severe crimes which had weapons involved

Data transformation and relabelling

Based on our data exploration above, we decided that transforming some variables such as time would be useful for our analysis. Most of our predictors were categorical variables rather than numerical. As such, we decided to create dummy variables for our analysis.

Variable	**Transformation**
CrimeCode	The CrimeCode column was changed from CrimeCode to Severity. This new column could only take two responses: Severe and Non-Severe. As mentioned in the introduction, the problem we are dealing with is a classification problem, therefore, the first variable we had to change was CrimeCode. Anything below 300, we considered a severe crime and have labelled them as 'Severe' because based on the description, these crimes included rape, murder etc. Anything above the 300 mark, we considered a not severe crime and have labelled them as 'Non-Severe'.
VictSex	The VictSex column was replaced by the Female column to Female. This new column could only take two responses: Yes and No. Under this new column, any records with a value of 'F' in the VictSex column was changed to 'Yes', and values of 'M' were changed to 'No'. This would indicate yes for females and no for males.
VictRace	The VictRace column was replaced by five new columns which are Asian, White, Hispanic, Black, and OtherRace. All of these new columns could only take two responses: Yes and No. The VictRace column required a lot more altering as compared to the VictSex column. For example, some of the records had Cambodian, Filipino, and Indian as their race, while some other records were labelled as Asian. Upon exploration, we realised that the biggest Asian group made up around 2.86% of the records, so, we decided that the best way to represent the other smaller Asian groups was to combine them all into one race group called 'Asian'. After

	combining these records, we created new columns for ‘Asian’, ‘White’, ‘Hispanic’, ‘Black’, and ‘OtherRace’.
WeaponCd	The WeaponCd column was changed from WeaponCd to Weapon. This new column could only take two responses: Yes and No. Earlier in the data cleaning, we assigned 0 to all crimes that did not have a weapon involved. Following this, for any record with 0, we assigned ‘No’, while records that were not equal to 0 were assigned ‘Yes’ indicating a weapon was involved in this particular case.
PremiseDesc	The PremiseCd and PremiseDesc columns come hand in hand. Both of these columns were replaced by 11 of the top premise descriptions. All of the new columns could only take two responses: Yes and No. We first ranked the number of records that had the same premise descriptions. After doing this, we extracted the top 10 premise descriptions and we found that they made up about 81.8% of the records. Because of this large proportion, we split the columns based on these top 10 descriptions and left the remaining under the ‘OtherPremise’ column.
Time	The time column was replaced by four new columns which are Morning, Day, Evening, and Night. All of the new columns could only take two responses: Yes and No. We split the 24 hour time into four intervals of six hours. The four intervals are as follows: Morning (6 am to 12 pm), daytime (12 pm to 6 pm), evening (6 pm to 12 am), and night (12 am to 6 am). Similar to the transformation of PremiseDesc and PremiseCd, we split the columns based on these intervals.
Area	The Area column was replaced by four new columns which are Valley, West, Central, and South. All of the new columns could only take two responses: Yes and No. In Los Angeles, each code represents an area that’s managed by a particular

	police station. Each of these police stations is located in one of four boroughs. We split all the records according to the boroughs they belonged in. For instance, any records with an area code of 3, 5, 12 or 18 were classified as South (Refer to Appendix Section 1 for the area map).

3.0 Data analysis

Decision trees

For the first model, we decided to fit a classification tree as it was easy for us to identify factors that influenced severe crimes. Decision trees are easy to explain, however, we should note that they are often less accurate than the other classification methods. For starters, we fitted our tree model using our dataset to see how our tree would look and decide where to go from there.

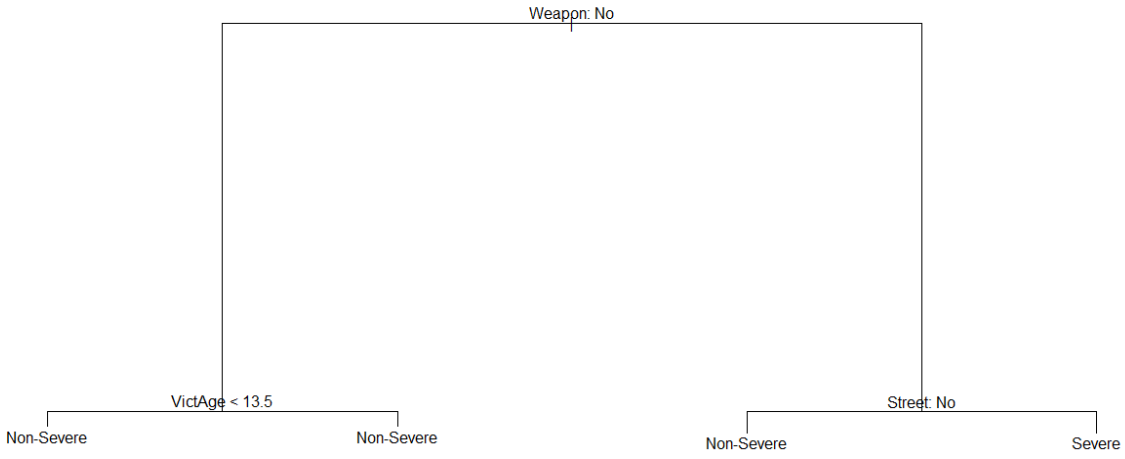


Figure 3.1.1: Decision tree using the whole crime dataset

Upon modelling our first tree, we noticed that our tree model did not tell us much about our predictors. Severe crimes only accounted for 15.6% of the records which lead us to an unsatisfactory result. Therefore, we decided to create a new set of training data from our dataset. This training data would consist of 10,000 records where 50% of them were classified as severe crimes, and the other 50% were non-severe crimes.

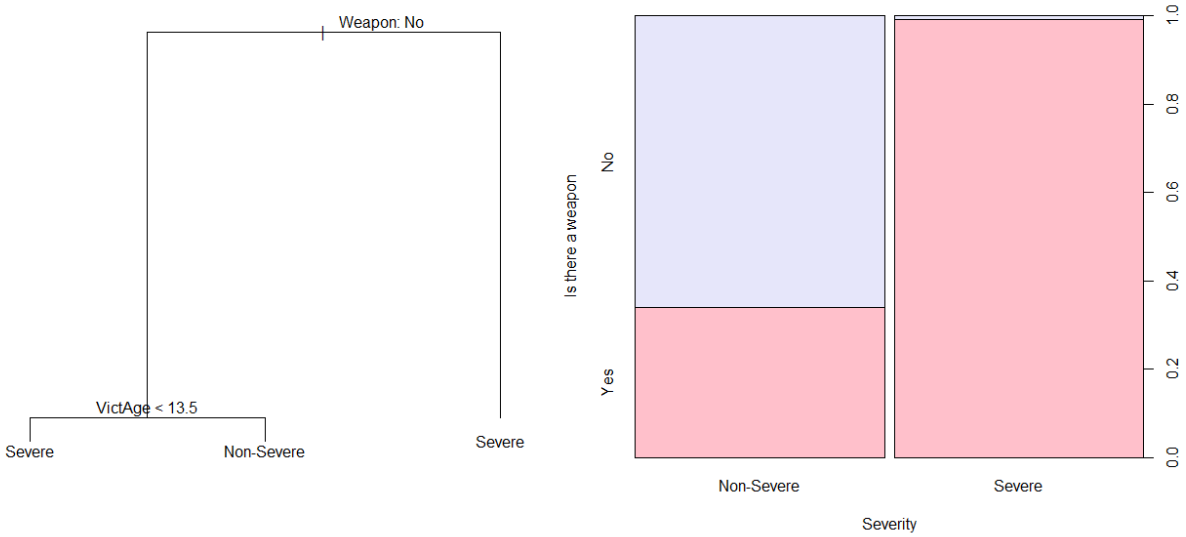


Figure 3.1.2: Pruned decision tree with the new dataset and the number of severe and non-severe crimes committed with weapons

Upon re-modelling with the training data and pruning, we identified that the Weapon predictor was an obvious, yet significant variable. For severe crimes in our training dataset, 99.2% were committed with weapons, while for non-severe crimes, only 34.0% were committed with weapons. It was extremely evident that weapons often influenced the severity of the crime. However, we wanted to explore other predictors in more detail without the bias of weapons, so we fitted another tree, this time, without the weapons predictor.

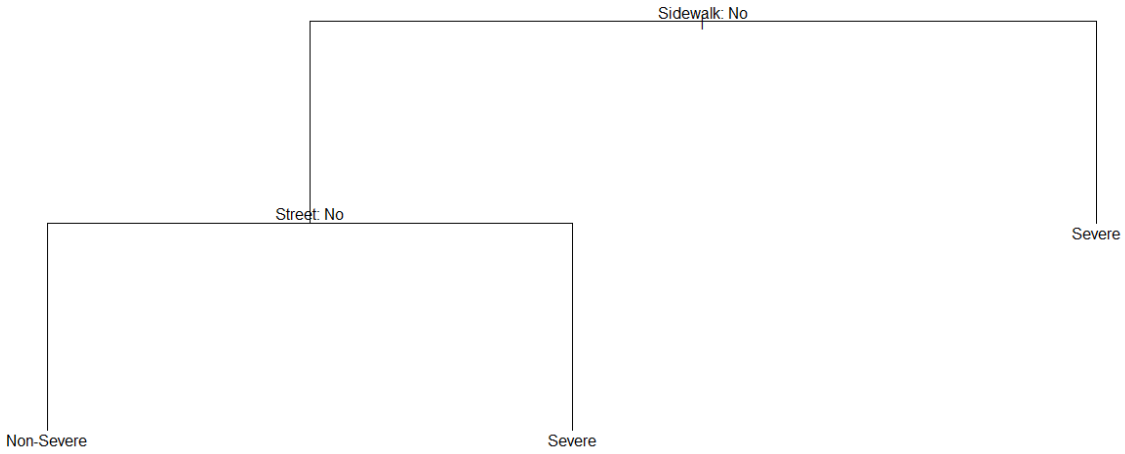


Figure 3.1.3: Pruned decision tree without weapons

After removing weapons out of the model, we found that Street and Sidewalk were the next most significant predictors in a severe crime.

Random forests

After decision trees, we decided to model our data using the random forests method. We chose the random forests method as this method uses a large iteration of decision trees to reduce overfitting. At each node of a random forest tree, a random set of predictors are considered for splitting which leads to diversity amongst the trees. (Deng, 2018). Hence, they are usually more accurate than decision trees.

Our random forests output while including weapons can be seen below:

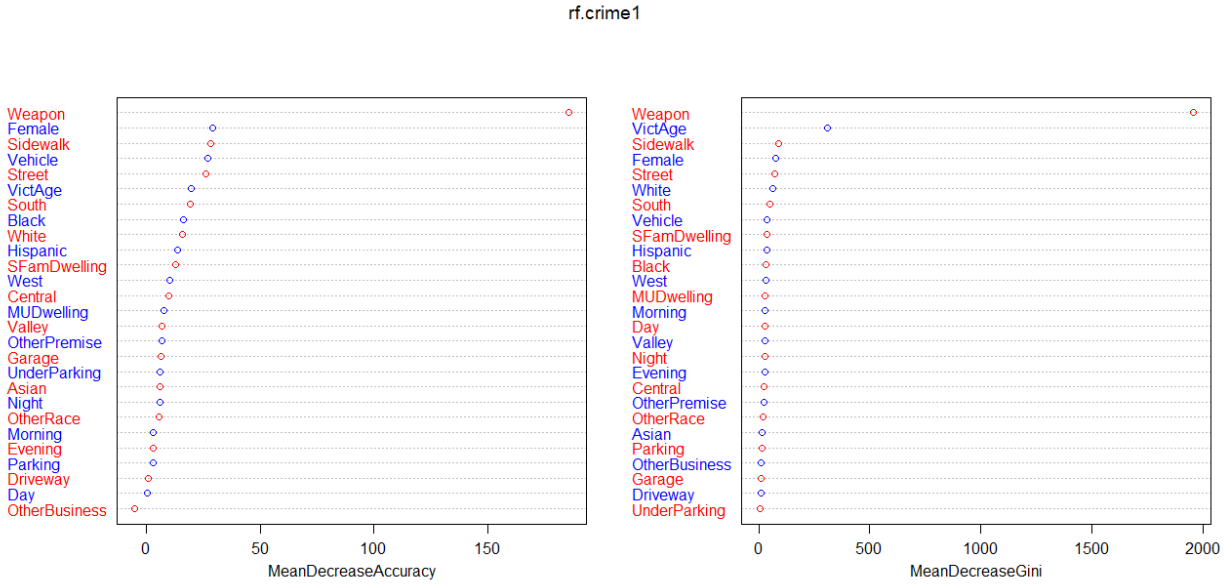


Figure 3.2.1: Random forest with weapons

From our model, we can see that the mean decrease in Accuracy and the mean decrease in Gini for Weapon far surpasses all other variables. These measures tell us how much the accuracy and node impurity of our model decreases if we remove a particular variable with the largest decrease signifying the most important variable (Navlani, 2020). Hence, we can say that Weapon is by far the most important variable in our model. However, to gain an understanding of the predictive power of the other variables, we decided to do another model without using the same training data but excluding Weapon.

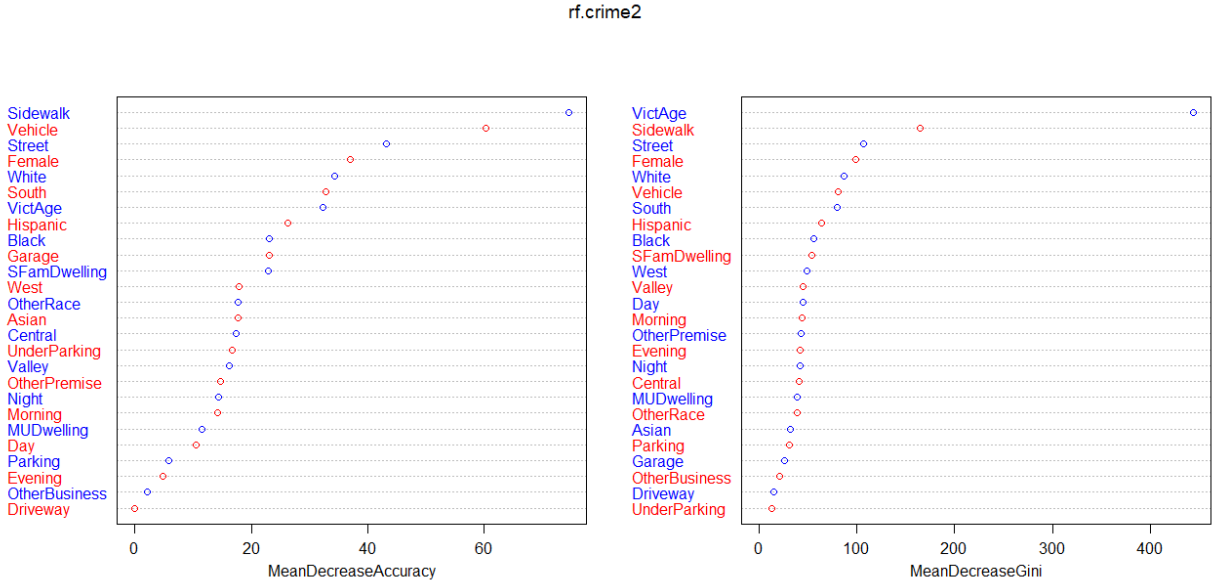


Figure 3.2.2: Random forest without weapons

From our second model, we can see that after removing weapons, Sidewalk, Street, VictAge and Vehicle have become the most important variables for our model. This aligns with what we saw in our decision tree as well.

Logistic regression

Lastly, we decided to also create a logistic regression model as our response variable (Severity) is binary and can take only two values.

A logistic regression model is used to estimate the odds of an event happening when there are several predictor variables. We are able to obtain the estimated impact of each of our predictor variables on the odds ratio. (Sperandei, 2014). We can use this model to once again look for the most important predictors that influence severe crime.

We decided to build two models, with Weapons and without Weapons. For both models, we initially built a model with all our predictor variables. We then removed variables until our final model contained only statistically significant predictor variables (with very low p-values). We continued to remove predictor variables one by one until the AIC value of our models was at its minimum.

The output of our models can be seen in Appendix Section 2. Judging from p values for our Weapons model, we can once again see that Weapons, Female, Street, and Sidewalk are the most significant variables. While modelling without weapons, we can see that VictAge, Street, Sidewalk, Female, Black, Hispanic, and South are the most significant with p values below 2.0e^-16.

ROC Curves

The ROC curves above display the trade-off between true-positives and false-positive rates. In other words, they give us information about the diagnostic ability of our classifiers. All our ROC curves can be found in Appendix Section 3, and we will be using the area under each of the curves to compare our models in the next section.

4.0 Evaluating our models

Below we have two tables encompassing the misclassification rates and area under the curve (AUC) from the ROC curves for all our models. Misclassification rate is a measure of the accuracy of our model based on the testing data we run through it. On the other hand, AUC measures the area under the ROC curve. The bigger the AUC, the more accurate the model.

It is worth noting that our misclassification rates and AUC values for the models without weapons are somewhat contradictory (figure 4.1.2) as the model with the highest misclassification rate has the lowest AUC and vice versa. We can look at figure 3.2.1 for a possible explanation. In this figure, the Weapon predictor has a value of around 170 for its mean decrease in accuracy. This could've led to the results below.

Model	**Misclassification rate**	**AUC from ROC**
Decision tree	0.305	0.822
Random forests	0.277	0.821
Logistic regression	0.278	0.887

Figure 4.1.1: Comparisons of models with weapons

Model	**Misclassification rate**	**AUC from ROC**
Decision tree	0.259	0.632
Random forests	0.317	0.673
Logistic regression	0.342	0.741

Figure 4.1.2: Comparisons of models without weapons

The logistic regression has the best performance amongst all our models for the training data set which includes Weapons. This logistic regression model has the highest AUC and roughly the same misclassification rate as the random forest model with weapons. The main goal of our analysis however was to find the factors that impact the occurrence of severe crime the most. As the logistic regression with weapons had the best performance on the testing data, we believe that the significant variables from this

logistic regression would be the most important factors for determining severe crime. However, it is also worth noting that all 3 models mostly agreed on the most significant predictor variables being Weapon, Sidewalk, Street, Female and Age.

5.0 Conclusion

Interpretation of results

Through our models, we have learned that Weapon, Sidewalk, Street, Female and Age are the factors that all of our models agree have the strongest link to severe crimes. The Weapon predictor is definitely consistent because having a weapon around would most probably lead to some sort of violence. We can also see that Race (Hispanic and Black) and Area (South) also have a somewhat strong link to severe crimes as shown in our logistic regression model.

Practicalities of our results

We will now discuss how our results could help potentially improve policing in LA and lead to the reduction of severe crimes. With weapons being a huge factor in severe crimes, introducing stricter laws with regards to carrying weapons would lead to a stark decrease in severe crimes. Furthermore, Street and Sidewalk seem to be important factors as can be seen from both our exploratory data analysis and our modelling. Hence, installing more CCTV cameras around streets and sidewalks could discourage severe crime. Area seems to be a somewhat relevant factor as well with our logistic regression and random forests model finding South as significant. This aligns with our exploratory data analysis where Areas 12 and 18 had the most severe crimes (both these areas are in the South borough). Increasing police presence and installing more CCTV cameras in these areas (perhaps around streets and sidewalks) would be a useful way of reducing severe crimes. Race can also be seen as a significant factor as our logistic regression found the victims being Black and Hispanic as important factors. Increasing police presence in areas with a larger population of these communities would also be a way of reducing severe crime.

Improvements

From the data transformation stage onwards, we noted down a couple of improvements to our analysis. For starters, our dataset consisted of mostly categorical data. Categorical data is harder to interpret at times. For instance, when we transformed the premise description column, we only took the top 10 premises and classified the remaining under the ‘OtherPremise’ category. It is possible that doing so affected our analysis. In a similar way, we also had WeaponCds available so it was possible to distinguish between types of weapons used. We believe that distinguishing between specific weapons would have led to a less one-sided result when performing our analysis on the training data with weapons. For instance,

knives or guns may be much more likely to contribute to severe crimes when compared to say a baseball bat.

Another possible way to improve this analysis would be to carry out separate analyses focusing on specific variables such as places, offenders and victims, and only combine them after analysing them separately. (RAND Corporation, 2013).

There is also the limitation of our dataset. As mentioned at the beginning, we decided to only use the 2020 - Present dataset due to the similarity in distributions. However, now looking at the results, merging both datasets may have told us a different story.

We also believe that we could potentially improve our analysis by performing bagging and boosting methods in our data analysis section. By including these results, we could have been more confident that the significant predictors we found were accurate.

6.0 Bibliography

Carmen Chan, What is a ROC Curve and How to Interpret it. [Online] Displayr. Available at: https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/ [Accessed 3 February 2022]

Deng (2018) - An Introduction to Random Forest [Online] - Available at: https://towardsdatascience.com/random-forest-3a55c3aca46d [Accessed 4th February 2022]

LAPD Patrol Area Maps (2009). [Online]. Available at: https://www.qsl.net/n6uru/lapd-maps.htm [Accessed 31 January 2022]

LAPD UCR vs CompStat Reporting Document (2018). [Online]. Available at: https://data.lacity.org/api/views/2nrs-mtv8/files/787064ac-a2f2-4cf3-a474-45036b7da937?download=true&filename=UCR-COMPSTAT062618.pdf [Accessed 24th November 2021]

James Le (2018), Decision Trees in R. [Online] Datacamp. Available at: https://www.datacamp.com/community/tutorials/decision-trees-R [Accessed 21 January 2022]

Magnus and Brandon (2022), Crime Trends in California. [Online] PPIC. Available at: https://www.ppic.org/publication/crime-trends-in-california/ [Accessed 25 January 2022]

Navlani (2018) - DataCamp - Random Forest Classifiers in Python [Online] - Available at: https://www.datacamp.com/community/tutorials/random-forests-classifier-python [Accessed 4th February 2022]

Neil, Nandish and Manan (2021), Crime forecasting. [Online] SpringerOpen. Available at: https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z#:~:text=Crime%20forecasting%20refers%20to%20the,record%20some%20unusual%20illegal%20activity. [Accessed 20 December 2021]

Alice Norga (2021), 4 Benefits And 4 Drawbacks Of Predictive Policing. [Online] Liberties. Available at: https://www.liberties.eu/en/stories/predictive-policing/43679 [Accessed 20 December 2021]

RAND Corporation (2013), Forecasting Crime for Law Enforcement. [Online] RAND Corporation. Available at:

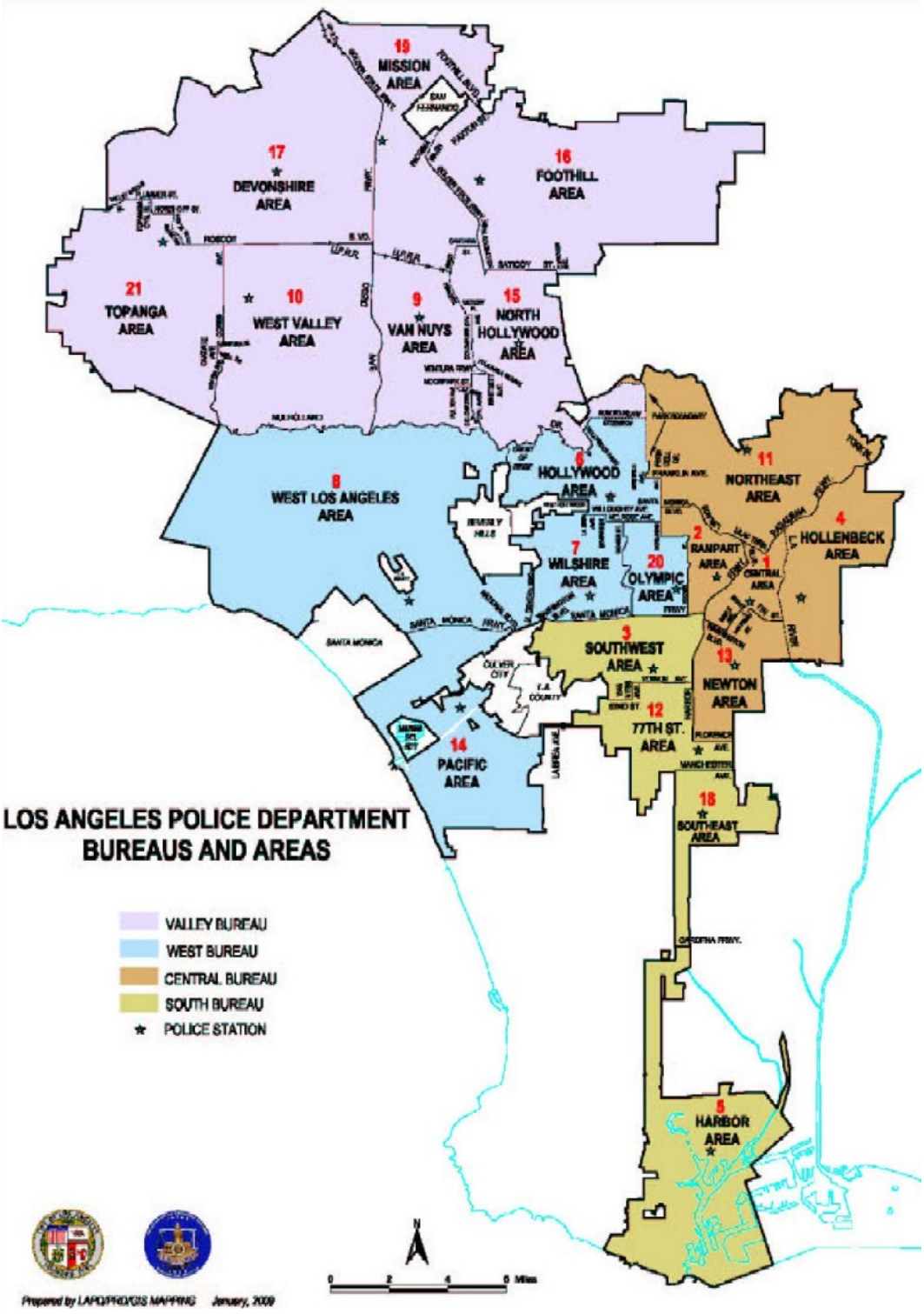
https://www.rand.org/content/dam/rand/pubs/research_briefs/RB9700/RB9735/RAND_RB9735.pdf [Accessed 6 February 2022]

Parveen Shupti - Los Angeles Crime Data 2010 - Present Kaggle Page [Online]. Available at: https://www.kaggle.com/sumaiaparveenshupti/los-angeles-crime-data-20102020?select=Crime_Data_from_2020_to_Present.csv [Accessed 20th November 2021]

Sperandei (2014) - Understanding Logistic Regression Analysis [Online] - Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/ [Accessed 5th February 2022]

7.0 Appendix

Appendix Section 1: Area code map of Los Angeles (LAPD Patrol Area Maps, 2009)



Appendix Section 2: Logistic regression results (to 3sf)

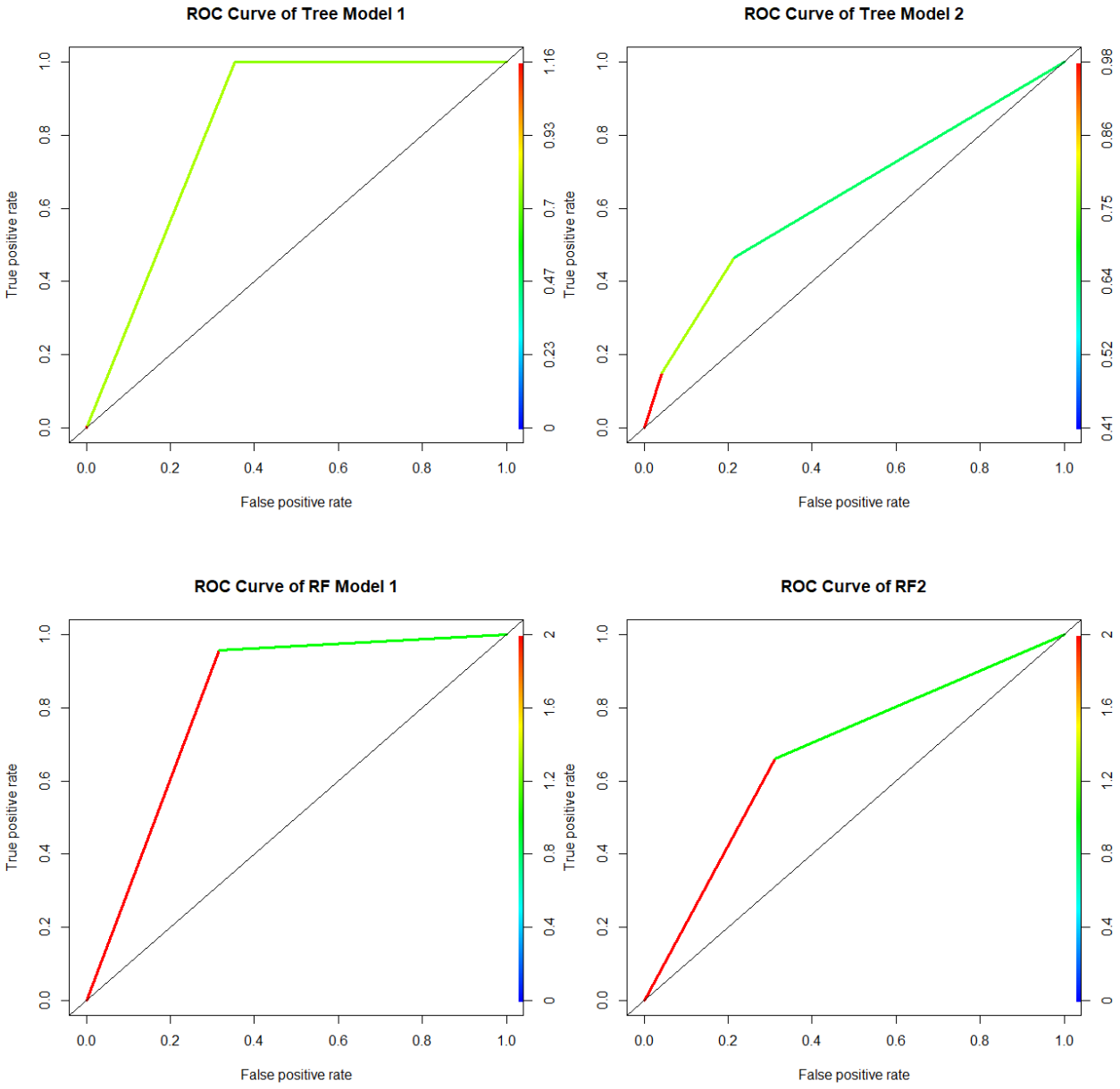
With Weapons:

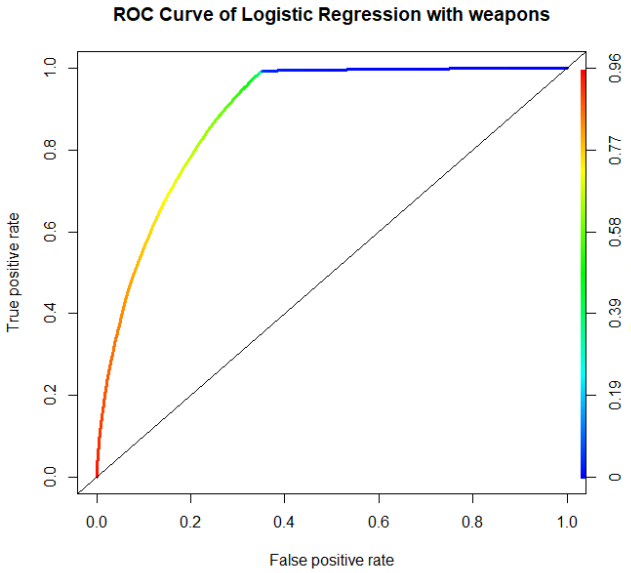
Coefficients	Estimate	p-value
Intercept	-4.01	<2e-16
VictAge	-0.00901	4.05e-06
Female	-0.686	<2e-16
Weapon	5.419	<2e-16
SFamDwelling	-0.366	4.07e-05
Street	0.721	5.54e-16
MUDwelling	-0.255	0.00601
Parking	0.217	0.0885
Sidewalk	0.757	7.57e-12
Vehicle	-1.08	5.09e-06
Black	0.506	6.08-09
Hispanic	0.277	8.55e-05
Morning	-0.281	0.000690
Day	-0.217	0.00278
Night	0.273	0.00246
Central	0.172	0.09842
South	0.615	8.21e-12
West	-0.221	0.263

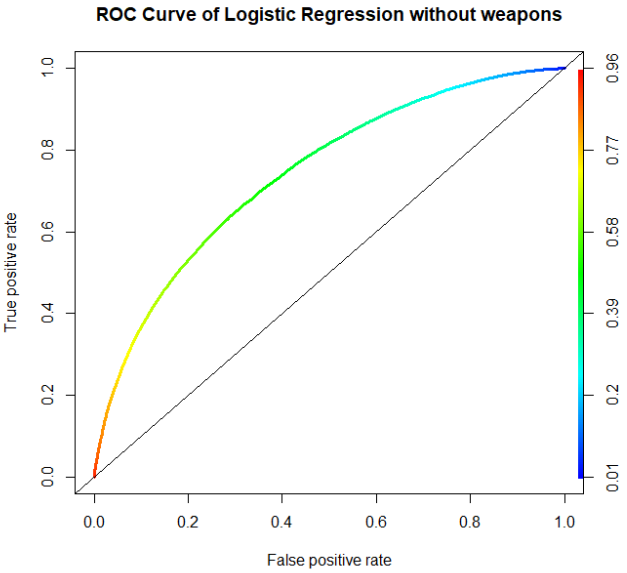
Without Weapons:

Coefficient	Estimate	p-value
Intercept	0.223	0.0134
VictAge	-0.133	< 2e-16
Female	-0.495	< 2e-16
SFamDwelling	-0.513	4.56e-14
Street	0.403	7.07e-11
MUDwelling	-0.246	0.000650
Sidewalk	1.30	< 2e-16
Vehicle	-2.16	< 2e-16
OtherBusiness	-0.291	0.038732
Garage	-2.00	2.47e-13
Driveway	-0.646	0.000401
UnderParking	-1.99	1.78e-07
Black	0.933	< 2e-16
Hispanic	0.784	< 2e-16
Morning	-0.403	2.29e-10
Day	-0.293	1.22e-07
Night	0.149	0.0257
Central	0.189	0.0143
South	0.773	< 2e-16
West	-0.172	0.0269

Appendix Section 3: ROC Curves for all models




ROC Curve of Logistic Regression with weapons
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0
0.19
0.39
0.58
0.77
0.96


ROC Curve of Logistic Regression without weapons
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.01
0.2
0.39
0.58
0.77
0.96