# Gun Violence in the United States: EDA, Predictive Modeling, and Strategic Insights for Safer Communities

McGill University
Statistical Foundations of Data Analytics - (MGSC 401)

Amine Kobeissi
amine.kobeissi@mail.mcgill.ca

December 10, 2023

## 1 Introduction

### 1.1 Background

The United States, with its distinctive cultural stance on gun ownership, stands apart from much of the world. Engraved at the heart of its constitution is the Second Amendment, granting citizens the constitutional right to bear arms. Although every American has the right to bear arms, is it the right thing to do? While this right is at the core of the American way of life, recent years have witnessed a concerning surge in gun-related incidents across the country, with many of these incidents being violent occurrences such as school shootings, drug and gang rivalries, and domestic abuse.

The juxtaposition of the constitutional right to bear arms and the rise of violent incidents necessitates a nuanced approach. It calls for a collective response from government officials, local law enforcement, the education system, and society at large, to ensure that Americans can maintain their constitutional rights, while drastically decreasing the amount of gun violence.

### 1.2 Goals

This project conducts an in-depth exploratory analysis and predictive modeling to comprehend the factors contributing to the severity of such incidents. The project delves into a comprehensive Kaggle data set documenting over 260,000 gun violence incidents in the U.S. from January 2013 to March 2018, with the aim of understanding and predicting the determinants of incident violence. The target of the investigation is the violent to total incidents ratio (Violence Severity Ratio), which is a metric constructed to serve as an indicator of the intensity and violence of gun-related incidents relative to the state and population.

To construct a model that would help explain the violent severity ratios, numerous statistical and machine learning techniques were used to understand and train the data, such as Principle Component Analysis (PCA), Random Forest Classifiers, Gradient Boosting Machines (GBM), and Logistic Regression. Furthermore, a variety of variables were analysed, such as demographic features, incident characteristics, geographical factors, and political metrics. By leveraging these techniques, our goal is to find the influential factors of the violent ratio, with the aim of creating evidence-based strategies that can effectively mitigate the severity of gun violence incidents across the USA.

# 2 Data Description

The data set that was used in this project was a Kaggle gun violence data set ranging from 2013 to 2018. It has a starting dimension of 239677 by 29, thus it is high dimensional. The data contained a variety of qualitative and quantitative variables, with important features such as a time-series component from the date, geographical data, political metrics, demographic variables, as well as qualitative and quantitative descriptors with categorical information on incidents.

## 2.1 Target Variable

In this data set, there were numerous candidates for target variables. Since the research was focused on determining the factors that impact violent gun incidents, it was natural to consider the number of killed and injured victims of gun incidents. However, there was not much variation in the data, with most values ranging between 0 and 5, and with a positive skewness, as can be seen in Figure 1. Additionally, it was difficult to tell the relative violence and significance of violent gun incidents when compared to each other due to the difference in population, and geography. Therefore, a new ratio was created as the target variable called the "Violence Severity Ratio". This ratio was calculated by considering the year, population for that year, violent gun incidents, and total gun incidents. The formula is found in Equation 1. Using this ratio allows us to have a more holistic understanding of gun violence over time across the US, so that comparing the number of violent incidents over different time periods and geographies is more consistent.

## 2.2 Time-Series Data

The time-series variable from our data set is the date-time, coming in the form of "month/day/year". To extract relevant insights from this variable, and to use it in the model, the "lubridate" package in R was used to extract the year, month, day, weak day, and quarter into new columns. Upon evaluating the distributions, data coming from the year 2013 was insignificant. Although the data started from the beginning of the year, only 278 incidents across all four quarters were reported in our data, therefore all 2013 observations were dropped. Furthermore, only the first quarter of the year 2018 was captured, however, it was not dropped from the analysis since it contained a similar frequency as the other years for quarter 1.

Upon analyzing the time-series data in terms of frequency to see when the most incidents occur, it became clear that between 2014 to 2018, absolute gun violence incidents were on the rise, with the middle quarters of each year being the most dangerous as can be seen in Figure 2. This suggests that as time moves forward, gun related incidents are becoming more frequent in the USA, and are most prevalent in the spring and summer months.

## 2.3 Geographical Data

The geographical features in our dataset are the state, city or county, address, longitude, latitude, and location description. This data provides insights into incident frequencies across various locations. Notably, a distinctive pattern emerges as there are a handful of states and cities that exhibit exceptionally higher volumes of gun incidents compared to the rest of the data, as can be seen in Figure 3. These cities, such as Chicago, Baltimore, and Washington, stand out due to their disproportionately elevated incident counts. This can be due to their relatively higher populations, urban centers, or political affiliations. Since in this data, "state" has 51 levels, and "city or country" has 12898 levels, the data is very complex and dense. Therefore, frequency-based grouping was performed so that only states with a frequency above

the mean frequency, and cities with a frequency greater than 2000 were kept as levels, binning the rest into "other".

Longitude and latitude come also in handy for plotting incidents. As can be seen in Figure 4, a plot of the US States map comparing the total incidents across states. This shows Illinois, Florida, and Texas are some of the states with the highest volume of gun incidents. Finally, the location description column contained 197588 NA's, or 82% of the data missing. Therefore this column was dropped from the analysis.

## 2.4 Political Data

The political features in the data are congressional district, state house district, as well as state senate district. These variables explain the political influence of a state, which is important in terms of gun laws. All three of these variables contained more than 50 levels to them. There were also numerous NA values, about 18% of values were missing for state house district. Due to these complexities, a similar frequency-based grouping was done, where the top 20 districts for these three labels were found, and binned into a binary feature, where if the incident is in the top 20 district it would be binned as 1, and if not it would be 0. Furthermore, to get a sense of how the president plays a role, a binary feature was added on whether the president was Trump or Obama.

## 2.5 Demographic Data

The demographic variables in this data set describe the participants of gun incidents, this includes participant age, gender, name, relationship, status (unharmed, injured, killed), and type (victim or suspect). Since there were more than half of the values missing for name, as well as over 90% of values missing for relationship, those variables were dropped. The rest of the data was stored in strings, where there were multiple people per incident. Since it is impossible to use the data as is in a model, 6 new continuous variables were created. To capture gender, a male count, female count, and total participants counts were created.

To capture age and type, the average age for the victims, the average age for suspects, and average overall participant age were calculated. Overall, the average participant age was 30 years old, the average suspect age, and average victim age was also 30. To try and better understand the impact of suspect ages on the violence of gun incidents, a K-Means clustering analysis was conducted, comparing the average suspect age with the total victims in each incident as can be seen in Figure 5. This result yielded findings that suspects around the cluster with an average age of 19 have the most violent crimes with an average of 1.07 victims per incident. Then, there is the cluster around the average age of 45, with an average of 0.95 victims per incident, followed by the 25 year old cluster with an average of 0.85 victims per incident. Finally, there is the 30 year old cluster, with an average of 0.64 victims per incident. Thus it seems like the youngest participants tend to be more violent.

## 2.6 Descriptive Data

The descriptive data is very rich in this data set. It includes a mix of continuous and textual features. The continuous features include the number of people killed, the amount of people injured, as well as the number of guns involved. Since this analysis is evaluating the violent gun incidents, a new feature called "total victims" was created by summing the total injured and total killed columns. Additionally, the number of guns involved column had 41% of its values as NA, an imputation technique was used, where if a key word indicating that a gun was involved such as "gun", "shooting", or "gun point" was found in the notes column, the NA cell would

be replaced by the number 1 since it implies that at least one gun was used, otherwise 0. The number of guns variable was normally distributed with most of the values falling between 1 and 8, and skewed to the right since some incidents such as raids, and confiscations had confiscated 400 guns in one incident.

The textual variables included gun type, if a gun was stolen, incident reports, and notes. Since the gun type, stolen status, and incident reports are stored as strings similar to the previous variables, 19 new binary variables were created. The variables created form the incident reports were done using the word cloud in Figure 6. From this word cloud, new binary features such as "drug related", "drive by", "possession related", "mass shooting" and more were created to try and understand how specific incident characteristics affects the violence of incidents. From the gun type variable, 5 dummy variables were created using the frequency counts of the most common guns used in the data. This included handguns, shotguns, rifles, 9mm, and 22 LR. So, if am incident involved one of these guns, the respective column would be assigned 1, otherwise 0. Similarly, if an incident involved a stolen gun, it would be assigned a binary encoding of 1.

In the end, after processing the data, dropping irrelevant variables and feature engineering new features, the final dimension of the data set is 239,395 entries, 73 total columns therefore the updated data set has an even higher dimension than the original.

# 3 Model Selection & Methodology

## 3.1 Dimensionality Reduction - PCA

To go about selecting a model, the first step that had to be done was to reduce the dimensionality and selecting the optimal predictors, since after creating new variables and extracting features from the original data, the data set ended up with 71 features. To begin, logic was used to eliminate variables from the data. For example, the features used in the creation of the target variable, such as "violent to total per 100k", as well as the original variables that were used for feature engineering such as "state house district".

Next, to use Principle Component Analysis (PCA), the data was filtered into quantitative continuous or binary variables, and qualitative categorical variables. Starting with all quantitative variables, a PCA analysis was conducted to visualize how much of an influence each variable had on the data set as well as on the target variable. While conducting the PCA, the analysis focused on looking for candidate variables to be removed. Therefore, there was an emphasis on vectors which were orthogonal to the target variable such as "drug related", indicating that it had low relation with the target variable, as can be seen in Figure 7, as well as vectors which were very close to each other, such as "President Indicator" and "Year", indicating potential collinearity. Furthermore, there was also a priority on keeping longer vectors, such as "top city binary", since they capture more variance in the data. Numerous iterations of PCA's were conducted until a shortlist of 14 quantitative variables were left, with a final principle component such that the first principle component explains 17.8% of the variance, and the second component explaining 13% of the variance, thus explaining a cumulative 31% of the variance with only the first two components. Additionally, as can be seen from the Percentage of Variance Explained output of PCA, with 7 components we would be able to recognize violent gun incidents with 80% accuracy. Finally, the PCA graph filtered for a Violence Severity Ratio of less than 0.50 as red, demonstrates that all of the final variables generally tend towards more violent observations.

After finalizing the list of quantitative variables, the related qualitative variables that were not compatible with PCA were added to the candidate list. For example, since both "top state binary" and "top city binary" performed well, the state and city variables were selected to get

a better understanding of which cities impact violence.

## 3.2 Base Model - Linear Regression

To first get an idea of which predictors are most influential, as well as to have a starting benchmark, a linear regression model was run on the top variables outputted from the PCA. The $AdjR^2$ was 29.78 to begin with, indicating mediocre predictive power, and that the selected predictors don't necessarily explain the variance of the Violence Severity Ratio extremely well. Thus more complex models will be tested.

## 3.3 Regression Trees - Pruning

Following the reduction of dimensionality through PCA, a candidate list of the potentially influential variables for the predictive modeling step were selected. To narrow down the list, tree based algorithms such as regression trees and the Random Forest algorithm were used, which is well-suited for this task due to its ability to handle potential collinearity among variables by using a random subset of predictors in each tree. This is important here, since there are many candidate predictors which have potential collinearity, such as "state" and "city", which are both geographic indicators. To determine an initial ranking of the variables, Regression Tree Pruning was used to narrow down which variables had the biggest impact on how observations get predicted. This was used as a first step since random forest algorithms are computationally expensive, and can have long run times when many predictors are introduced. To begin, a regression tree trained on the entire data set with a low complexity parameter of 0.000001 was created. This was done to construct a full-depth decision tree, so that the tree can capture all the relationships between the predictors and target variable. Once a complex tree was created, pruning was implemented to refine the initial tree such that it finds a balance between model complexity and predictive power. By simplifying the decision tree, the most relevant predictors, such as city and state, became more prominent, and the decision criteria helps to interpret how each variable play a role in predicting the violence severity. As can be seen in Figure 8, it suggests that states such as Florida, Georgia, and New York, cities like Columbus, Jacksonville, and Milluwake, as well as the year 2015, acts as a threshold and indicator of how violent incidents are, suggesting that they are important variables to keep.

## 3.4 Random Forest - Feature Selection & Model

Now that a more intuitive understanding of the relationship between predictors and the target variable has been established, the list of potential predictors has been narrowed down by eliminating variables that were not used in the construction of the trees. Now that only 11 variables were left, to determine feature importance the Random Forest model was used. This was relevant since in each tree the model uses a different subset of predictors, in this case each tree would select between 3 to 4 predictors. This helps the model determine the importance of each variable based on it's contribution to the predictive accuracy, while accounting for potential collinearity. Figure 9 shows the variable importance plot given by the Random Forest, the variable importance scores, computed based on the Gini impurity and mean decrease in accuracy, quantify the extent to which each variable influences the overall predictive performance. For example, "Top State" is the variable that contributes the most to the MSE, such that if it were to be removed, it would increase the MSE by 60%. Similar to the pruned decision tree, the Random Forest model suggested that the city, state, and year were top predictors, along with shooting related incidents, possession related incidents, and age cluster. Furthermore, our RF model performed relatively well compared to the base linear regression model, with an $R^2$

of 44.27, and an MSE of 0.01522. The initial performance of the Random Forest model is very strong to begin with.

To ensure that these results were reliable, the "out of the bag" (OOB) error was evaluated on every 50 trees to determine the optimal number of trees, and establish how well the model performs. The experiment concluded that the model's performance stabilizes as the number of trees increases, with the MSE and $R^2$ not changing after 200 trees, with an MSE of 0.015 and $R^2$ of 53%. This suggests that adding more trees beyond 200 will likely not improve the models predictive performance on unseen data, and the error rates suggests that the model is not over fitting with the given number of trees.

### 3.5 Gradient Boosting Machines - GBM

Finally Gradient Boosted machines were tested as well since they perform better than large complex trees by combining the predictive power of many small weak trees. The reason why GBM's were tested was to test if more complex patterns could be captured in the data. To test different hyper parameters, three GBM models were run with 1000 trees, but with different depths of 3, 4, 5. Meaning that each tree that was created had 3 to 5 branches in these models.

## 4 Results

In the end to determine the optimal model, a 70-30 train test split was done one the data, each model was re run on the 70% train split, and the MSE's were calculated based on their performance on the 30% test split. The optimal model that was selected was the GBM, with parameters of 1000 trees, and a depth of 4. Although this model had very similar performance to other models such as the Regression Tree and other GBM's, as can be seen in Table 1, it was selected since it was overall less complex. The Regression tree had a very low complexity parameter to yield the MSE of 0.01471, and the other GBM's had deeper trees, pushing these models to over fit. To ensure that the results of the selected GBM were not over fitting either, a 20 K-Fold cross validation was run on the GBM models. The final GBM MSE after the 20 K-Fold CV was 0.01470, which is still the best performing model as can be seen from Table 1. The final predictors for the model were a combination of categorical and continuous variables, selected from the filtering stages described above. Categorical variables were the age cluster (1 to 4), the city, state, congressional district, and if there was a possession tag in the report. The continuous variables were year and the number of male participants involved in the incident.

### 4.1 Variable Importance

Since GBM's are complex models, and the interpretation is not as straight forward as linear regressions, feature importance results, partial dependency plots, as well as regression trees can be used to help interpret our model. As can be seen in Table 2, the GBM model has an inherent function to output the variable importance in terms of its relative influence to the target variable. Overall in the final GBM model, the "Top State" variable was the most important, with a relative importance of 59.84%. This suggests that there is a very strong association between the state and how violent incidents are based on the severity ratio. This is concurrent with the heat map created of the United States, as can be seen in Figure 4, where there are clearly certain states with more gun incidents than others, such as Illinois, California, and Florida. Similarly. "Top City" was the other geographical feature, which was the third most important variable with a relative importance of 11.64%. Thus overall, the geography of an event is quite important in determining how violent an incident will be, as well as how frequent gun incidents will occur.

The second most important variable was the year exhibiting a relative influence of 22.41%. This highlights the ongoing trend seen in the US, where overtime the number of gun incidents is increasing. Thus it will be important to evaluate the change in gun violence over the following years, while keeping a year-to-year benchmark to determine how much of an increase or decrease in gun incidents there was to track and quantify progress.

Furthermore, the "Possession Related" variable, accounting for 2.96% of the model's relative importance, and the "Males Count" variable accounting for 2.11% of the relative importance were kept in the model due to their ability to describe how gender and incident characteristics impact the violence of an event. The "Age Cluster" variable had an importance of 0.73%, suggesting that specific age groups may not influence the predicted outcomes as much as other predictors, however was kept to grasp an understanding of how age groups impact violence overtime. Finally, the "Top Congressional District" variable, had the lowest relative importance of 0.32%, but was kept to highlight the potential regional influences on the Violence Severity Ratio.

Finally, as can be seen in Table 3, the prediction results of the first 10 observations can be analysed. It seems as if when the city is a top city, or if the state is a top state, the prediction is much more accurate then when not. Furthermore, the residual errors tend to be quite small, however this might also be due to the fact that the predicted value is bounded between 0 to 1, thus making mistakes seem less severe. Therefore, in future work, it is suggested that the evaluation criteria is revised. Furthermore, testing what the severity score would be for an incident in 2019, in New York, Albany, with the a top congressional district, a possession tag, age cluster of 4 and 5 males (Table 4), predicts a severity ratio of 0.73. This is a concerning result since overall in our current data, for Albany New York, the average severity ratio was 0.68. This means that overtime, it seems like many cities will start facing more violent gun incidents if actions are not taken.

# 5 Conclusion

Overall, the model that has been created will be able to help statisticians and managers in various departments and organizations across the USA to make inference and data-driven decisions on how to decrease gun violence in the US. Based on the final model as well as the analysis, it is important to target younger generations and teach them about the dangers of guns since the 19 year old age group is the most violent one across all ages. Thus, it is crucial for the government and for teachers to encourage the youth to avoid violence as much as possible, and if individuals would still like to bear arms, then decision makers should provide accessible courses on safe gun use and best practices. Furthermore, form the final PCA, it was observed that gang related incidents and drug related incidents tend to follow more violent incidents, leading to a higher Violence Severity Ratio. This means that local law enforcement and communities should invest in efforts to diminish the influence of gangs, specifically in less affluent neighborhoods, as well as having more resources in place to decrease drug use.

In terms of the predictive models, it seemed as if the state and city matter on violence a lot. It is crucial for the government and respective states to focus on cities that have high Violence Severity Ratios on average such as Philadelphia, Chicago, Jacksonville, and Baltimore. Policymakers should consider implementing targeted interventions in states and cities with disproportionately high incident volumes. Resources and initiatives should be created to address specific regional challenges to achieve improvement. Furthermore, districts that are frequently observing high volumes of gun incidents should take action and discuss strategies with their political parties to re discuss local gun laws, and consider setting restrictions on certain gun types. It would be wise for the local law enforcement to keep an eye on individuals with possession charges, and to offer targeted services and support programs for those individuals

to have a lower probability of participating in a gun incident. Finally, it also seems that when the number of males increases in an incident, the more violent the incident gets. Therefore, it is important to create new resources for men, specifically towards men's mental health and discussion groups.

In conclusion, this study lays the groundwork for ongoing research into the dynamics of gun violence. Future investigations could dive deeper into specific regional challenges, explore the impact of policy changes, and assess the effectiveness of interventions implemented in response to predictive modeling insights.

# 6 Appendix

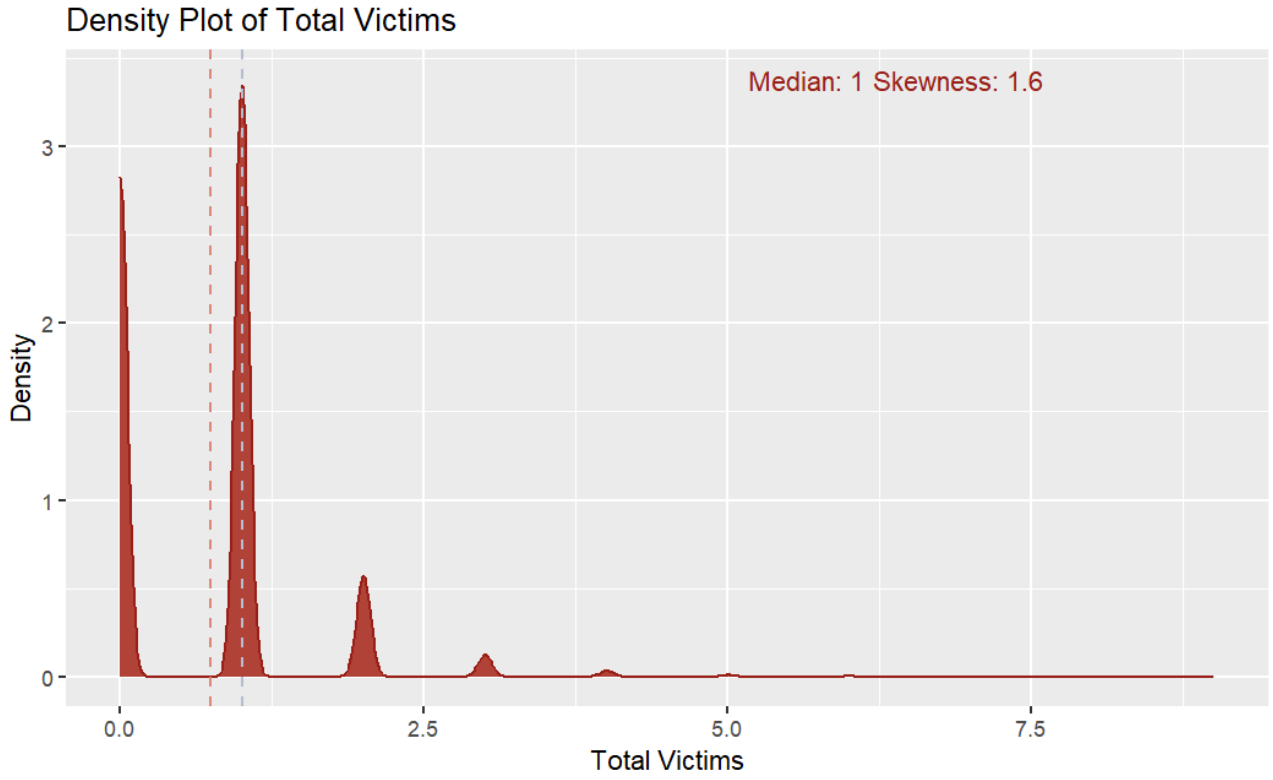


Figure 1: Density Plot of Total Victims cut with a max of 10 victims for visualization purpose. Most observations fall between 0 to 5 victims

**Violence Severity Ratio Formula**

Let $TISY$ be the total incidents by state per year, $VISY$ be the number of violent incidents by state per year.

Let $n$ be the number of incidents, $v$ be the number of violent incidents, $i$ be the year, and $j$ be the state.

$$TISY = \sum_{i=2014}^{2018} \sum_{j=1}^{51} n_{i,j}$$

$$VISY = \sum_{i=2014}^{2018} \sum_{j=1}^{51} v_{i,j}$$

$$TISY\,PopAdj = \frac{TISY}{population_{i,j}} \times 100000$$

$$VISYPopAdj = \frac{VISY}{population_{i,j}} \times 100000$$

$$ViolentSeverity = \frac{TISYPopAdj}{VISYPopAdj}$$

(1)

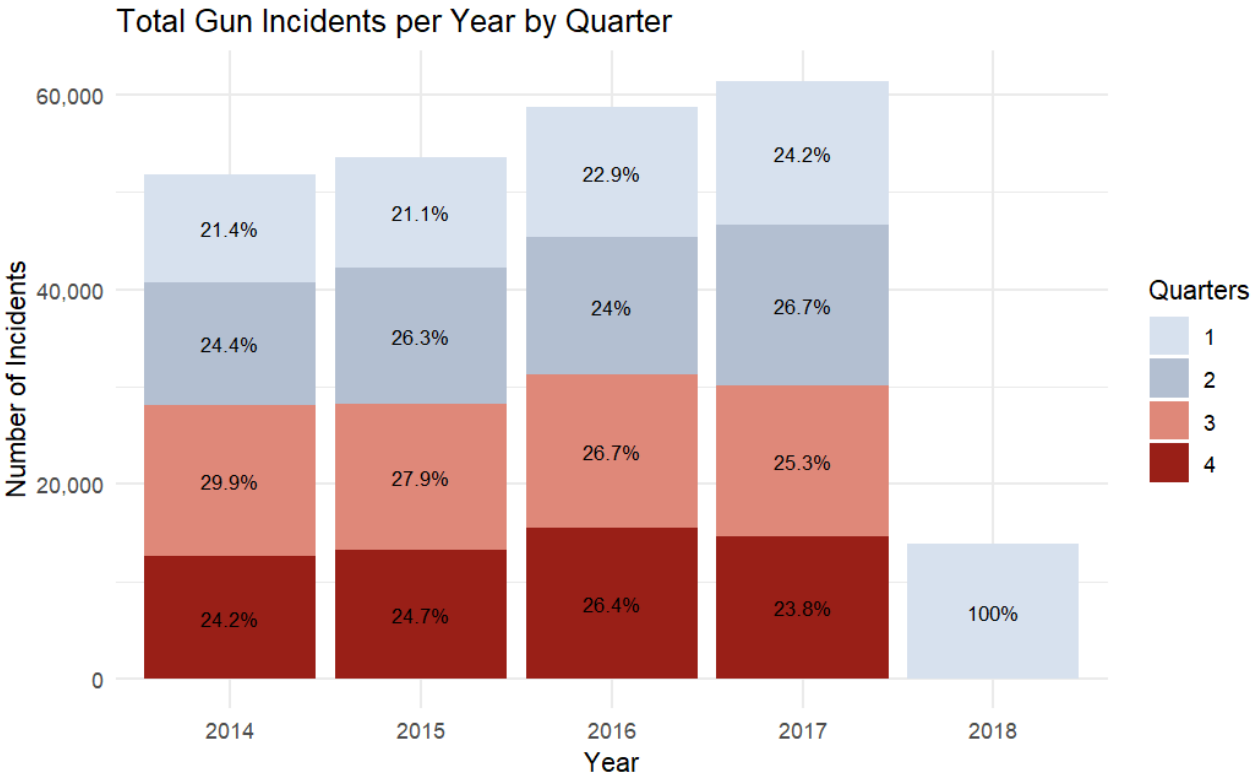


Figure 2: Year-Over-Year gun incidents per quarter. Since 2014, gun violence has been on the rise, with the middle quarters having the highest frequncy

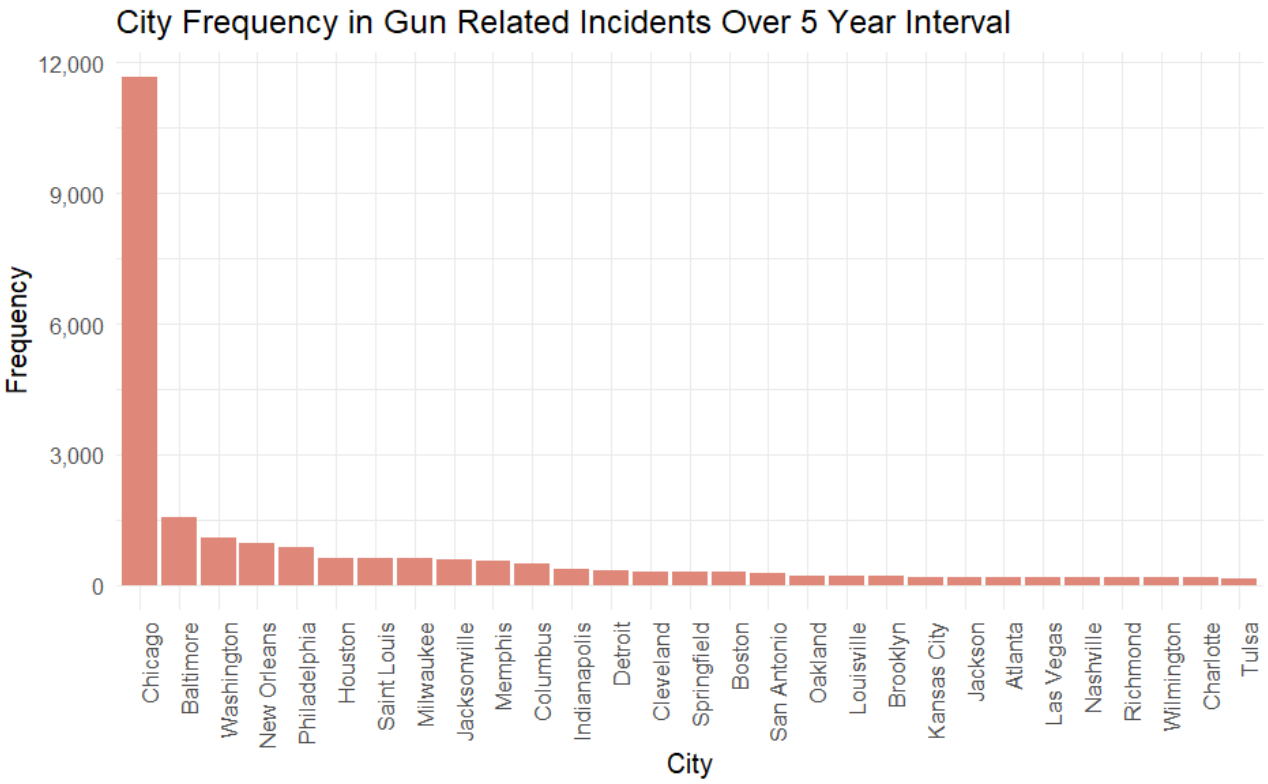


Figure 3: Frequency of Cities in the Gun Incident Data Set. Chicago is clearly a violent city

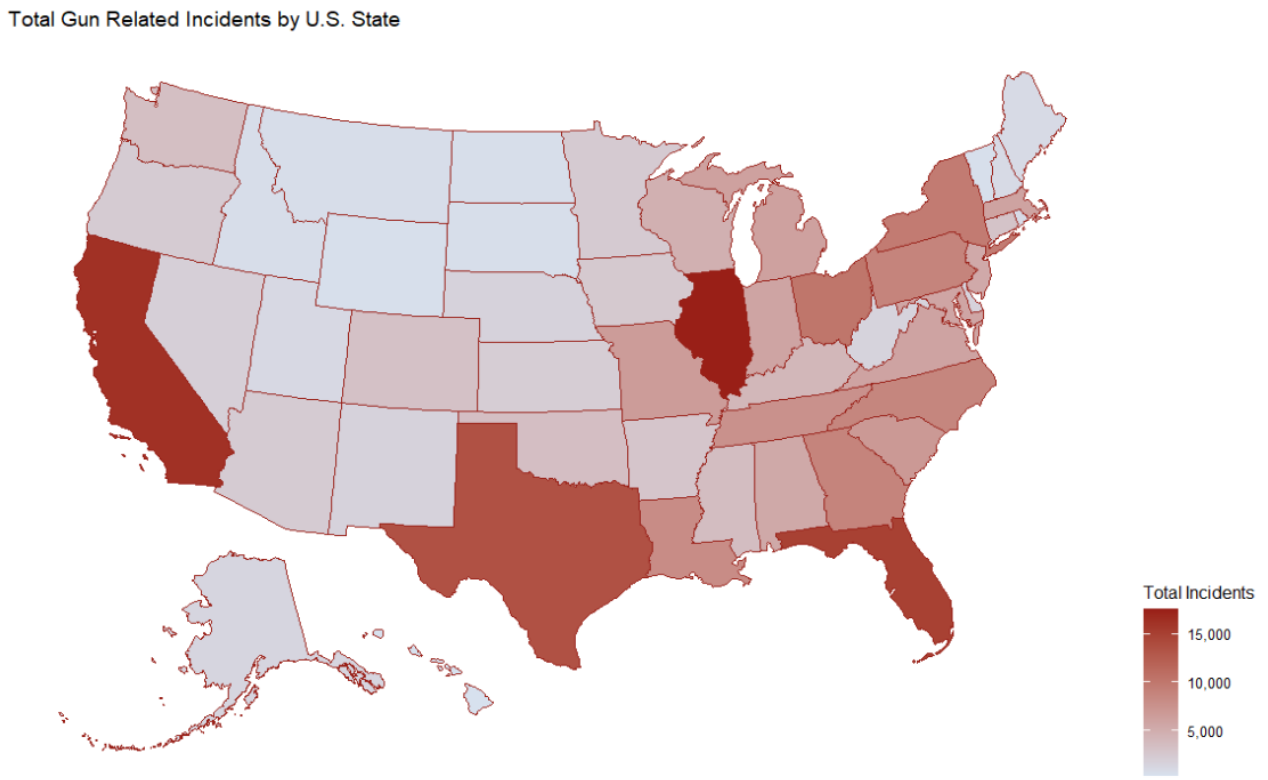


Figure 4: US Map Plotting Gun Related Incidents Across States

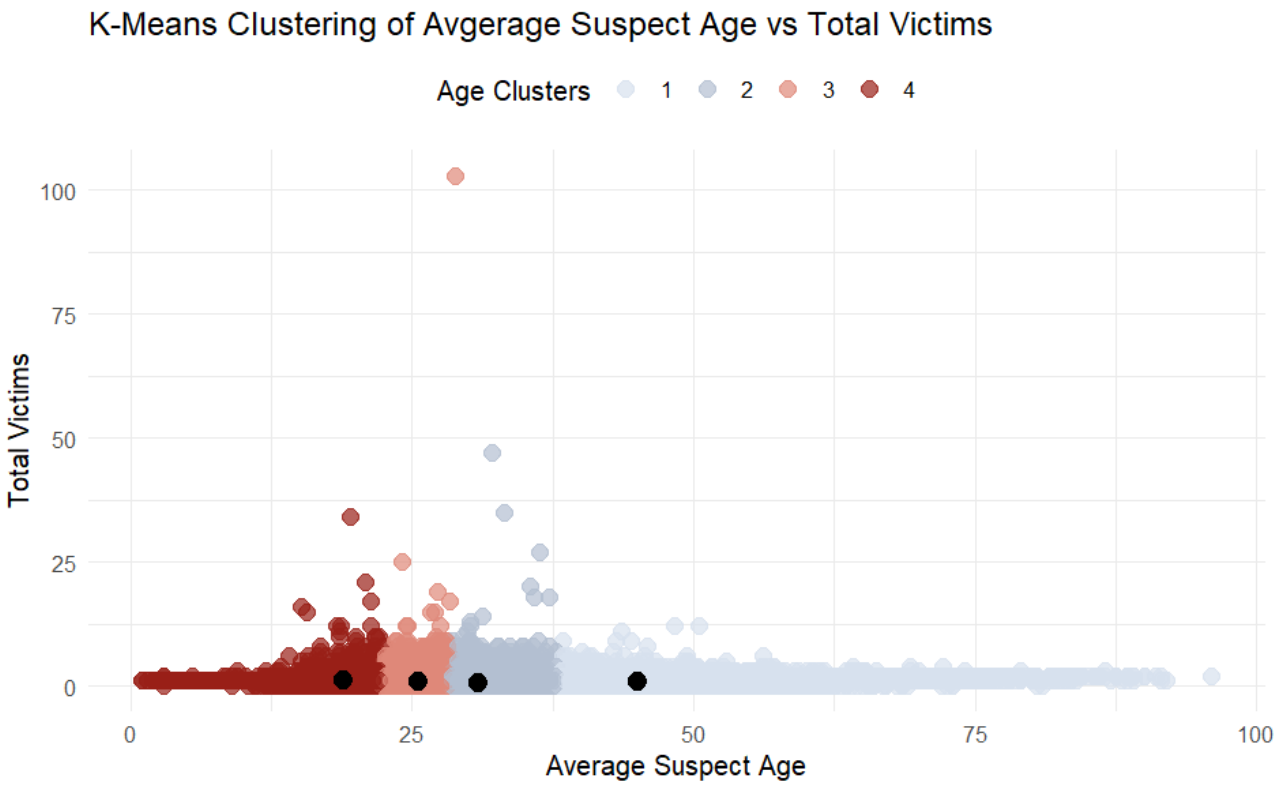


Figure 5: K-Means Clustering of Average Suspect Age vs Total Victims. Most Violent age group is 19 Year old, with average of 1.07 victims per incident. Then it is 45 year old overage age, with an average of 0.95 victims per incident, followed by 25 years old with 0.85 victims per incident. Finally there is the 30 year old age group with 0.64 vicitms per incident

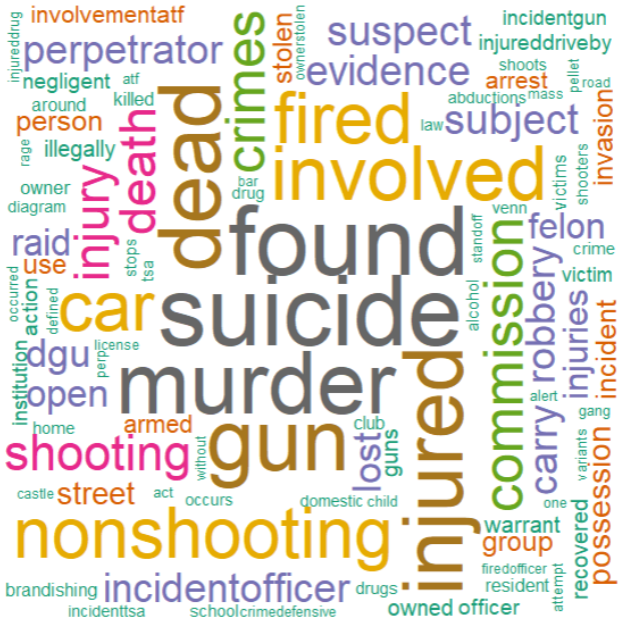


Figure 6: Word Cloud on Incident Characteristics Column

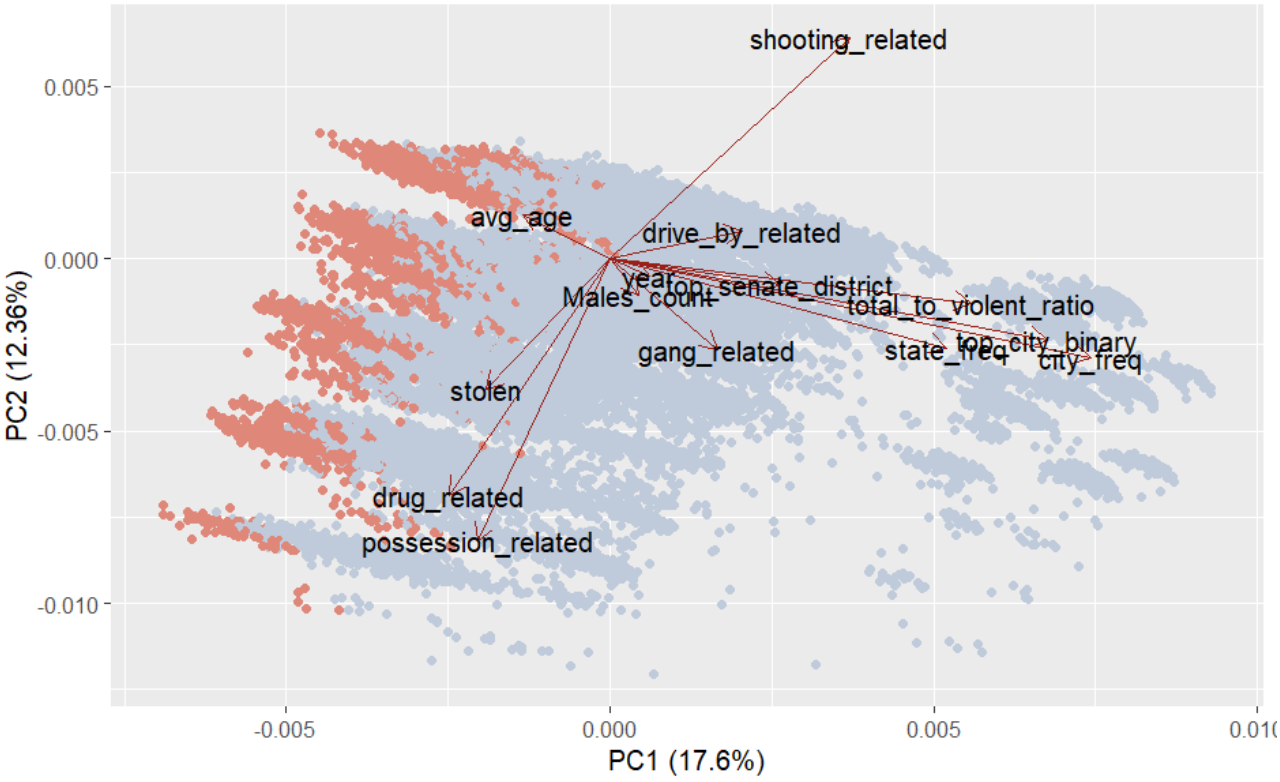


Figure 7: PCA Plot on Last Iteration

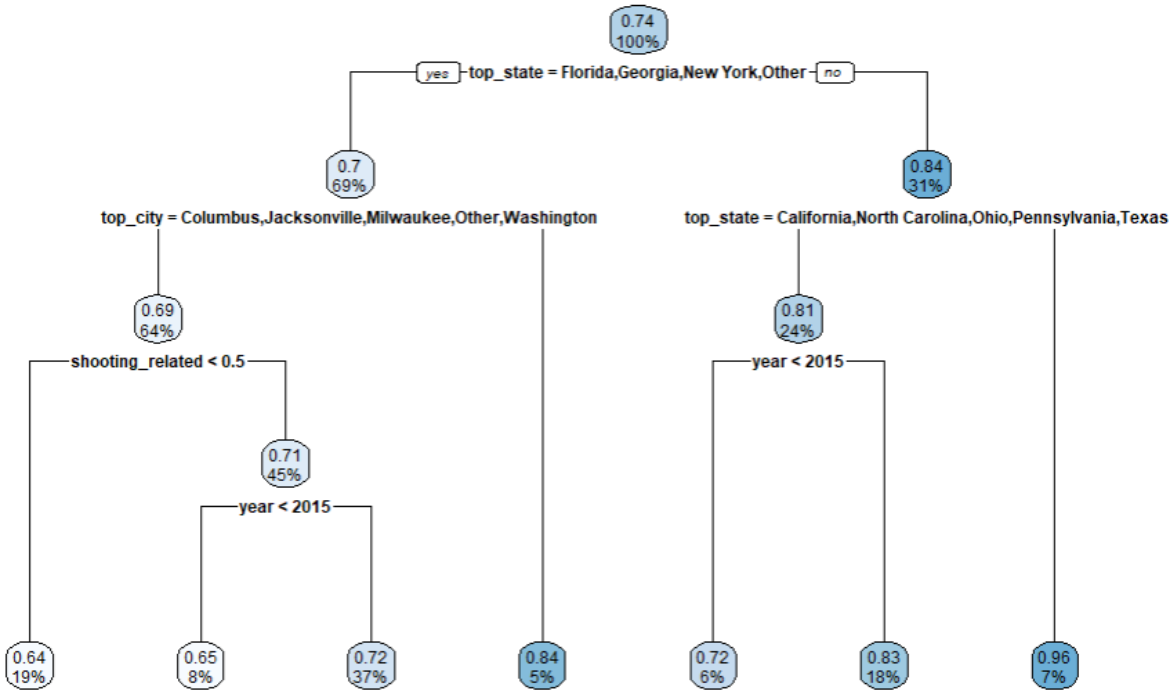


Figure 8: Pruned Regression Tree Identifying Top Predictors

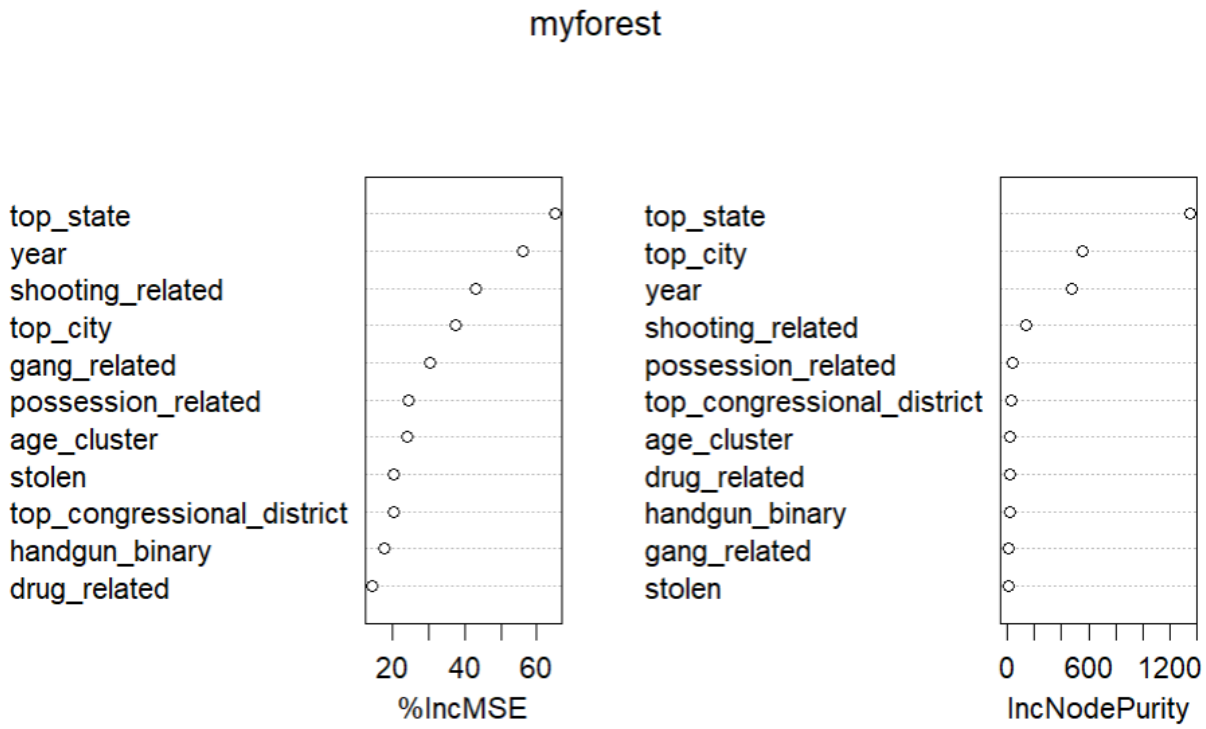


Figure 9: Random Forest Variable Importance Plot

| Model | Test MSE |
|---|---|
| Linear Regression | 0.01944 |
| Polynomial Regression | 01748 |
| Regression Tree | 0.01471 |
| Random Forest | 0.01543 |
| GBM 1 | 0.01474 |
| GBM 2 | 0.01470 |
| GBM 3 | 0.01470 |
| GBM 4 | 0.01514 |

Table 1: Final predictive performance of 8 models tested on MSE, with the Gradient Boosted Machine 2 being the optimal model for MSE

| Variable | Relative Influence |
|---|---|
| Top State | 59.838 |
| Year | 22.406 |
| Top City | 11.637 |
| Possession Related | 2.960 |
| Males Count | 2.108 |
| Age Cluster | 0.732 |
| Top Congressional District | 0.318 |

Table 2: Variable importance from the final GBM model. Most important variables are "top state", "year", and "top city"

| Residuals | State | City | Year | Top CD | Possession | Age Cluster | Males |
|---|---|---|---|---|---|---|---|
| 0.046 | Other | Other | 2,014 | 1 | 0 | 2 | 0 |
| 0.076 | Other | Other | 2,014 | 1 | 0 | 2 | 0 |
| 0.001 | Pennsylvania | Philadelphia | 2,014 | 1 | 0 | 3 | 1 |
| 0.033 | Other | Other | 2,014 | 1 | 0 | 2 | 0 |
| 0.176 | Other | Other | 2,014 | 1 | 0 | 3 | 2 |
| 0.001 | Other | New Orleans | 2,014 | 1 | 0 | 2 | 1 |
| 0.010 | New York | Other | 2,014 | 0 | 0 | 3 | 1 |
| 0.024 | Other | Other | 2,014 | 1 | 0 | 1 | 3 |
| 0.074 | Other | Other | 2,014 | 1 | 0 | 4 | 6 |
| 0.054 | Other | Other | 2,014 | 1 | 0 | 4 | 2 |

Table 3: 10 Predictions from Final GBM Model with Predictors

| Observation | Variable |
|---|---|
| New York | Top State |
| Albany | Top city |
| 2019 | Year |
| 4 | Top Congressional District |
| 1 | Possession |
| 4 | Age Cluster |
| 5 | Males count |

Table 4: Prediction Results on a new observation, with a predicted severity score of 0.73