MGSC 661 Final Project
Gun Violence

Contents

1 Introduction

Gun violence has emerged as a critical and deeply troubling issue in the United States in recent years, marked by a disturbing rise in mass shootings, homicides, and suicides involving firearms. The prevalence of guns and the ease of access to them have been central to debates and discussions on public safety and gun control policies. Moreover, the impact of gun violence extends beyond the immediate loss of life, causing long-lasting trauma and grief to families and communities. Utilizing data from the gun violence dataset spanning the years 2013 to 2018, this project aims to delve into the underlying factors contributing to fatal gun violence incidents. Key objectives are to develop predictive model that can classify an incident as fatal or not and estimate the number of fatalities in each incident, thereby providing valuable insights that could inform policy decisions and intervention strategies.

2 Data Description

2.1 Gun Violence Demographics

Gun violence in the United States is not evenly distributed but varies significantly from one state to another. This variation can be influenced by a range of factors, including local gun laws, socio-economic conditions, cultural attitudes towards firearms, and law enforcement practices. Fig.1 highlights that states like Illinois, California, Washington D.C., and Texas are among those with the highest rates of gun violence incidents in the nation.

Upon close examination of the state-wise data, Fig.2 and Fig.3 reveal that Arizona, Nevada, and Texas stand out for having the highest average number of fatalities resulting from gun violence incidents. Meanwhile, when it comes to non-fatal outcomes, Illinois, Virginia, and Tennessee emerge as the states with the highest average number of individuals injured in such incidents. These findings highlight the varying nature and severity of gun violence across the United States, underscoring the need for region-specific approaches to address this critical issue.

2.2 Numerical Data Overview

Several variables have been constructed to deepen our analysis of gun violence. The ’Subject-Suspect count’ variable, extracted from the ’participant type’ column, quantifies the count of individuals involved in an incident who are not victims. Additionally, ’Victim age’ and ’Non-victim age’ variables are calculated by averaging the ages of the victims and non-victims, respectively, involved in each incident.

Counterintuitively, Fig.4 reveals a negative correlation between the number of injuries and fatalities per incident, suggesting that events with higher injury counts do not necessarily correspond to a greater number of deaths. Furthermore, the number of firearms involved in an incident doesn't show a significant correlation with the likelihood of the event being fatal. Similarly, the number of injuries has a negative correlation with an event's fatality, challenging common assumptions about the dynamics of gun violence. These insights provide critical perspectives on the complex nature of violent incidents and their outcomes.

2.3 Categorical Data Overview

The dataset is rich in textual content, particularly within the ”incident characteristics” column, necessitates the derivation of several categorical variables to facilitate more nuanced analysis. Key variables extracted from this column include ”teens involved,” ”drug involved,” ”gang involved,” ”bar,” ”home invasion,” ”school,” and ”drive by.” These variables, distilled from the detailed descriptions, offer vital categorical distinctions, enabling a more structured and focused examination of the data. The percentage of occurring can be found in Fig.5.

Fig.6 offers a revealing snapshot of the patterns associated with gun violence incidents. Drive-by incidents, along with those involving drugs or teenagers, show a marked propensity for fatal outcomes, underscoring the deadly nature of these events. Contrary to what one might expect, gang-related incidents are shown to have fewer fatal outcomes, suggesting an alternative pattern of violence that warrants further investigation. When examining the locations commonly associated with gun violence, drive-by events, which often occur randomly, along with home invasions and incidents at bars, are more likely to result in fatalities compared to those occurring at schools. This information could be crucial for law enforcement and policymakers, providing data-driven guidance for the allocation of resources and the development of targeted preventive strategies.

3 Model Selection and Methodology

Two models are developed to classify the fatality of an incident and estimate the number of people being killed. To circumvent the challenges posed by inconsistent null values across various derived predictors, which can impede model performance, our analysis will be primarily focused on the subset of data pertaining to Illinois as it has the highest gun violence events. This targeted approach allows for a more consistent and reliable examination of the factors at play, ensuring that our model is informed by a robust data subset with greater internal consistency.

3.1 Data Preprocessing

Amidst the extensive location data available, including city, county, state senate district, and location description..etc, the decision was made to employ K-means clustering based on longitude and latitude coordinates. After testing cluster numbers ranging from 2 to 20, an optimal partitioning was identified at $k = 4$. This selection was informed by an analysis of the elbow plot displayed as Fig.7, which indicated that $k = 4$ achieves a satisfactory level of within-cluster compactness, marking it as a point where the marginal gain in reducing within-cluster variance begins to diminish.

3.2 Derived Variables in Model

The dependent variable in the classification model, which classifies an incident as fatal, is derived from the count of individuals killed. An incident is deemed fatal if the 'number of people killed' exceeds zero. As previously discussed, the model incorporates a variety of predictors, including the 'number of subject suspects' involved, the involvement of teenagers, and the average age of victims per incident. Additionally, the presence of drugs or gangs, as well as the type of location—such as bars, schools, homes (home invasion), or random locations (drive-by shootings)—are factored into the analysis. Temporal variables, specifically the month and day of the week when the incident occurred, are also included to capture potential time-related patterns in the data.

3.3 Classification Model - Fatal or Not

In gun violence incidents, where the classification of an event as fatal or non-fatal is of paramount importance, the Random Forest algorithm stands out for its robust and versatile modeling capabilities. This method is adept at managing datasets replete with a multitude of variables that characterize the complex nature of gun-related incidents, a task further complicated by the frequent occurrence of missing data. Random Forest's inherent mechanism to handle such missing values ensures that the accuracy of the model's predictions is not compromised. Crucially, this algorithm also provides an invaluable feature—variable importance ranking, which sheds light on the predictors that most significantly influence the likelihood of fatality in an incident. The interpretability of a Random Forest model offers clear insights into the factors that escalate an incident's severity, thus informing preventative strategies and policy-making.

3.4 Regression Model - Number of People being Killed

Random Forest Regression emerges as a highly effective model for predicting the number of fatalities in gun violence incidents. Its strength lies in its accuracy and adeptness at navigating complex, non-linear relationships within data, especially in gun violence statistics. Furthermore, the model employs an ensemble methodology, combining the outcomes of multiple decision trees, which substantially diminishes the likelihood of overfitting.

4 Results

4.1 Classfication Model Hyperparameter Tuning

In the process of optimizing the Random Forest Classification model, the analysis to determine the optimal number of trees is conducted, testing a range from 100 to 500 without pre-selecting features. To ascertain the most effective tree count, two distinct approaches were employed: analyzing the Out-of-Bag (OOB) error and implementing cross-validation with a train-test split of 0.3. Findings(Fig.8) indicate that the model achieves peak accuracy, marked at 0.94, with 200 trees (tested at intervals of 100). Furthermore, a tree count of 450 stands out, as it results in the lowest OOB error, recorded at a mere 0.43%. Consequently, it is concluded that the optimal number of trees for the model is 150.

4.2 Classification Model Feature Selection

Feature importance in our analysis is determined using two distinct methodologies: the intrinsic feature importance scores generated by the Random Forest model, and Principal Component Analysis (PCA).

The Random Forest model's importance scores (Fig.9) have revealed that specific features are particularly predictive of whether an incident will be fatal. These include the number of people injured, the number of guns, the involvement of teenagers and drug, the average age of victims, and whether the incident occurred during a home invasion. Interestingly, the involvement of gang does not significantly impact the fatality of an event, nor does the general location of the incident.

In Principal Component Analysis, while the detailed results are outlined as Fig.10, it is crucial to highlight that the analysis does not demonstrate a pronounced drop in the percentage of variance explained by successive principal components. The absence of a distinct 'elbow'—a point where the rate of variance explanation significantly decreases—coupled with concerns regarding interpretability, led to the decision not to employ PCA as a primary method in the final model.

4.3 Classification Model Performance

Post feature selection, there was a marginal but notable improvement in the model's predictive accuracy, as evidenced by the Out-of-Bag (OOB) error rate decreasing from 6.53% to 6.40%. This reduction, although seemingly small, is indicative of the model's enhanced generalization capability. The corresponding confusion matrix (Table1 & Table2) further elucidates this progress, revealing a reduction in both false positives and false negatives. Such a decrease is significant as it reflects the model's improved precision in distinguishing between the classes—reducing the instances where non-fatal incidents are mistakenly classified as fatal (false positives) and fatal incidents are incorrectly labeled as non-fatal (false negatives). This refinement in the predictive performance underscores the effectiveness of feature selection in honing the model's ability to more accurately identify the true nature of gun violence incidents.

4.4 Regression Model Hyperparameter Tuning and Feature Selection

Random Forest Regression was employed to estimate the count of fatalities per incident. The optimal number of trees was determined by evaluating within the spectrum of 100 to 500. It can be identified from Fig.11 that the minimum Out-of-Bag (OOB) Mean Squared Error (MSE) of 0.099 was attained with a forest consisting of 400 trees.

The Random Forest model's importance scores (Fig.12) have revealed that specific features are particularly predictive of the number of fatalities in gun violence incidents. These include the number of people injured, the number of subject suspects, the involvement of teenagers, the average age of victims, and whether the incident occurred at a bar or during a home invasion. Interestingly, the number of guns involved does not significantly impact amount of death, nor does the general location (cluster) of the incident. Contrary to common perceptions, factors like gang involvement, drug involvement, and scenarios typically associated with traumatic mass shootings, such as school or drive-by incidents, surprisingly do not significantly contribute to a higher amount of death. These insights challenge some of the usual assumptions about the dynamics of gun violence and highlight the complex nature of predicting fatal outcomes.

The significance of various predictors diverges notably when comparing their impact on determining the fatality of an incident versus estimating the total number of deaths. Analysis of feature importance in both classification and regression models reveals that while the number of guns and drug involvement are key factors in predicting the likelihood of an incident being fatal, they do not provide substantial insight into the expected number of fatalities. Conversely, incidents occurring at bars, which are less influential in predicting fatality, surprisingly emerge as more significant when estimating the death toll. This distinction underscores the nuanced role different predictors play across various modeling objectives of gun violence.

4.5 Regression Model Performance

The regression model incorporates previously identified key features for enhanced predictive accuracy. In Table 3, a slight decrease in RMSE and increase in R squared suggests that the model with selected features is slightly more accurate and better at explaining the variance in the data than the model using all features. The increase in MSE is quite large and could potentially indicate that there's an issue with the model after feature selection. It could be possible that while certain features that heavily influenced the prediction error were removed, leading to better RMSE and R squared, the overall consistency of the model suffered, hence the higher MSE. The minimal changes in MAE and OOB error rates suggest that the overall average prediction error and the model's ability to generalize have not changed significantly with the feature selection.

5 Classification / Predictions and Conclusions

In the analysis of gun violence incidents, two Random Forest models were developed, focusing on different aspects: one to classify the fatality of an incident and another to estimate the number of fatalities. The findings revealed distinct predictors in each model. For the fatality classification model, key predictors included the number of people injured, number of guns, involvement of teenagers and drugs, average age of victims, and whether the incident was a home invasion. Surprisingly, gang involvement and the incident's general location had less impact. After feature selection, the model's predictive accuracy improved marginally but notably, as seen in the reduced Out-of-Bag (OOB) error rate and fewer false positives and negatives in the confusion matrix. This improvement indicates a more precise classification of fatal and non-fatal incidents.

In contrast, the model predicting the number of fatalities found the number of people injured, the number of suspects, involvement of teenagers, the average age of victims, and incidents occurring at bars or during home invasions to be significant predictors. Interestingly, factors like the number of guns, general location, gang involvement, drug involvement, and scenarios typically associated with mass shootings, like school or drive-by incidents, did not significantly contribute to a higher death toll.

The regression model incorporates previously identified key features for enhanced predictive accuracy. Selected features refined the models, yielding a minor RMSE decrease and R squared increase, suggesting improved accuracy and variance explanation. However, a substantial MSE increase may signal issues with consistency post-selection. Changes in MAE and OOB were negligible, implying stable prediction error and generalization capability.

For management and policy-making, the implications are clear: targeted intervention strategies must be prioritized. Efforts should focus on the strong predictors identified—particularly, implementing educational and support programs to address teenage involvement and drug use, and mitigating risks in identified critical locations such as homes and bars, which are prone to higher fatality counts. Although the number of guns did not significantly predict the death toll, it was closely associated with whether an incident would be fatal, suggesting that gun control measures could be effective in reducing fatality risks.

Furthermore, these results necessitate a re-evaluation of certain assumptions, such as the influence of gang activity and the general location of incidents, which appeared to be less significant than expected. Policies and resources may need to be redirected from these traditionally overestimated risk factors to those highlighted by the data. Community engagement and educational initiatives can play a crucial role here, especially given the significance of teenage involvement as a predictor. Such programs, aimed at educating youth about the consequences of gun violence, could be instrumental in prevention.

In conclusion, integrating these data-driven recommendations into a comprehensive strategy that includes law enforcement, community outreach, and policy amendments is essential. This multifaceted approach is crucial for effectively addressing the intricate issue of gun violence, aiming to reduce both the likelihood of fatal incidents and the number of fatalities when they occur.

6 Appendix

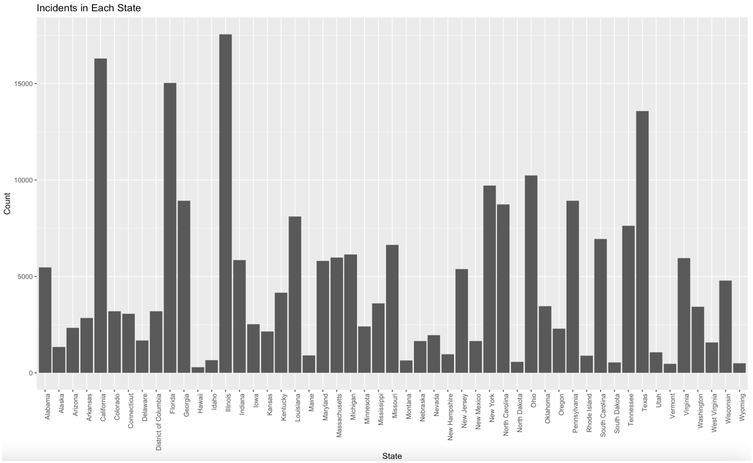


Figure 1: Number of Gun Violence Incidents by State

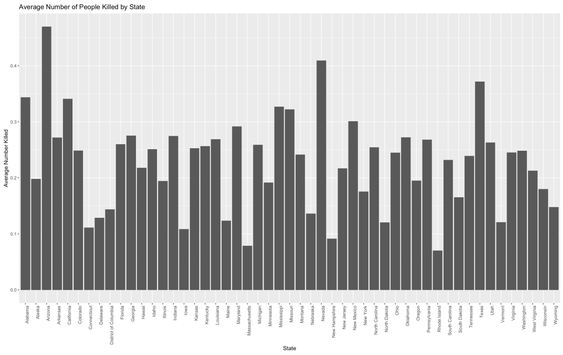


Figure 2: Average Number of People Killed by State

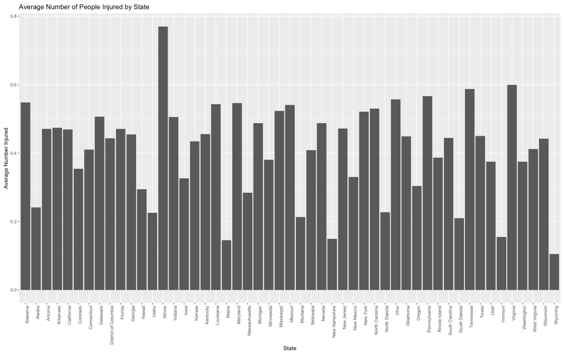


Figure 3: Average Number of People Injured by State

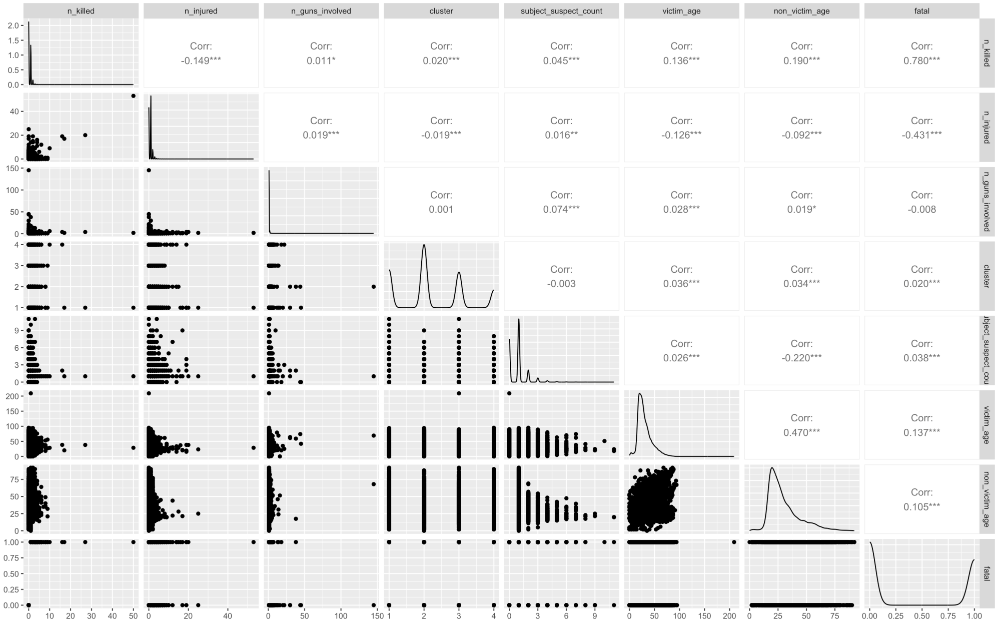


Figure 4: Distribution and Correlation of Numerical Variables

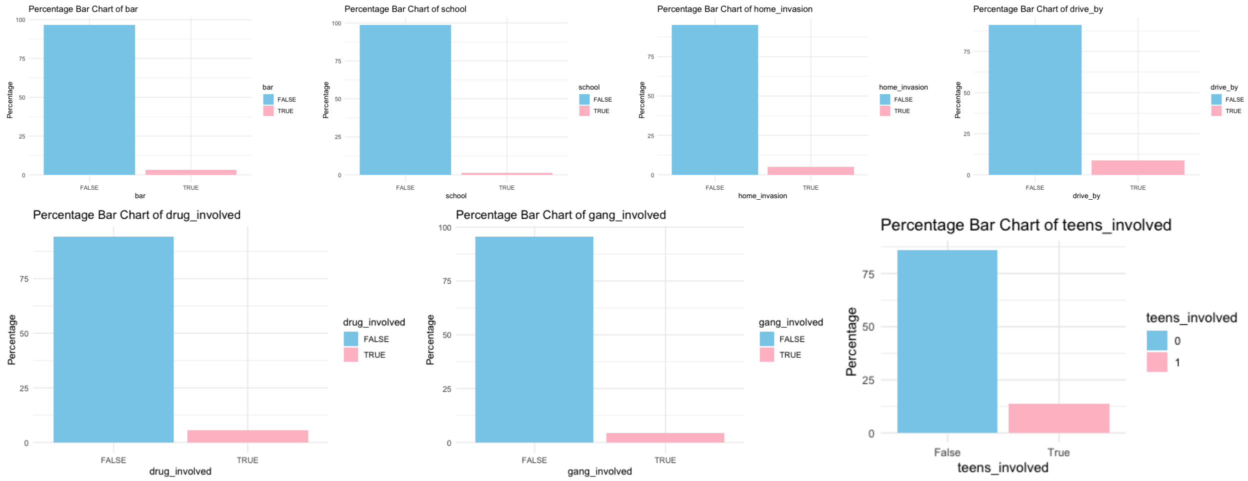


Figure 5: Percentage of Each Categorical Variables

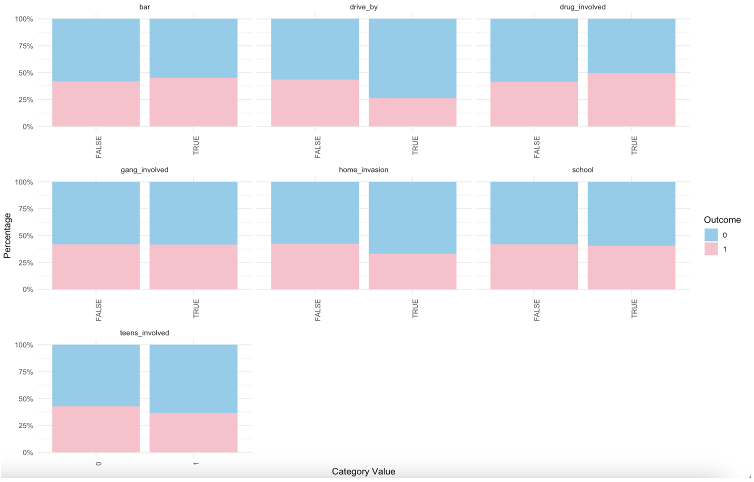


Figure 6: Percentage of Fatal Incidents by Categorical Variables

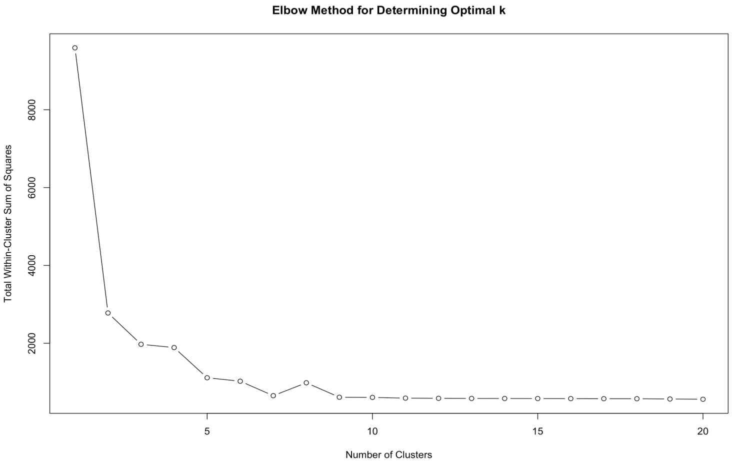
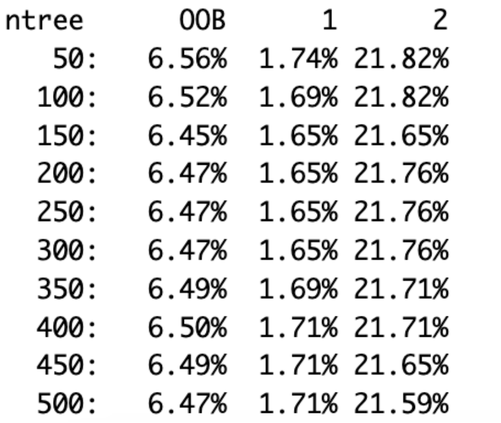
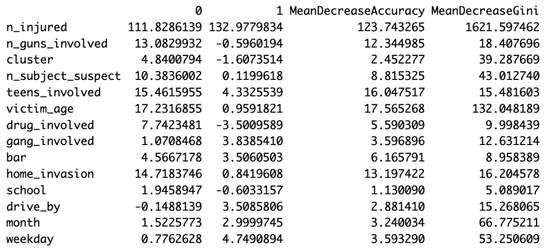


Figure 7: Optimal K for Clustering Incident Locations



```
ntree      OOB      1      2
   50:   6.56%  1.74% 21.82%
  100:   6.52%  1.69% 21.82%
  150:   6.45%  1.65% 21.65%
  200:   6.47%  1.65% 21.76%
  250:   6.47%  1.65% 21.76%
  300:   6.47%  1.65% 21.76%
  350:   6.49%  1.69% 21.71%
  400:   6.50%  1.71% 21.71%
  450:   6.49%  1.71% 21.65%
  500:   6.47%  1.71% 21.59%
```

Figure 8: Optimal Number of Trees for Classification Model



```
                          0           1 MeanDecreaseAccuracy MeanDecreaseGini
n_injured       111.8286139 132.9779834           123.743265      1621.597462
n_guns_involved  13.0829932  -0.5960194            12.344985        18.407696
cluster           4.8400794  -1.6073514             2.452277        39.287669
n_subject_suspect 10.3836002  0.1199618             8.815325        43.012740
teens_involved   15.4615955   4.3325539            16.047517        15.481603
victim_age       17.2316855   0.9591821            17.565268       132.048189
drug_involved     7.7423481  -3.5009589             5.590309         9.998439
gang_involved     1.0708468   3.8385410             3.596896        12.631214
bar               4.5667178   3.5060503             6.165791         8.958389
home_invasion    14.7183746   0.8419608            13.197422        16.204578
school            1.9458947  -0.6033157             1.130090         5.089017
drive_by         -0.1488139   3.5085806             2.881410        15.268065
month             1.5225773   2.9999745             3.240034        66.775211
weekday           0.7762628   4.7490894             3.593290        53.250609
```

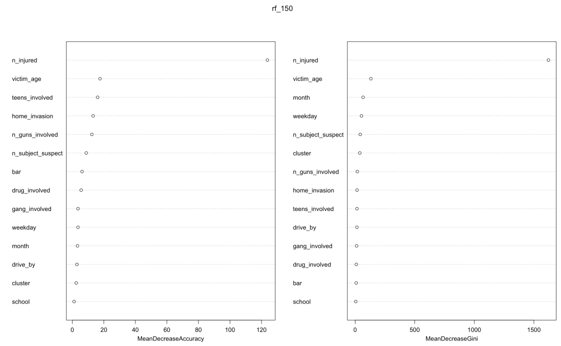
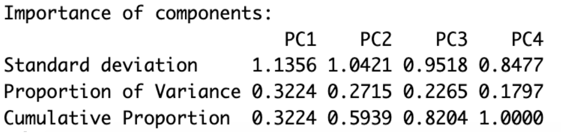


Figure 9: Random Forest Feature Importance Score for Classification model



```
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.1356 1.0421 0.9518 0.8477
Proportion of Variance 0.3224 0.2715 0.2265 0.1797
Cumulative Proportion  0.3224 0.5939 0.8204 1.0000
```

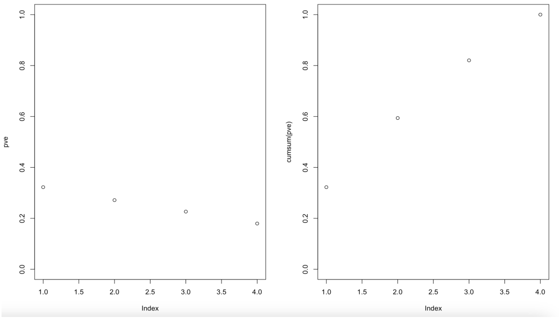


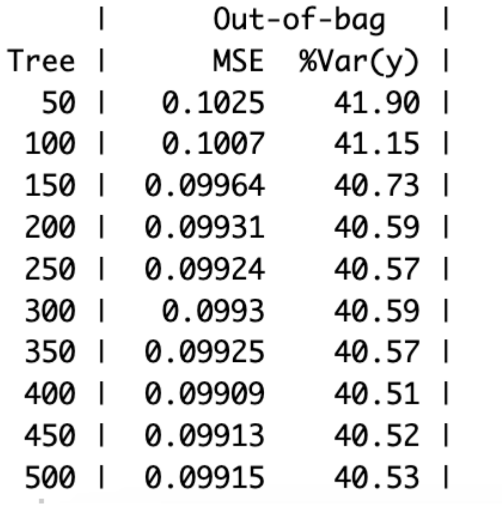
Figure 10: Principal Component Analysis for Classification model

	0	**1**	**Class.error**
0	5292	97	0.01799963
1	366	1334	0.21529412

Table 1: Confusion Matrix of Original Classification Model

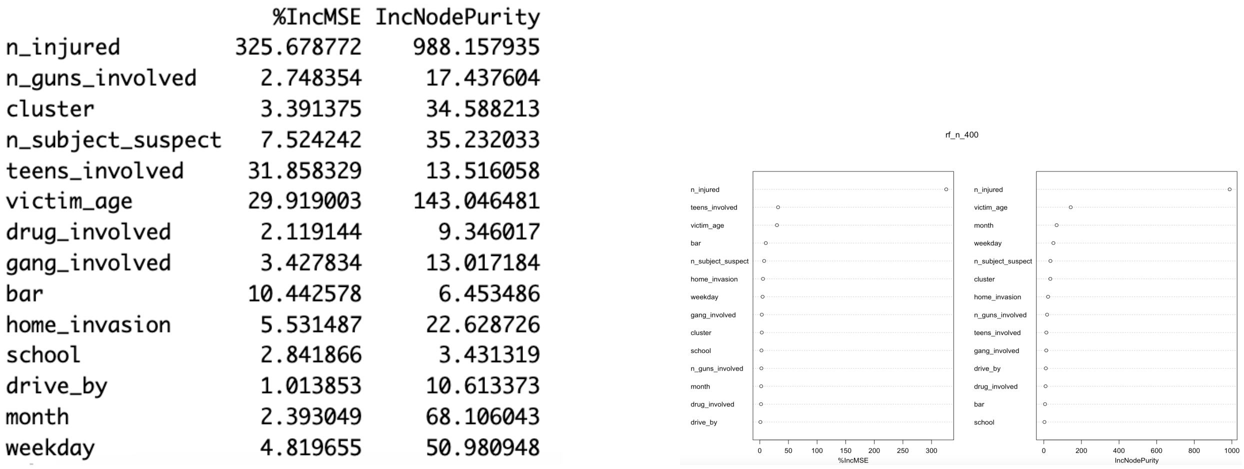
	0	1	Class.error
0	5300	89	0.01651512
1	365	1335	0.21470588

Table 2: Confusion Matrix of Feature Selected Classification Model



```
     |      Out-of-bag   |
Tree |      MSE  %Var(y) |
  50 |   0.1025    41.90 |
 100 |   0.1007    41.15 |
 150 |  0.09964    40.73 |
 200 |  0.09931    40.59 |
 250 |  0.09924    40.57 |
 300 |   0.0993    40.59 |
 350 |  0.09925    40.57 |
 400 |  0.09909    40.51 |
 450 |  0.09913    40.52 |
 500 |  0.09915    40.53 |
```

Figure 11: Optimal Number of Trees for Regression Model



```
                     %IncMSE IncNodePurity
n_injured         325.678772    988.157935
n_guns_involved     2.748354     17.437604
cluster             3.391375     34.588213
n_subject_suspect   7.524242     35.232033
teens_involved     31.858329     13.516058
victim_age         29.919003    143.046481
drug_involved       2.119144      9.346017
gang_involved       3.427834     13.017184
bar                10.442578      6.453486
home_invasion       5.531487     22.628726
school              2.841866      3.431319
drive_by            1.013853     10.613373
month               2.393049     68.106043
weekday             4.819655     50.980948
```

Figure 12: Random Forest Feature Importance Score for Regression model

	RMSE	MSE	MAE	R squared	OOB
All Features	0.2852933	0.08139227	0.1300527	0.6362989	0.09917
Selected Features	0.2838925	0.37449972	0.1354784	0.6409369	0.09921

Table 3: Model Performance Metrics-Regression Model