Analytics in Gun Violence: A Predictive Model for Gun Violence Incident Severity

1. Introduction

Gun violence, according to Amnesty International, is defined as "violence committed with firearms, such as handguns, shotguns, or semi-automatic rifles." This presents a threat to over 600 individuals daily who lose their lives due to firearms as "up to 71% of all homicides globally involve gun violence.[1]" In 2019, the United States placed second in terms of overall total gun deaths,[2] emphasizing the magnitude of gun-related issues and the need for comprehensive approaches to address public safety concerns.

This project therefore aims to classify gun violence incidents based on severity using a comprehensive dataset containing over 230k US incidents from 2013-2018. 'High' severity is identified when it is a mass shooting, that is, "3 or more killings in a single incident,"[3] 'medium' severity when the number of individuals killed is less than 3 but more or equal to 1, and 'low' severity when there are no deaths at all. This could potentially provide insight into identifying high-risk areas in need of more attention regarding targeted policy interventions and further help develop policies tailored to the key contributing characteristics of mass shootings. In summary, the goal of the project is to answer the following question: Can we effectively predict the severity of a gun violence incident in the United States based on a variety of features? Additionally, what key factors contribute to each incident severity level?

2. Data Description

The gun violence dataset contains 239,677 gun violence incidents spanning from 2013 until 2018, and 29 different variables characterizing them. From a visual perspective, as shown in the

[1] https://www.amnesty.org/en/what-we-do/arms-control/gun-violence/
[2] https://worldpopulationreview.com/country-rankings/gun-deaths-by-country
[3] https://www.britannica.com/topic/mass-shooting

first row in Figure 1, there has been an increasing trend in the number of incidents from 2013-2017 and more notably, a significant hike from 2013 to 2014 which has also been observed in external research, not signalling any anomaly in the data.[4] It is worth noting that one can observe a dip in the number of incidents in 2018, and this is mainly due to data collection only including events up until March 31st for this year, therefore, occurrences for 2018 have been dropped only for visualization purposes (see bottom of Figure 1). A more accurate representation of the monthly trends can now be seen, showing that July and August have the most incidents from 2013 to 2017, and Illinois having the most incidents relative to its peers (see Figure 2).

Furthermore, adding back 2018 incidents, due to the numerous variables in the dataset, a few were ultimately removed as they did not provide any relevant information for statistical analysis. Throughout the analysis, dropped columns include *incident_id*, *city_or_county*, *address*, *incident_url*, *source_url*, *incident_url_fields_missing*, *gun_type*, *incident_characteristics*, *location_description*, *notes*, *participant_age*, *participant_name*, *participant_status*, *participant_type*, and *sources*. While they do provide additional information on the incidents, it is rather difficult to apply any statistical techniques due to the nature and complexity of their data types. Additionally, *congressional_district*, *state_house_district*, and *state_senate_district* were dropped as this information is reflected in the *longitude* and *latitude* variables which were kept. Despite the correlation matrix only showing correlations of |0.23|, this information was ultimately deemed as repetitive since the district numbers are associated with the state.

Overall, the remaining columns include *date* (separated into *year* and *month*, dropping *day*), *n_killed* (later dropped as the information was incorporated for creating the y-variable), *n_injured*, *latitude*, *longitude*, *n_guns_involved*, *participant_age_group* (further processed to extract the

[4] https://www.pewresearch.org/short-reads/2023/04/26/what-the-data-says-about-gun-deaths-in-the-u-s/

main group involved such as ‘Adult’, ‘Teen’, ‘Child’, and ‘Mixed’, the latter meaning various age groups were involved), *participant_gender* (further binary encoded as to whether a female was involved or not), *participant_relationship* (further binary encoded as to whether the participants involved in the incident knew each other or not), and *gun_stolen* (further binary encoded as to whether any gun involved in the incident was stolen or not). As a last step, these predictors were evaluated with a correlation matrix to remove any collinear variables.

In summary, the final dataset contains one y-variable and 12 predictors. The dependent variable, *severity*, is a categorical variable containing three possible values: ‘low’, ‘medium’, and ‘high.’ The 12 independent variables include binary variables and dummy variables as well as numerical variables. These were chosen to reflect various facets that could affect mass shootings, such as geography (i.e. certain geographic locations might affect the probability of a mass shooting incident) and time (i.e. there appears to be an increasing trend in incidents, therefore there may be an overall increase in mass shootings during specific months). Additionally, characteristics of the incidents were incorporated such as information on the participants involved in the events (i.e. perhaps more mass shootings were targeted towards individuals in the perpetrators’ close relationship network, females, etc.) and the firearms used (i.e. having more stolen guns could potentially be a characteristic of mass shootings). The frequency of these characteristics can be observed in Figures 3, 4 and 5 to have a better understanding of the preprocessed data.

3. Model Selection & Methodology

Random Forest. A Random Forest was chosen to perform hyperparameter tuning to find the most relevant predictors for the subsequent models. This also helps with decreasing the likelihood of creating an overfitting model and keeping the variables that are most important for classifying gun violence incidents into ‘low’, ‘medium’, and ‘high’ severity cases.

Quadratic Discriminant Analysis (QDA). Quadratic Discriminant Analysis (QDA) is the chosen modelling approach for predicting the severity of gun violence incidents classified as 'low', 'medium', or 'high' severity. Since it has a multi-category outcome variable, QDA proves to be a suitable choice for this classification task. Unlike logistic regression, which is well-suited for binary outcomes, QDA applies to situations with more than two possible outcomes. The decision not to opt for logistic regression aligns with the categorical nature of the severity classification. QDA, based on Bayes' theorem and prior probabilities, provides one with an effective framework for estimating the probabilities of incidents falling into each severity category. Additionally, compared to the Linear Discriminant Analysis (LDA) model, QDA assumes that the standard deviation across different classes is likely to be different, which allows one to capture more complex relationships between the various predictors being used for the model. Overall, this approach allows one to explore how incident characteristics influence the likelihood of outcomes, providing a better understanding of the factors influencing the severity of gun violence incidents and contributing to a nuanced understanding of their diverse nature and impact.

Classification Tree. While the QDA builds a model around predictors one wants to examine for classifying incidents as 'low', 'medium', and 'high' severity, the classification tree improves overall user interpretability to understand the hierarchy and relationships between said variables. Unlike QDA, which relies on statistical methods, a classification tree presents a visual representation of decision rules, making it easier for individuals in general to understand the decision-making process. This graphical representation allows for a clear understanding of how each predictor contributes to the classification outcome and provides a transparent framework for supporting decisions. It is a rather simple and intuitive method, overall making it an effective tool for improving comprehensibility and user-friendliness in terms of interpretation.

4. Results

Random Forest. As seen in Table 1, the error rate of this model building 500 trees results in an error rate of 24.26%, which is relatively low. This means that the predictions of the model are decent. Additionally, by running a Random Forest model, one can evaluate the importance of the variables included in the model, which can further serve as a means for hyperparameter tuning to refine the classification tree (see Figure 6). Three predictors are noticeably insignificant in terms of their average decrease in the accuracy of the model: *month*, *participant_age_groupChild*, and *participant_age_groupMix.* They were therefore dropped when running the subsequent models. Also, since all participant age groups appear to be of less importance, with the only remaining one being the *participant_age_groupAdult*, the latter was dropped as well.

QDA. From the QDA output in Table 2, one can see that the prior probabilities of each severity class are 59.62%, 38.26%, and 2.12% for 'low', 'medium', and 'high' severity cases respectively. These prior probabilities indicate that, before incorporating the effect of the predictors on the model's classification probabilities, these are the initial probabilities of each incident happening. Furthermore, one can observe the average of each variable for each incident severity type. If one were to apply this to a case i.e. a gun violence incident in June 2023 in Chicago (41.8781° N, 87.6298° W) with 10 injured, 2 non-stolen guns involved, at least one female present and where all individuals involved somewhat know each other, the probabilities of this specific case being a high severity incident are 99.69%, 0.03% being a low severity incident, and close to 0% of being a medium severity case. This somewhat shows one the characteristics that increase the likelihood of there being a mass shooting with 3 or more individuals killed of high severity.

Classification Tree. To enhance one's understanding of the model, a classification tree was plotted using a complexity parameter (cp) of 0.007, resulting in a tree as seen in Figure 7. Finding

the optimal cp was attempted (see Figure 8), but this resulted in an optimal cp of 0.001 with an overfitted and uninterpretable tree as seen in Figure 9. Due to time constraints, the assumption of an optimal cp of 0.007 is used for analyzing the subsequent results. Based on this assumption, in terms of the most important factors affecting the severity of a gun violence incident, the most important one is the number of individuals injured during the event, followed by whether the participants in said incident knew each other prior (i.e. acquaintances, co-workers, family, friends, neighbours, and/or significant others). In terms of overall interpretation, highly severe incidents (i.e. mass shootings) do not dominate any of the nodes as they are relatively rarer. Paying attention to the seven terminal nodes, and more specifically, the nodes at both extremities, 45% of sample observations fall into 'low' severity incidents while 38% fall into 'medium' severity incidents. Out of those samples, in the extreme-left node, for example, 1% of the observations are classified as being 'low' severity when in fact they are 'high' severity, showing the error rates, but 87% of 'low' severity cases are correctly classified as 'low.' This same analysis can be applied to all nodes. More conclusively, if there are no injured individuals during an incident and the relationship between these individuals is known, the model predicts that it will be classified as a 'medium' severity case (i.e. at least one individual killed and at most 2 individuals) with a 65% likelihood. Another example of interpretation going through more nodes would be the following: if there are no injuries, the participants in the incident do not know each other, and the guns used have been stolen, this incident would be classified as 'low' with a likelihood of 83%, which overall makes sense intuitively. In sum, this model allows for a better understanding of the predictors affecting the classification of gun violence incidents, notably the number of people injured, the relationship between all participants, whether the gun was stolen or not, as well as the latitude of the occurrence.

5. Conclusion & Limitations

In conclusion, the classification model is the most insightful for understanding the QDA model for predicting the severity of gun violence incidents. By classifying incidents into 'low', 'medium', and 'high' severity categories, the model contributes to identifying high-risk areas and formulating targeted policy interventions. The Random Forest model, with an error rate of 24.26%, demonstrates decent predictive performance. Additionally, the importance analysis highlights key predictors, helping with refining subsequent models. For example, since a key factor involves the perpetrator previously knowing the individual, awareness programs aiming to educate individuals about recognizing warning signs of potential perpetrators within their close circle can be implemented. One can also observe that 'medium' severity cases are more likely in regions in the USA between 33° and 40° in latitude (i.e. the lower half of the USA including states like Alabama and New Mexico, to name a couple[5]). They could potentially benefit from implementing tighter gun laws and violence policy interventions. However, there are limitations to consider. The dataset covers past incidents from 2013 to 2018, which may not accurately represent current trends and dynamics. Dropping and modifying certain variables was mainly for simplification purposes, potentially overlooking relevant information. Furthermore, the optimal cp was not found which may impact the accuracy of the classification tree and would need to be examined moving forward. Despite these limitations, the project lays a foundation for understanding the factors influencing gun violence severity. Combining machine learning models like Random Forest and QDA with traditional statistical techniques like classification trees contributes to a nuanced understanding of incident characteristics. The findings can inform policymakers in developing effective strategies to mitigate the impact of gun violence and preserve the fundamental right to life.

[5] https://www.mapsofworld.com/lat_long/usa-lat-long.html

Appendix

Figure 1. *Temporal Trends of Gun Violence Incidents by Year and Month from 2013-2018 and 2013-2017.*

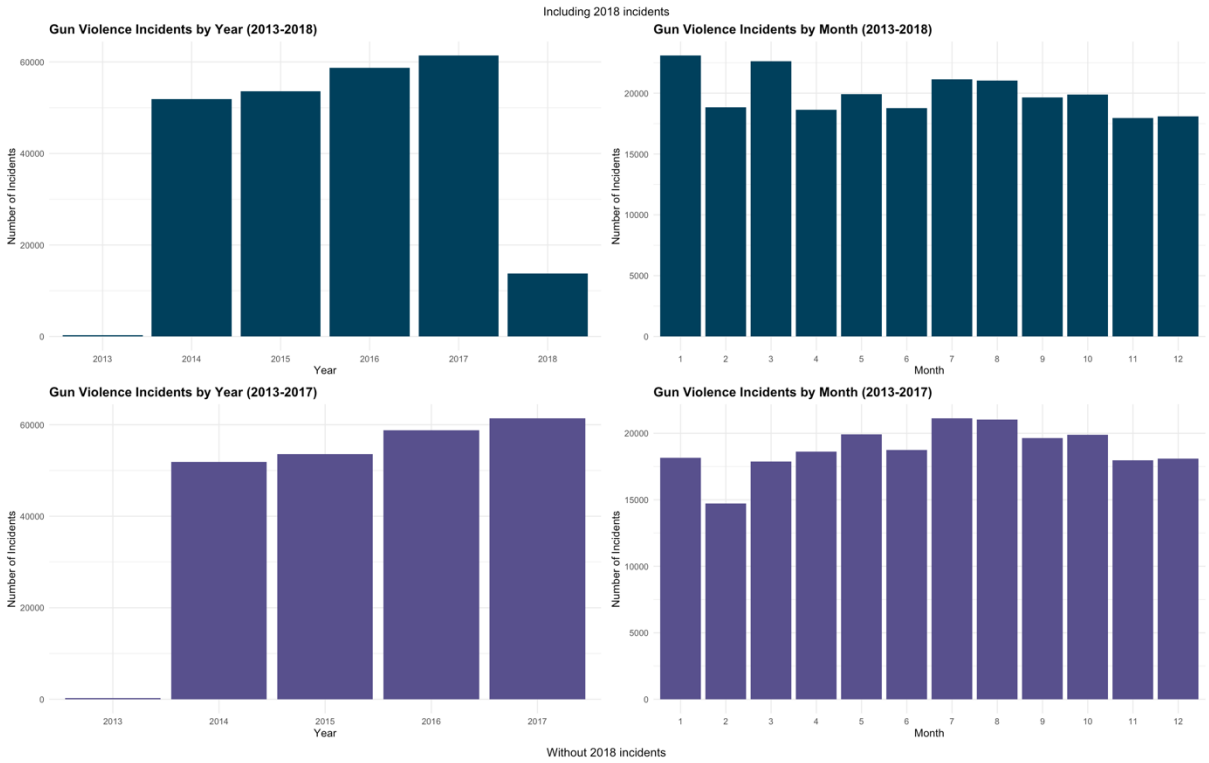

Figure 2. *Top 10 States with the Highest Number of Gun Violence Incidents from 2013-2017.*

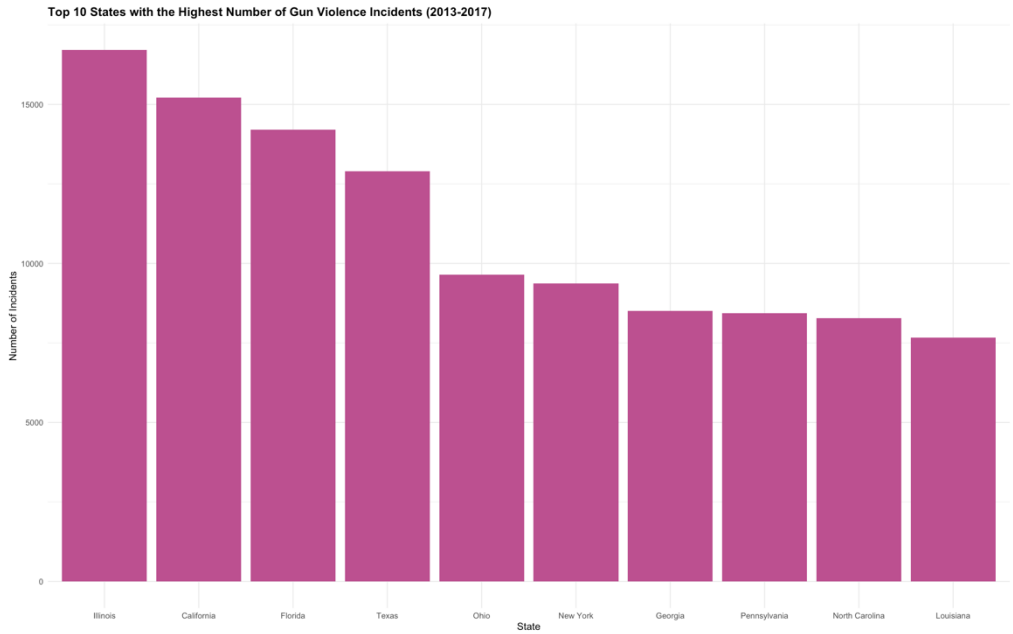

Figure 3. *Frequency of the Number of Incidents for Different Numbers of Individuals Killed and Injured from 2013-2018.*

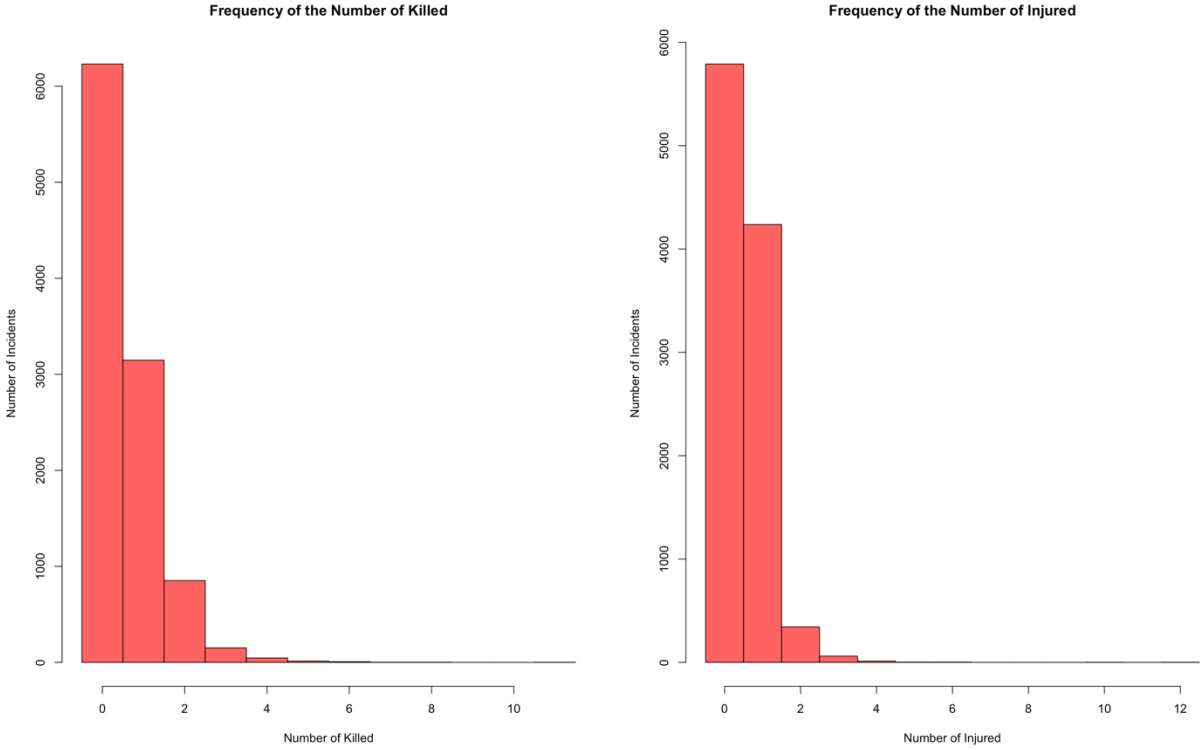

Figure 4. *Frequency of Each Age Group Category Involved in Incidents from 2013-2018.*

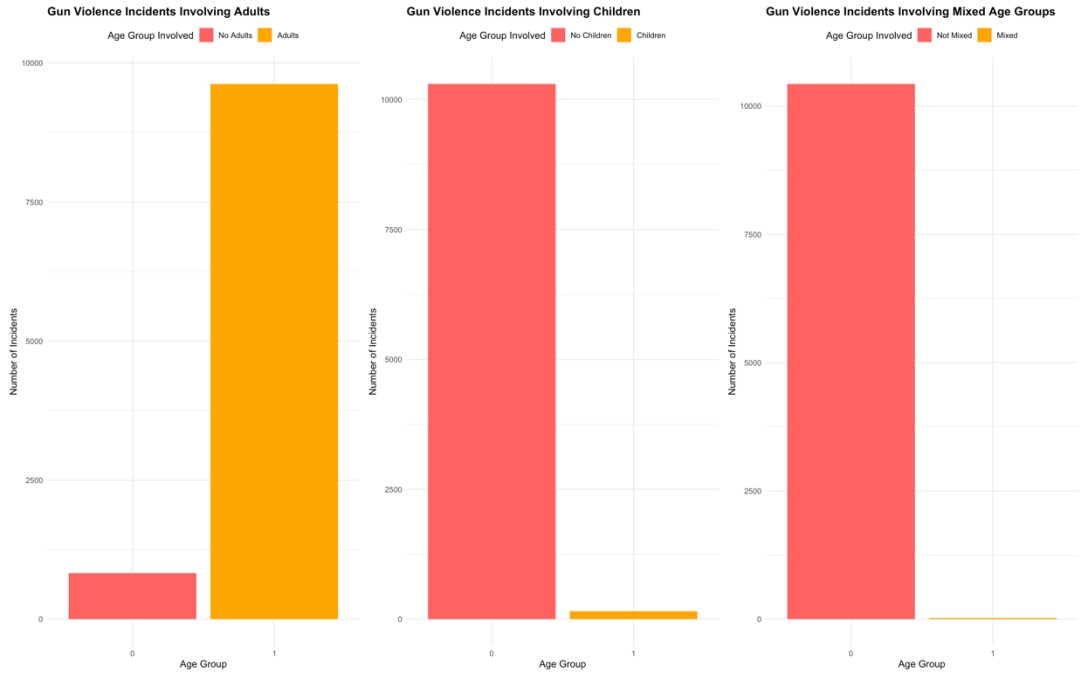

Figure 5. *Frequency of Other Model Variables (Gender, Relationship Status, and Gun Status) Involved in Incidents from 2013-2018.*

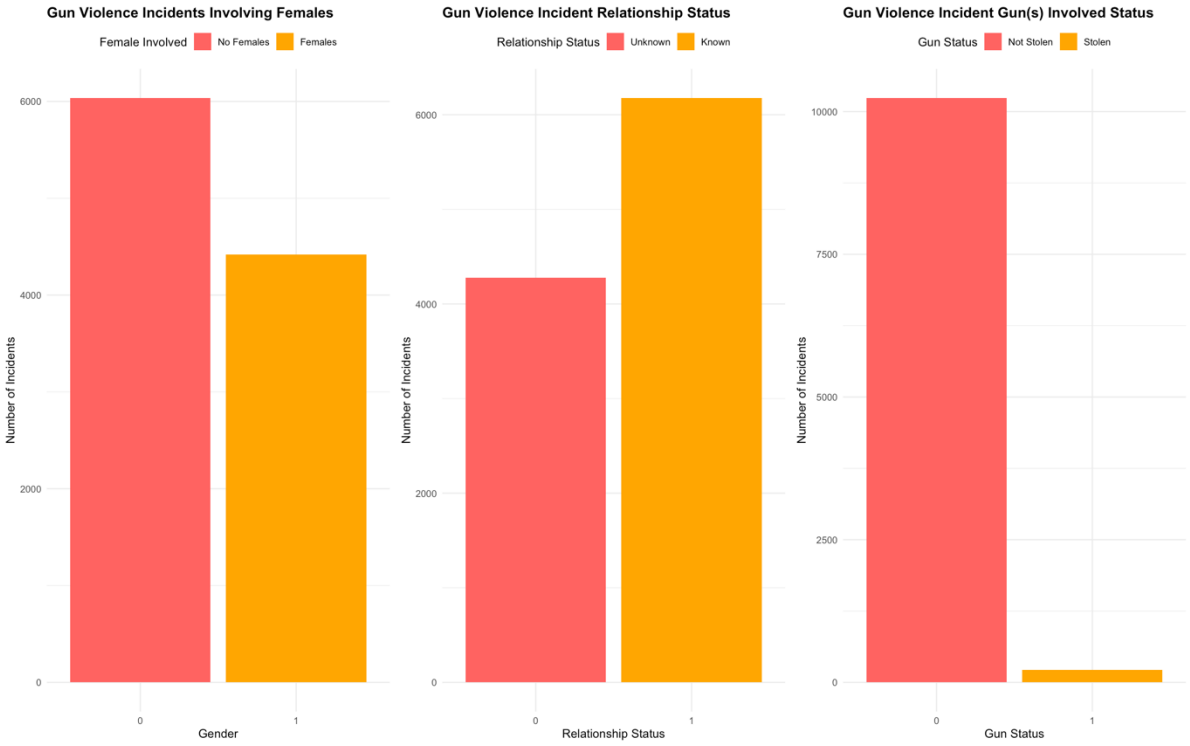

Figure 6. *Importance of Each Variable in the Random Forest Model.*

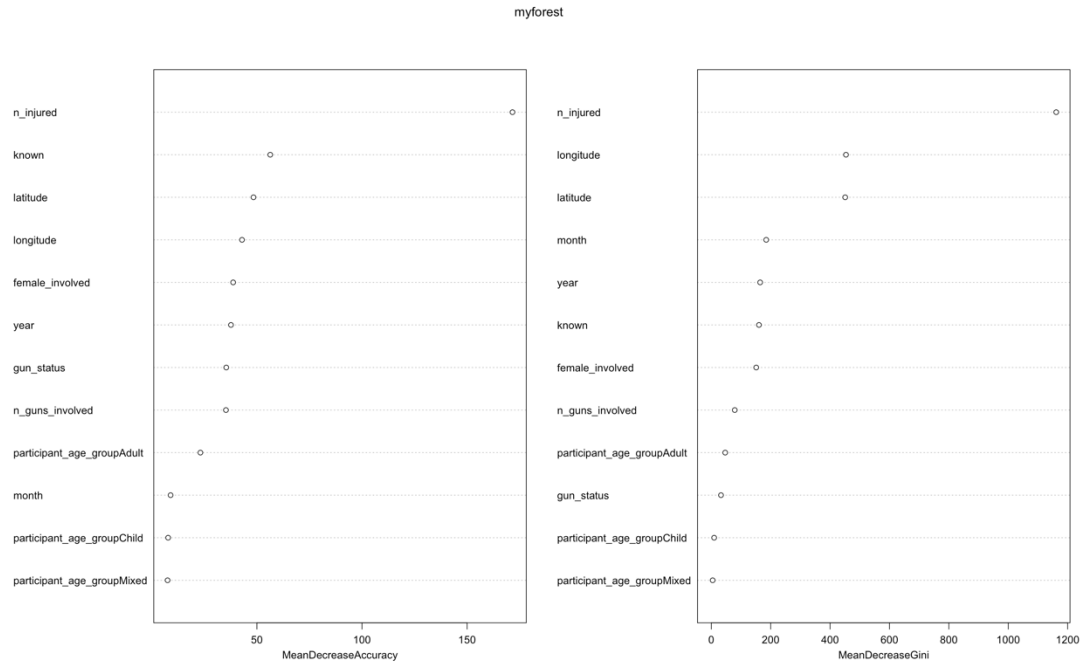

Table 1. *Random Forest Algorithm Results.*

```
Call:
 randomForest(formula = severity ~ year + month + n_injured +      latitude + longitude + n_guns_involved + participant_age_groupAdult
+      participant_age_groupChild + participant_age_groupMixed +      female_involved + known + gun_status, data = gun_violence,
ntree = 500, importance = TRUE, na.action = na.omit)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 24.26%
Confusion matrix:
       high  low medium class.error
high      0   59    163   1.0000000
low       0 4976   1255   0.2014123
medium    1 1058   2940   0.2648162
```

Table 2. *Quadratic Discriminant Analysis Results.*

```
Call:
qda(severity ~ year + n_injured + latitude + longitude + n_guns_involved +
    female_involved + known + gun_status)

Prior probabilities of groups:
      high        low     medium
0.02123995 0.59615385 0.38260620

Group means:
           year n_injured latitude longitude n_guns_involved female_involved     known  gun_status
high   2016.014 0.2702703 36.59304 -94.09207        1.373874       0.9279279 0.8873874 0.009009009
low    2016.298 0.7188252 37.30638 -91.36585        1.192264       0.3344567 0.4943027 0.028085380
medium 2016.287 0.1680420 36.60009 -92.81082        1.096524       0.5323831 0.7246812 0.010002501
```

Figure 7. *Classification Tree Using a Complexity Parameter (cp) of 0.007.*

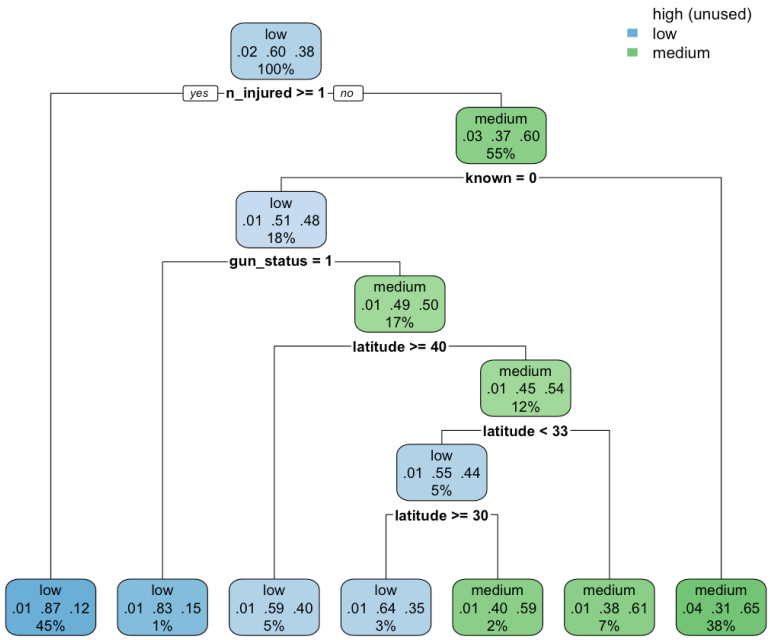

Figure 8. *Out-of-sample Performance Error Plot to Determine the Optimal Complexity Parameter (cp).*

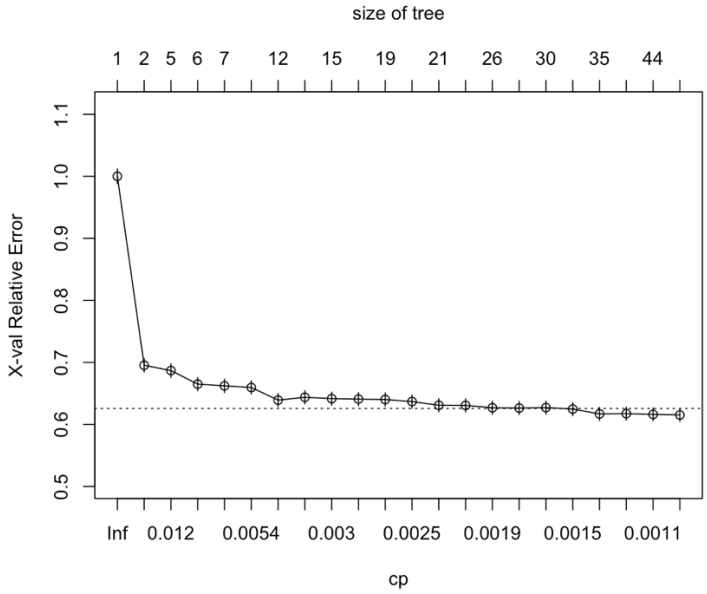

Figure 9. *Optimal Classification Tree Using the Optimal Complexity Parameter (cp) of 0.001 (Overfitted Tree).*