Machine Learning Analysis of Gun Violence

Aakash Akkineni *

September 2021

Abstract

In the U.S, gun violence is a substantial and divisive issue that legislators have spent decades trying to solve by passing various laws, restrictions, and bans using data analysis and research to support their ideas. This paper aims to use machine learning models such as logistic regression, random forest classifiers, gradient boosting classifiers, (Support Vector Machine) SVC classifiers, and Multi-layer Perceptron (MLP) classifiers to analyze different pieces of shooting incidents. Using the feature weights of the best-performing model, we aim to find which factors have the largest effect on shooting incident lethality - whether anybody died or not during the incident. Based on this, lawmakers can focus on the factors that most contribute to lethality. The first models in the paper aim to predict whether a shooting was lethal or not based on the age and gender of the shooter(s), state poverty rate, percentage of people in that state with only a high school education rate, shooting location, year, etc. The other models aim to predict the shooter's gender, the state that the shooting was in, and the political trifecta of that state based on similar factors. The precision, recall, and f-score of these models, as well as the feature weights of the random forest classifiers and gradient boosting classifiers, are reported. We used tree-based models to analyze the feature weights because of their high performance and because of Sk-learn's limitations in finding feature weights for other model types. Additionally, we include an analysis and possible reasoning for our results.

1 Introduction

The United States has one of the highest gun fatality rates across developed nations, with over 30,000 gun-related deaths annually [1]. As a result, gun violence has been a primary political and social issue for decades. One of the largest factors impeding effective legislation regarding gun homicides is a lack of knowledge on the specific issues that make gun-related incidents lethal or not. With more information on which specific components are most related to lethal gun violence, reducing gun deaths would be more feasible as politicians and activists could focus on the particular features that actually contribute to gun violence. Although there is an abundance of data on the thousands of gun-related incidents nationwide, effectively sorting through all the information and computing which factors are the most important would be extremely difficult and time-consuming. Machine learning can help to practically sort through these tens of thousands of incidents and find out which components of gun violence are most lethal.

Machine learning models are some of the most powerful data science tools ever built. Because of this, they can be used effectively to learn more about the specific factors that cause shootings to be lethal. Using different machine learning models, we can predict incident lethality based on features such as the gender and age of the shooter, the number of people injured in the encounter, the political trifecta of

*School for the Talented and Gifted, research mentor: Rida Assaf

the shooting incident's state, etc. Additionally, we can see which of these features the models we create found to be the most useful in predicting lethality. Logically, the factors that are the most useful are the ones that have the greatest influence on shooting lethality. For instance, if a model was able to accurately predict shooting lethality solely using the age of the shooter, it would make sense for legislators to focus on age when attempting to improve the United States gun violence situation.

Another way to tackle this problem using machine learning is to try and predict a categorical value, such as the state of a shooting incident, using the number of people killed as a feature in addition to the other features included in the data-set. We call these models feature prediction models, as these models aim to predict one of the features of our lethality prediction models, which are described above. An example of this would be training a model to predict whether a shooter is male or female based on all the other factors related to the shooting plus the number of people killed during that shooting. If the model were to accurately classify the shooters into males and females, that would prove that there is a predictable difference between male and female shooters. Using the feature importances of this model, we would also be able to see which specific features are different between the sexes. Alternatively, if the model was not able to accurately predict the shooter's sex, that would show that male and female shooters are largely similar.

Our goal is to use machine learning to help shed light on what the most important factors are that determine incident lethality. Using the feature weights for the best-performing model, we can see which features make the largest difference between lethal and non-lethal gun-related incidents. To do this, we will train multiple models to predict different features related to a shooting and analyze the feature importances of each. We will also use models to predict certain features, such as gender and state, using all the usual factors plus the number of people killed. These models are meant to shed light on whether certain groups or locations are predictably more violent or lethal than others. The last question we aim to answer with this study is which models are best for which purposes, as in what models are best for predicting lethality and what models are best for the feature prediction models.

2 Results

2.1 Prelude

In this section, we reported in a table the precision, recall, and f-score of every model we created. In a separate table, we showed the feature importances of the random forest classifier and gradient boosting classifier models for the features used in that subsection. Finally, we include a brief description of the results and important takeaways of that subsection.

The following is a concise explanation of precision, recall, and f-scores taken from towardsdata-science.com [2].

The precision is the ratio $tp/(tp + fp)$ where tp is the number of true positives and fp is the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

The recall is the ratio $tp/(tp + fn)$ where tp is the number of true positives and fn is the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where

an F-beta score reaches its best value at 1 and worst score at 0.

Notes: To account for any imbalances in the data, the f-score is weighted based on data imbalances. Also, RF = Random Forest Classifier and GB = Gradient Boosting Classifier in the tables and paragraphs below. Finally, in the subsection summaries below, the numbers in the parentheses are the relevant feature weights taken from the tables.

2.2 Single Shooter Model Results

Table 1: We present precision, recall, and f-score of the single-shooter machine learning models' abilities to predict lethality:

	Precision	Recall	F-score
Logistic Regression	0.71	0.71	0.71
Random Forest Classifier	0.75	0.75	0.75
Gradient Boosting Classifier	0.75	0.75	0.75
MLP Classifier	0.71	0.70	0.70
SVC Classifier	0.49	0.49	0.49

Table 2: We present feature importances of the random forest classifier and gradient boosting classifier models for the single-shooter section:

	state	city	lat.	long.	n_injured	congr_district	n_guns	state_house_district	state_senate_district	shooter_age	shooter_gender	year	pov_rate	high_school_education_only_rate
RF	0.03	0.06	0.11	0.11	0.24	0.04	0.10	0.07	0.06	0.09	0.02	0.03	0.03	0.02
GB	0.01	0.03	0.06	0.09	0.37	0.01	0.28	0.02	0.02	0.05	0.03	0.01	0.01	0.01

Judging by the f-scores in Table 1, the best performing model is the random forest classifier with a f-score of 0.75, followed closely by the gradient boosting classifier and logistic regression model. Using the information in Table 2, we can see that the random forest classifier model's strongest feature is the number of people injured (RF: 0.24), with the latitude, longitude, number of guns involved, and shooter age having weights near 0.10, implying a slight importance in those features. The other features had little to no weight. For the gradient boosting classifier, the number of people injured was even more important (GB: 0.37) and no other factor had substantial importance.

2.3 Multiple Shooter Model Results

Table 3: We present precision, recall, and f-score of the multiple-shooter machine learning models' abilities to predict lethality:

	Precision	Recall	F-score
Logistic Regression	0.84	0.84	0.84
Random Forest Classifier	0.87	0.86	0.86
Gradient Boosting Classifier	0.87	0.86	0.86
MLP Classifier	0.81	0.79	0.79
SVC Classifier	0.53	0.53	0.53

Table 4: We present feature importances of the random forest classifier and gradient boosting classifier models for the multiple-shooter section:

	state	city	lat.	long.	n_ injured	congr_ district	n_guns	state_ house_ district	state_ senate_ district	avg_ shooter_ age	male_ shooters	female_ shooters	n_ shooters	pov_ rate	high_ school_ education_ only
RF	0.01	0.03	0.05	0.05	0.39	0.02	0.03	0.03	0.03	0.04	0.09	0.02	0.19	0.01	0.01
GB	0.00	0.01	0.02	0.01	0.34	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.58	0.00	0.00

The best performing model by f-score in this case, as seen in Table 3, is just barely the random forest classifier, with a f-score of 0.86. All of the models, except for the SVC classifier, have about a 10% increase in performance over the single-shooter equivalents. The most important features, as seen in Table 4, for the best performing models, namely the random forest classifier and gradient boosting classifier, were the number of people injured (RF: 0.39/GB: 0.34) and the number of shooters (RF: 0.19/GB: 0.58). Other factors had little to no weight.

2.4 Shooter Gender Prediction Model Results

Table 5: We present precision, recall, and f-score of the shooter gender prediction models' abilities to predict the shooter's gender:

	Precision	Recall	F-score
Logistic Regression	0.58	0.58	0.58
Random Forest Classifier	0.58	0.58	0.58
Gradient Boosting Classifier	0.58	0.57	0.57
MLP Classifier	0.59	0.59	0.59
SVC Classifier	0.53	0.53	0.47

Table 6: We present feature importances of the random forest classifier and gradient boosting classifier models for the gender prediction section:

	n_injured	n_killed	n_guns_ involved	shooter_ age
RF	0.17	0.10	0.01	0.72
GB	0.38	0.13	0.01	0.48

As shown in Table 5, all of the models in this section performed very similarly, with f-scores just under 0.60. Judging by the feature importances of the random forest classifier and the gradient boosting classifier, which are outlined in Table 6, the most important features were the shooter age (RF: 0.72/GB: 0.48), number of people injured (RF: 0.17/GB: 0.38), and the number of people killed (RF: 0.10/GB: 0.13). The number of guns involved was irrelevant to both models (RF: 0.01/GB: 0.01).

2.5 State Prediction Model Results

Table 7: We present precision, recall, and f-score of the state prediction models' abilities to predict different states:

	Precision	Recall	F-score
Logistic Regression	0.10	0.24	0.12
Random Forest Classifier	0.06	0.09	0.07
Gradient Boosting Classifier	0.05	0.08	0.05
MLP Classifier	0.09	0.22	0.10
SVC Classifier	0.07	0.22	0.09

Table 8: We present feature importances of the random forest classifier and gradient boosting classifier models for the state prediction section:

	n_injured	n_killed	n_guns_involved	avg_shooter_age	n_shooters	male_shooters	female_shooters
RF	0.07	0.06	0.06	0.74	0.03	0.03	0.02
GB	0.09	0.10	0.08	0.42	0.03	0.16	0.12

Unlike the past few models, as shown in Table 7, the highest performing models are the logistic regression model, MLP classifier, and SVC classifier, all with f-scores around 0.10. Although these are the highest performing models, the model overall did poorly at predicting different states. Since we cannot find the feature weights for the other models using Sk-learn, we will analyze the random forest classifier and gradient boosting classifier's feature importances. Using the information found in Table 8, we can see that the most important feature according to the random forest classifier and gradient boosting classifier is the average shooter age (RF: 0.74/GB: 0.42). For the random forest classifier, none of the other features held much weight, but for the gradient boosting classifier, the number of male and female shooters held some importance, with weights of 0.16 and 0.12 respectively.

2.6 Political Trifecta Prediction Model Results

Table 9: We present precision, recall, and f-score of the political trifecta models' abilities to predict lethality:

	Precision	Recall	F-score
Logistic Regression	0.44	0.58	0.42
Random Forest Classifier	0.42	0.48	0.42
Gradient Boosting Classifier	0.46	0.50	0.42
MLP Classifier	0.48	1.0	0.65
SVC Classifier	0.35	0.27	0.26

Table 10: We present feature importances of the random forest classifier and gradient boosting classifier models for the political trifecta prediction section:

	n_injured	n_killed	n_guns_involved	avg_shooter_age	n_shooters	male_shooters	female_shooters	year
RF	0.06	0.06	0.05	0.64	0.03	0.03	0.01	0.11
GB	0.09	0.18	0.06	0.36	0.12	0.06	0.03	0.11

For this final model, judging by the f-scores seen in Table 9, the best performing model is the MLP Classifier with an f-score of 0.65. Conversely, the SVC classifier was the worst-performing model with an accuracy of 0.25. The remaining three models all performed similarly with f-scores around 0.40. Since we cannot find the permutation importances for the MLP classifier using Sk-learn, we will analyze the random forest classifier and gradient boosting classifier's feature importances. For the random forest classifier, the most important features are the average shooter age (RF: 0.64) and the year (RF: 0.11). For the gradient boosting classifier, the most important features are the average shooter age (GB: 0.36), the number of people killed (GB: 0.18), the number of shooters (GB: 0.12), and the year (0.11). The common thread between the two is the avg shooter age and the year.

2.7 Prediction Results Analysis

2.7.1 Single Shooter Model

With this initial model, our main goal was to answer the question of which features most contribute to the lethality of a shooting incident with only a single shooter. According to the most accurate models, which are the random forest classifier and the gradient boosting classifier, the strongest feature is the number of people injured in a shooting. This means that when looking at a shooting incident, the most vital factor that determines if the incident is lethal or not is the number of people that were injured during the shooting. The other feature that remained consistently high-weighted between the two high-performing models, although not as high-weighted as the number of people injured, is the number of guns involved in the shooting. Past this, every other feature had a comparatively low weight - except for the latitude and longitude, which implies that some locations are predictably more or less lethal than others. These irrelevant features include the state, city, congressional district, state house district, state senate district, state poverty rate, percentage of people with only a high school education in that state, shooter age, shooter gender, and year. According to our results, when a shooting happens, none of these features seem to matter when deciding whether that shooting will be lethal or not, including shooter gender, shooter age, state poverty level, and state education level. Since the models are only about 0.75 accurate according to the f-score, these factor weights are somewhat unreliable, but since the models performed with an f-score above 0.50, which is what the models would get if they were just guessing, there is substance to these results.

2.7.2 Multiple Shooter Model

The multiple shooter model aims to answer a similar question to the single shooter model - namely which factors contribute the most to incident lethality - but the data-set used to train this model includes shooting incidents with multiple shooters. Like the single shooter model, the random forest classifier and gradient boosting classifier were the most accurate machine learning algorithms. The single-shooter model and the multiple-shooter model are also similar in that the most important feature is the number of people injured in the shooting, which reinforces the idea that the number of people injured in a shooting has a large impact on the predicted lethality of a shooting. Unique to the multiple shooter model are the male shooters, female shooters, and number of shooters columns, which count the number of male, female, and total shooters involved in the shooting. The fact that the number of shooters column had a high weight, but the number of male and female shooter columns did not, implies that, although the total number of shooters does matter, the gender of those shooters does not. Interestingly, in the random forest classifier, despite still having a low weight, the male shooter factor had three times the weight of the female shooter factor, showing that the predicted lethality of shooting depends more on the number of male shooters involved than the number of female shooters involved. Overall, the performance is 10% better than the

single shooter model, which allows us to be more secure in this model's results, yet the f-scores are still less than 0.90. One drawback of the multiple shooter model is that, because many shooting incidents with multiple shooters have missing information on at least one of the shooters involved in the incident, many of the shooting incidents are unusable. Due to this, the training data used for the multiple shooter model is a third of the size of the single shooter model.

2.7.3 Gender Prediction Model

The question we aimed to answer with this model was how different are males and females in terms of gun violence. How pronounced is the difference between the two sexes, and what features might cause that potential difference? Starting with the f-scores of the models, all of the models had very similar results. If the models were to guess, the expected f-scores would be around 0.50, but since the models had f-scores nearer to 0.60, there must be a slight correlation between gender and the features used for these models. Judging by the random forest classifier and gradient boosting classifier feature weights, the age of the shooter is very important in predicting lethality, which implies that there is a correlation between the age of a shooter and their gender. The low weights of the other features, namely the number of people killed, the number of people injured and the number of guns involved, as well as the overall low f-scores, shows that male and female shooters are not much different when it comes to these factors. Even though almost 86% of shootings are caused by males, the fact that our models could not distinguish between male and female shooters proves that when females incite a gun shooting incident, they are not predictably distinct from male shooters [3].

2.7.4 State Prediction Model

The state model was built to answer the question of whether different states are distinguishable in terms of gun violence. Our best performing algorithms for this model based on f-score were the SVC and MLP classifiers. Although their f-scores are relatively low compared to other models in this study, with f-scores around 0.09, these classifiers performed four times as well as they would have if they were purely guessing. Therefore, due to the slight success of our models, our results imply that some states are predictably more or less violent, have younger shooters, have more male than female shooters, etc. than other states. It could alternatively imply that all states are slightly predictable in terms of shootings, or that the data-set we used was skewed towards some states. The lack of accuracy overall shows that states do not differ a large amount in terms of gun violence. The most important feature distinguishing different states' shootings is the average shooter age, which could convey that the age of shooters in some states is predictably different than the age of shooters in other states.

2.7.5 Political Trifecta

Because gun violence is such a politically divisive issue between the two main US parties - and therefore also divisive between different states of different parties - we wanted to answer the question of whether or not a machine learning model can predict a state's political trifecta based on factors related to gun violence [4]. Using the feature weights of the random forest classifier and the gradient boosting classifier, the two most important features are the average shooter age and the year of the shooting. The high weight associated with the average shooting age implies that different places with different political trifectas have differently aged shooters. The importance of the year shows that from 2013 to 2018, different political trifectas change in terms of gun violence in different years. The best-performing

model, which is the MLP classifier, had an f-score accuracy of about 0.65. Since there are three political trifectas, the accuracy would be around 0.33 if the model were just guessing, so there is a predictable difference between gun violence in states with different political trifectas, although the feature importance analysis above might be inaccurate if the MLP classifier's permutation importances for do not line up with the feature importances of the gradient boosting classifier and random forest classifiers. Additionally, the 1.0 recall of the MLP classifier could possibly show that the MLP classifier's results are unreliable.

3 Methods and Materials

3.1 Generally Applicable Information

The information and methods described in this section are applicable to all the models we trained using this data-set. The data-set we used is a compilation of data on over 230,000 different gun-related incidents from 2013 to 2018 built by James Ko and posted on the machine learning social platform Kaggle [5]. It combines data from a variety of smaller gun-violence data-sets collected from the Gun Violence Archive's website. The Gun Violence Archive (GVA) is a non-profit organization that aims to improve public access to information on gun violence in the United States. The data-set includes demographic information like the gender, age, relationship status, and name of many of the victims and shooters/suspects, as well as locational details like the address, latitude, longitude, state, city, type of building (i.e., club, school, etc.,), senate district, and the house of representatives district of the area each shooting took place in. Other incident-specific factors include the date, number of people killed, number of people injured, type of gun, whether the gun was stolen or not, miscellaneous notes on the shooting, and a brief description of the incident characteristics. Using 11 of the almost 30 available factors, we started processing the data into a format fit for five classifying machine learning algorithms including logistic regression, random forests, SVC classifiers, MLP classifiers, and gradient boosting classifiers. Once the data was modified, we used Sk-learn libraries in Python to create the models. In order to use categorical factors such as the city, state, and shooter gender, we imported OrdinalEncoder and assigned a numerical value to each value in these columns. We included only the year of every shooting when training the model with the date of the shooting as a feature. For the city and county column, we iterated through each incident and deleted rows that included the county instead of the city. When training the model, we split the data into training and testing data using an 80-20 split. These results were averaged with a "weighted" parameter.

We used GridSearch to fine tune the model's hyperparameters. GridSearch allows us to cycle through every possible hyperparameter combination within the arrangement we set. For the logistic regression models, we edited only the regularization parameter and the max iterations. For the random forest classifier, we changed the minimum samples per leaf, the minimum samples per split, and the number of estimators. For the gradient boosting classifier, we edited the same parameters plus the learning rate. Our SVC classifier's GridSearch algorithm cycles through different values for the regularization, kernel, and max iteration parameters. For the MLP classifier, our grid search algorithm modifies its alpha value, hidden layer size, and learning rate.

3.2 Lethality Prediction Models

3.2.1 Single Shooter Model

We used machine learning model to classify shootings as lethal or non-lethal. We built several versions of this model, the first of which we called the single-shooter model. The single-shooter model is named as such because every shooting included in the data-set for this model only included one shooter. When preparing the data-set for this model, we did not include the factors that either hurt the model's accuracy, had inconsistent formatting, or were irrelevant to the lethality, such as the incident URL column. The columns we used for the single-shooter model were the state, city, latitude, longitude, number of people injured, congressional district, number of guns involved, state house district, state senate district, year, shooter gender, shooter age, poverty rate, and percentage of people with only a high school education. The year, shooter gender, and shooter age are all columns we added to the data-set and manually filled in by iterating through each incident and taking what we needed. To fill in the poverty rate and percentage of people with only a high school education columns, we took state data from the USDA and filled in those columns according to the state of the shooting incident [6;7]. We dropped rows/incidents that held empty values for any of these columns, excluding columns we added, before iterating through the rows. We then also dropped null values in the columns we created, such as shooter gender and shooter age, after iterating through the rows and filling in those columns. Additionally, we ran through the participant type value for every remaining row and counted the number of "shooter-suspects" type participants in that row. If the row had more than one shooter-suspect, we deleted the row, which made sure we only included shooting incidents with one shooter. We also edited the number of people killed column so that any shooting that had one or more deaths involved had a value of 1, indicating that the shooting was lethal. Otherwise, we let the value of the number killed remain zero. After modifying both the training factors and the prediction factors, we trained the model on the five machine learning algorithms mentioned above.

3.2.2 Multiple Shooter Model

Our other lethality prediction model includes rows with multiple shooters. Instead of just taking one shooter's gender and age, we replaced the shooter gender column with two new columns which count the number of male and female shooters respectively and exchanged the regular shooter age column with the average shooter age column. Through the participant type value in each row, we added every shooter-suspect id to an array called ids. We used that array to count the male and female shooter-suspects from the participant gender column and added those values to their respective columns. We also counted the number of shooter ids present in the participant type column and inputted those values into a new column/factor that counted the total number of shooters. Using values from this column, we were able to calculate the average shooter age by dividing the total age, which was found by going through the participant age value and adding all the shooter's ages, by the total number of shooters. To make sure our data was accurate, if the sum of the male and female shooters column or the number of shooter ages present in the participant age column was not equal to the number of shooters, we deleted that row. We filled in the poverty rate and percentage of people with only a high school education rate columns the same way as with the single shooter model. In addition to those two features, the features we used with the five machine learning algorithms listed above were the state, city, number of people injured, congressional district, number of guns involved, latitude, longitude, state house district, state senate district, average shooter age, number of male and female shooters, and the total number of shooters. We dropped features with any null values in these factors, excluding columns we added, before iterating through the rows. Similar to what we did before, after iterating through the rows and filling in the new columns, we dropped columns with null values in the columns we added.

3.3 Feature Prediction Models

3.3.1 Shooter Gender Model

Unlike the previous models, these next few models use the number of people killed as a feature and predict a certain feature or column. For example, the first model we built of this type is meant to predict the gender of the shooter based on the number of people injured, the number of guns involved in the shooting, the number of people killed, the year, and the shooter age. It is based on the single-shooter model from section 2.2 in the way that it also drops rows with multiple shooters and fills in the newly added shooter gender and shooter age columns in the same way as does the single shooter model. Specifically, it does this by finding the shooter id from participant type and cross-referencing that with the participant age and participant gender columns. Like the single-shooter model, If there are multiple shooter-suspects ids in the participant type column the row is dropped. Unlike the single shooter model, instead of changing the number of people killed to 0s and 1s to label shootings as purely lethal or non-lethal, the number of people killed was kept as it was in the original data-set. Also, fewer, more gun violence-related features were used to predict the shooter's gender. Finally, since the training data after deleting null values was extremely imbalanced towards men, with about 10,000 men to 1,500 women, we deleted all but around 1500 of the rows with men to balance the data. Using these features, we trained five models with the different machine learning algorithms outlined in 2.1 to classify shooters as male or female.

3.3.2 State Prediction Model

We built the next model of this type to predict the state of every usable incident. We first unedited the initial data-set to have the original values for the number killed column. Using the multiple-shooter model as a base due to its higher accuracy, we added four columns to the data-set for the shooter's average age, the number of male shooters, the number of female shooters, and the total number of shooters. We discarded most non-gun or shooter-related features, as we did in the gender prediction model, using only the number of people injured, the number of people killed, the number of guns involved, the average shooter age, the total number of shooters, the number of male shooters, and the number of female shooters as factors. After this, we dropped null values from the features we planned to use and assigned numeric values to the states. When the data-set was ready, we trained both models on the same five algorithms as before to predict the state of an incident on the number of people injured, the number of people killed, the number of guns involved, the average shooter age, the number of shooters, the number of male shooters, and the number of female shooters.

3.3.3 Political Trifecta Model

We created the last model to predict whether or not a gun-related incident happened in a democratic, republican, or divided state. The way we measured a state's political association is through their government trifecta, found using Ballotpedia [8]. A political trifecta is when a single political party has control of the executive branch and both branches of a bicameral legislature. For example, a state has a democratic trifecta if the governor, statehouse, and state senate are controlled by the democratic party. If the three positions are occupied by different parties, the state has a divided trifecta. Other modifications we made were keeping the number killed column's original values and adding and filling in the year, number of male and female shooters, total number of shooters, and average shooter age columns like we did in the multiple shooter model.

We also dropped null values in any of the features we added and took out features not specifically related to gun violence or shooter demographics, choosing to train the model based on the number of people injured, the number of people killed, the number of guns involved, the average shooter age, the total number of shooters, the number of male/female shooters, and the year with the same five classifying algorithms.

4 Discussion

4.1 General

We used machine learning algorithms to investigate feature importances and found which features contribute the most to lethality and whether certain features/groups are distinct enough to be classified apart based on shooting-related features, the results of which are detailed in section 2. Our secondary question was also answered: what model is best for which task when it comes to applying machine learning to gun violence. What we've found is that logistic regression, random forests, and gradient boosting classifiers are the most accurate when it comes to predicting lethality. However, for feature prediction models like the shooter gender model, MLP classifiers usually perform slightly better. The largest obstacle in this process has been missing information. Because of this missing information, many of the columns, such as the type of gun involved and the participant relationship factor, had too much missing data to effectively use. Even in the columns we used, there is a lot of missing data, which made preparing the data-set a challenge. It also meant that we had to cut thousands of shooting incidents from the training data. Additionally, in order to make the data usable, we had to translate the strange formatting of the spreadsheet into something a computer can read through. For example, we were forced to numericize the categorical variables such as the state and city of the shooting. Another issue we had is that, at first, we aimed to predict the number of people that died in a certain shooting, rather than just classifying incidents as lethal or non-lethal. The problem with this is that the vast majority of the remaining incidents, after eliminating rows with null values in any of the stated factors, had only 1 or 0 kills, which hurt the accuracy of the model since it is difficult for machine learning models to predict outliers [9]. To remedy this we asked a different question: using this data, can we predict whether a shooting is lethal or not? This is why, for the single and multiple shooter models, we modified the number of people killed column to only have 1s and 0s, representing lethal and non-lethal shooting incidents. Finally, because of unequal sizes for different columns within data-sets, we used a weighted f-score to get an accurate performance report.

4.2 Further Research

Although we have explored many questions and found many results, with this data-set, there is still lots to explore.

One possible avenue of further research could be exploring how important the socioeconomic status of the city is in predicting lethality. This is possible since every shooting incident has the city as a data-point. In order to do this, however, you would need some sort of data-set with every city's socioeconomic status. You could also do the same thing with a city's educational status, police presence, etc., assuming you had the data.

Another way to continue this research would be to fill in the missing information directly from the gun violence archive. This would allow you to train on more data with more features, which would create more accurate models. You could alternatively normalize the data-set's formatting, which allows for the same thing - more training data.

5 Citations

[1] Stark, David E., and Nigam H. Shah. "Funding and Publication of Research on Gun Violence and Other Leading Causes of Death." JAMA, vol. 317, no. 1, Jan. 2017, pp. 84–85. Silverchair, https://doi.org/10.1001/jama.2016.16215.

[2] Ghoneim, Salma. "Accuracy, Recall, Precision, F-Score amp; Specificity, Which to Optimize on?" Medium, Towards Data Science, 8 Apr. 2019, https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124.

[3] Firearm Violence Prevention —Violence Prevention—Injury Center—CDC. 5 May 2021, https://www.cdc.gov/

[4] Amid a Series of Mass Shootings in the U.S., Gun Policy Remains Deeply Divisive." Pew Re-search Center - U.S. Politics Policy, 20 Apr. 2021, https://www.pewresearch.org/politics/2021/04/20/amid-a-series-of-mass-shootings-in-the-u-s-gun-policy-remains-deeply-divisive/.

[5] Gun Violence Data. https://kaggle.com/jameslko/gun-violence-data.

[6] "Poverty." USDA ERS - Data Products, https://data.ers.usda.gov/reports.aspx?ID=17826.

[7] "Education." USDA ERS - Data Products, https://data.ers.usda.gov/reports.aspx?ID=17829.

[8] "State Government Trifectas." Ballotpedia, https://ballotpedia.org/State government trifectas. Accessed 19 Sept. 2021.

[9] Swalin, Alvira. "How to Make Your Machine Learning Models Robust to Outliers." Medium, Heartbeat, 24 Sept. 2021, https://heartbeat.comet.ml/how-to-make-your-machine-learning-models-robust-to-outliers-44d404067d07.

Link to Program Code:
https://colab.research.google.com/drive/1atGpdnh8kFeLVTplfyAAhT5x7XTnbdoZ?usp=sharing