Northeastern University
Khoury College of
Computer Sciences

Professor Benjamin Nye

Analyzing Mass Shootings

Final Report Paper

Ryan Liang

Spring 2021
4/29/2021

Class Section: TTh 11:45AM - 1:25PM
360 Huntington Ave., Northeastern University, Boston, MA 02115

Analyzing Mass Shootings

DS4420 Machine Learning Final Project

TABLE OF CONTENTS

I. Abstract

The following report is an analysis of major mass shootings reported over the past five decades. The data provided contains information about both the victims and shooters of these mass shootings, as well as their location, summaries, and other descriptives. After some data processing and cleaning, several machine learning models were run on the data for different tasks in relation to the goal of deriving insights and conclusions to aid in mass shooting research.

Some important conclusions drawn from the project involve spending more time and research studying mental health, figuring out what leads to deteriorating mental health, and providing more help to those suffering from it. Additionally, it was found that an overwhelming majority of mass shooters tend to be male rather than female, and therefore research should be conducted on why this case may be and how it can be helped. Other actionable insights into mitigating mass shootings can be found within the report.

II. Introduction

Problem Description

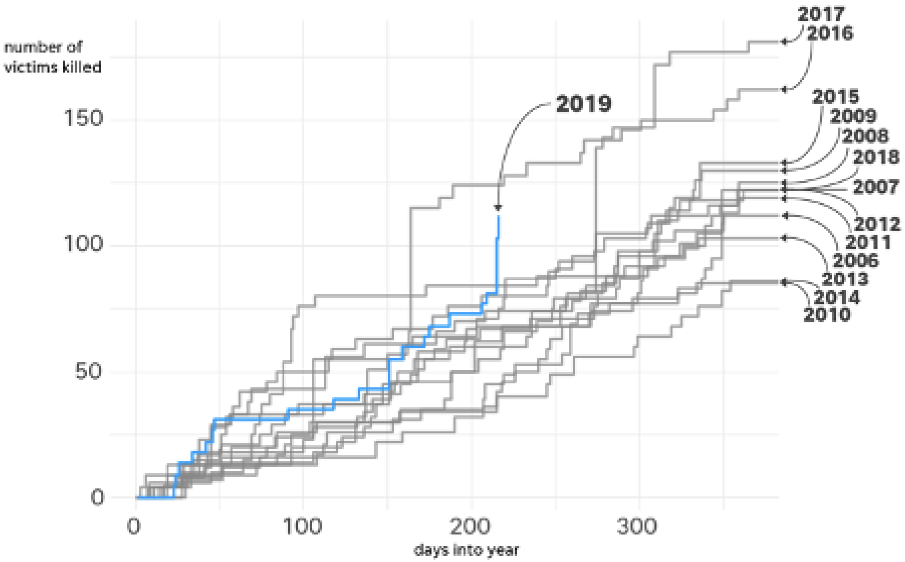
There are two approaches to the problem at hand regarding how machine learning will be conducted. The plan is to process the problem mainly as a classification problem with a later attempt at regression.

Classification: The machine learning models will take in the variables that indicate the shooter's race, gender, and mental health status, and classify the data into each category based on these variables.

****Regression**: A potential regression task can attempt to predict the total number of victims based on the variables being used (mentioned above)

Motivation

Mass shootings in the U.S. are a serious problem, with 2019 showing record high numbers of both fatalities and shootings in just the first half of the year (as pictured in the graph). As we progress further into 2021, the effort to reduce the number of mass shootings has been aided negligibly, and almost nothing is getting better. The goal of the project will be developing key insights about the perpetrators of mass shootings, and mainly analyzing the mental health, race, and gender of the killers. In the future, the location, motivation, and any other important information can be used to analyze data even further. When tackling a project of this magnitude, it's important to keep in mind the sensitivity of it, and how important it is to not to draw generalizations and conclusions about race and/or gender, as that is what fuels stereotypes and leads to racial profiling, an extremely dangerous threat to society. However, it is also important to study and analyze key patterns in order to get a full comprehensive picture and to understand the problem better, especially when the issue is one of this caliber, and especially when enough resources haven't been allocated to studying the problem in the first place.



III. Dataset

There is one major dataset that I'm using from Kaggle titled MassShootingsDatasetVer5.csv. In it, we have information about the shooter, their race, their age, and other general demographics. We also have several columns including location, description, summary, fatalities, injured, and number of total victims. We encounter a problem when we want to run several machine learning models on the dataset, which happens to contain a lot of string variables such as race.

The way I dealt with this problem was by encoding those several variables as numerical values. For example, the variable "Mental Health Issues" was a column in our dataset that indicated whether the shooter suffered from Mental Health Issues. This was previously encoded as a variable with values "Yes", "No", "Unknown", etc., which is not translated well when running through a machine learning model. So in the form of preprocessing, I created a new list that would contain all the entries, but this time decoded the variable into numerical values, and then iterated through the Mental Health Issues column. If the variable at the current entry was "Yes", I would encode that entry as "1", and for "No" it would be decoded as "0". I did this for the entire column, and I decoded/preprocessed other variables similarly such as "Race" and "Gender". Both Mental Health and Gender ended up being binary variables (save for the "Unknown / 2" variable)

Another slight issue with the dataset is the volume of records. The dataset was not rich enough to effectively split up into training and testing, and to subsequently allow the data to be trained well enough to reach optimal performance. In the future, a richer and denser dataset would be ideal.

For this project, it was important that the data be cleaned AND preprocessed/normalized properly in order to measure the performance of the machine learning models. I used the built-in StandardScaler() model to accomplish this task.

Finally, I am splitting up the project by creating different tasks/problems for me to solve. The way I have split up this project is by the following tasks:

1) Predicting the gender of the shooter (***Classification***)
2) Predicting the mental health status of the shooter (***Classification***)
3) Predicting the race of the shooter (***Classification***)
4) Predicting the total # of victims of a mass shooting (***Regression***)

And I split the data up this way:

1. ***Training/Testing data*** (easily changeable)
    a. This was done using the built-in **train_test_split()** method from sci-kit learn
2. ***Features/Labels***
    a. i.e., Problem #1 is split up as:
        i. *Features* = Race, Mental Health Issues, Total victims
        ii. *Label* = Gender

IV. Methods

The main methods that I have chosen for this project are running several classification models from scikit-learn and measuring their performance. The models that I have chosen are:

- Naive Bayes
- Logistic regression
- kNN
- Decision Tree
- SVM (Support Vector Machine)

And for Regression:

- Linear Regression

After running these models through the training data features and labels and then testing them on the testing data, analysis is done through several common metrics. So, before jumping into analysis, the metrics I had decided to use are:

- Accuracy
- Error
- Precision
- Recall
- ...which culminates into a Confusion Matrix
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

I also decided to measure ***feature importance*** for the specific task at hand by producing the coefficient matrix (for machine learning models that would allow it).

V. Results

***Some of the results are based on different compilations of the code, but all results are similar.*

The problem/tasks are split up in the following ways:

- Task/problem
- Problem type
- Context
- Models used
- Results
- Analysis

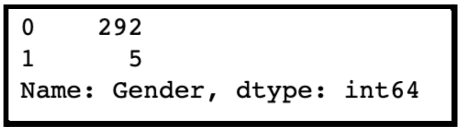
Task/Problem #1: *Predicting the* ***gender*** *of the shooter*

Problem type: *Classification*

Models used:

- Naive Bayes
- Logistic Regression

Context: the data contains about 98% male shooters (**292/297**) and 2% female shooters (**5/297**)**



```
0    292
1      5
Name: Gender, dtype: int64
```

Results:

Model Performance:

Model	***Data*** (% = size)	***Accuracy***	***Error***
Naive Bayes	Training (*35%*)	**~96%,** (0.9611650485436893)	~4%, **(**0.03883495145631066)
	Testing (*65%*)	**~97%,** **(**0.9663461538461539)	~3%, **(**0.033653846153846145)
Logistic Regression	Training & Testing	**~98%,** **(**0.979381443298969)	~2%, (0.020618556701030966)

Feature Importance (for Logistic Regression):

Feature	***Coefficient***
Race	-0.633614
Mental Health Issues	-0.476776
Total Victims	-0.151343

Analysis:

Both classification models were able to predict the gender of the shooter with very high accuracy (typically over ~95%). Additionally, after interaction with the data, it can be seen that relatively high portions of the data show that the shooter's gender tends to be male.

Initially, other metrics were included to measure the performance of the models, including precision, recall, F1 score, and AUC/ROC scores. However, for a problem like this, there was an **overwhelming amount of** data that ended up being *male* (only five records were female shooters), and so even when the models assigned a label of *female* to unknown data (on the test set), the false positive rate (FPR)/true positive rate (TPR) would not be a good way to measure the model because there would be very few instances where a TPR or True Positive would not be 0, resulting in both division by 0 errors and a 0 for precision, recall, and F1 score, etc. What I mean by this is that let's say that the five records where the shooter was female was in the held out test data set. If the model(s) predicted that five records in the test data to be *male* rather than *female* (a likely scenario, considering the training data would be all male), then the TPR/FPR would end up being 0 because we never predicted the female records accurately.

In a *feature importance analysis*, the prediction of the gender of the shooter places highest importance on the factor of **Race**, which has the largest coefficient of -0.633614.

Considering the very high amount of male shooters and the comparably high accuracy of the models, it can be said that the models don't effectively learn a good amount of information and are not being trained well on the data for this task.

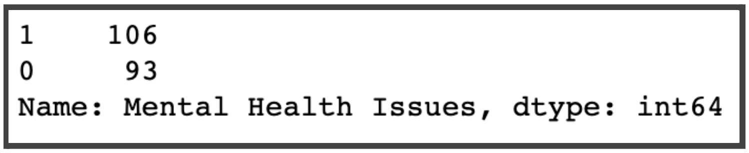
Task/Problem #2: *Predicting the* ***Mental Health Status*** *of the shooter*

Problem type: *Classification*

Models used:

- Naive Bayes
- Logistic Regression
- kNN (k-Nearest Neighbors)

Context: the data contains about **53%** shooters with mental health issues (**106/199**) and **47%** shooters that don't (**93/199**)**



```
1    106
0     93
Name: Mental Health Issues, dtype: int64
```

Results:

Model Performance:

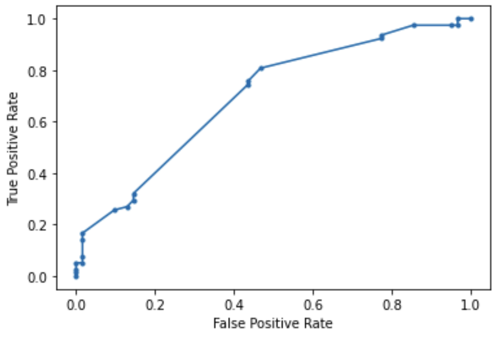
Model	***Data*** (% = size)	***Accuracy***	***Error***
Naive Bayes	Training (*30%*)	**~66%,** (0.6610169491525424)	~34%, (0.3389830508474576)
	Testing (*70%*)	**~61%,** (0.6142857142857143)	~39%, (0.3857142857142857)
Logistic Regression	Training & Testing	**~56%,** (0.5642857142857143)	~44%, (0.4357142857142857)

kNN (*k* = *9*)	Training & Testing	**~54%,** (0.5428571428571428)	~46%, (0.4571428571428572)

Feature Importance (for Logistic Regression):

Feature	***Coefficient***
Race	-0.005854
Gender	-0.370167
Total Victims	0.908748

Additional Metrics:

Model	***Metric***
Logistic Regression	Precision = **~74%,** (0.7428571428571429)
	Recall = **~33%,** (0.3333333333333333)
	F1 Score = **~46%,** (0.4601769911504425)
	AUC = **~70%,** (0.6950992555831265) 

Analysis:

The classification models were able to predict the Mental Health Status of the shooter with less accuracy then gender, but not with negligible results. Our models performed with about ~55% - ~65% accuracy. What this tells us is that we are able to determine whether a shooter is suffering from mental health issues (based on the given variables) with moderate accuracy, and therefore it is not a factor to rule out. Based on an initial exploration of the data, more often than not, the shooter is suffering from some type of mental health issue. This is important because this means that mental health care is a sector that we should invest more resources into, as it is determined to be one of the factors that can lead to mass shootings and gun violence.

The kNN model was run through multiple iterations, with different sizes *k* to ensure that we were using the optimal size to maximize performance. The accuracy and error of each *k* value was recorded and reported, and then the maximum accuracy was stored and used. In the Naive Bayes model, the AUC/ROC score was used to measure the FPR and TPR for the model. The AUC is an important metric to plot true and false positive rates and further back up the performance of the model. An AUC score close to 1 is an ideal/optimal rate, and a 0.5 score indicates a random classifier. Our model scored ~0.7, which is a considerably good AUC score. This backs up the decent performance of the Naive Bayes for this specific task. Finally, for Logistic Regression, there were >~5 instances of either label/class (unlike gender), and so we were able to report our precision, recall, and F1 score. All of these metrics are reported above under *Results*.

In a *feature importance analysis*, the prediction of the mental health status of the shooter places highest importance on the factor of **Total Victims**, which has the largest coefficient of 0.908748.

Task/Problem #3: *Predicting the **race** of the shooter*

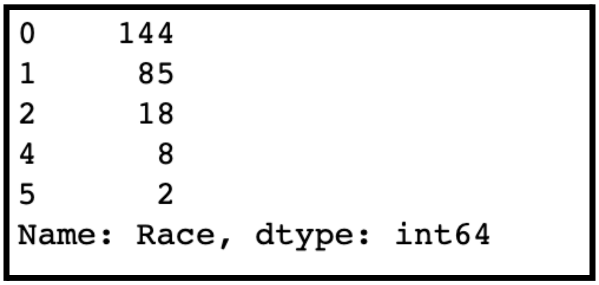
Problem type: *Classification*

Models used:

- kNN (k-Nearest Neighbors)
- SVM (Support Vector Machine)

Context: the data contains the following:

- 56% of the shooters were White (144/257)
- 33% of the shooters were Black (85/257)
- 7% of the shooters were Asian (18/257)
- 3% of the shooters were Native American (8/257)
- <1% of the shooters were 2 or more races (2/257)



```
0    144
1     85
2     18
4      8
5      2
Name: Race, dtype: int64
```

Results:

Model	***Data***	***Accuracy***	***Error***
kNN (*k=2*)	Training & Testing	**~59%,** (0.5891472868217055)	~41%, (0.4108527131782945)
SVM	Training & Testing	**~52%,** (0.5193798449612403)	~48%, (0.48062015503875966)

Analysis:

Our classification models were able to predict the race of the shooter with moderate accuracy, similar to our ability to predict the Mental Health Status. Our models performed with about ~52% - ~59% accuracy, and so what this tells us is that, similar to mental health, we can determine the race of the shooter solely based on the variables we are given (mental health status, gender, and total number of victims) with moderate accuracy. Interestingly enough, we actually can eliminate the variable of ***gender*** because, as previously explored, over ~98% of the data ended up being male shooters, and so this would not allow our race target variable to be accurately predicted based off of the gender variable. This leaves us mental health status and total number of victims as our real feature/predictor variables.

In the future, to get a better grasp on this specific task, we should use more variables than just the ones that we encoded. Some variables of interest within the dataset may be *target*, *location*, *employed*, *age*, and *cause*. Additional variables will give the models we train more features to base our predictions off of and may improve performance. Given our current data, a lot of these variables are not rich enough to be included in our model evaluation, as a lot of the data is missing for most of the records.

Task/Problem #4a: *Predicting the* ***total number of victims*** *of a mass shooting*

Problem type: *Regression*

Models used:

- Linear Regression

Results:

Model Performance:

Data (% = size)	***Metric***	***Value***
Training Data (50%)	MSE	2101.6067217792
	RMSE	45.84328436946027
Testing Data (50%)	MSE	175.3739841254949
	RMSE	13.24288428271934

Feature Importance (for Linear Regression):

Feature	***Coefficient***
Race	-5.410165
Gender	-0.005663
Mental Health Issues	4.797411

Analysis:

The performance of the Linear Regression model is measured through the MSE and RMSE metric. The MSE is the average of all of the residuals (difference between actual and predicted values) squared. In terms of our actual values, the MSE for both of our data sets is not great, but the RMSEs have redeemable scores. The RMSE measures how concentrated the data is around the regression line, and relates to how far the data points vary from the predicted values on the regression line. The RMSE values for both are not awful - considering what the task at hand was. We were trying to predict the total number of victims, and that number can vary a lot. On average, if we only vary about ~13 on our testing data set, that is not something to be ashamed of. However, the model can definitely be fine-tuned in the future to improve performance.

Additionally, from our Linear Regression model, we were able to derive a coefficient table, which tells us the importance/weight of each variable/feature on the regression line. From this coefficient table, we can see that the variable with most weight on the regression model was Mental Health Issues. This is important and backs up a conclusion from before stating that Mental Health is a really important area to invest more research and resources into, because on multiple occasions it has proven to be a factor that plays a role in increased gun violence and mass shootings.

In a *feature importance analysis*, the prediction of the total number of victims of the shooting places highest importance on the factor of **Race**, which has the largest coefficient of -5.410165. This tells us that when attempting to determine how many victims a future mass shooting will have, the race of the shooter interacts the most with the target variable.

Task/Problem #4b: *Predicting the* ***total number of victims*** *of a mass shooting*

Problem type: *Classification*

Models used:

- Logistic Regression

Results:

Model	***Data***	***Accuracy***	***Error***
Logistic Regression	Training & Testing	**~78%,** (0.7835051546391752)	~22%, (0.21649484536082475)

Analysis:

The regression task that involved predicting the total number of victims of a mass shooting was transformed into a classification task through some further data processing. The current variable of total victims was placed into *buckets*, i.e., if a record had 14 victims, it was placed into a bucket of 10-19 victims. This was then assigned a value (the value being the index of the list, i.e., 10-19 victims = index 1)[1]. Through this process, we were able to encode the victims column as a variable represented by a single value. This was also important because, as previously mentioned, the total number of victims can be really hard to precisely predict, as the model will be considered wrong if you are just one number off. However, it would be nice to consider this case because the model got pretty close but not exactly right, which is why we are using a bucket technique.

After this process, we ran the task as a typical classification problem. Running a Logistic Regression model through this newly processed data, we see a great model performance of **~78%**. This is pretty good performance because we are able to predict the right "bucket" of victims based on the few variables that we have provided to us with over 75% accuracy. Keeping this in mind, in the future this can hold importance because we can use other variables like location and "open" or "closed" location to predict in the real world what areas of the world we should be careful of or keep extreme security on.

[1] This is all explained further in the Jupyter Notebook

VI. Conclusions

After looking over the results of the various machine learning models used and the performance of them for each task, we can draw several important conclusions. For one, there is a consistently large percentage of mass shooters that tend to be male. Out of all of the data records, over 98% of the data ended up being a male shooter. This is an important factor to keep in mind when focusing on mass shooting research and aid; a pattern like this is too prominent to disregard.

Similarly, we were also able to gather important information about the mental health status of the shooters. The perpetrators of these mass shootings tended to be suffering from mental health issues more often than not. About ~53% of the data contained shooters that were suffering from these issues, and our model was able to predict with about ~55% - ~65% accuracy whether a shooter of a mass shooting was going to be suffering from these issues. This is a really important observation because this tells us that we should devote more time and resources to studying the effects of bad mental health, its causes, and figuring out ways to help. It is also an important note to realize that not all victims of bad mental health become perpetrators of mass shootings but more often than not a perpetrator of a mass shooting suffers from bad mental health.

Something we also noticed within the report was that at least two of variables that were being predicted placed a heavy emphasis on Race as a factor. Considering a good amount of the models place a heavy emphasis on this factor, if we predict a variable such as total number of victims, then we can say that the models used the Race variable as a factor of predicting whether the total number of victims that would result from a mass shooting. Looking into context, over half of the records contained mass shooters that were White (~56%). This number is small enough not to be an overwhelming majority (and subsequently conclude that the models did not train well on the data), but also not small enough to be ignored. Using this context and the feature importance of the models, we can spend time and research to determine how race plays a factor with mental health issues, and why certain races suffer more from mental health issues than others, or how each affects some more than others. This is a field/sector that has a lot of potential to be explored, and based on this research is worth the time and effort.

Overall, mass shootings are an extremely volatile and sensitive topic, with chaotic patterns and consistently unpredictable factors. However, through data exploration and analysis, we can attempt to make some sense of the phenomenon before us. It was found that more research and time should be dedicated to understanding how mental health and race play a role in the perpetrators of mass shootings, as these were the variables that ended up having the largest weight in predicting various features of mass shootings. In general, these are sectors that should have more attention devoted to it, and with a problem of this magnitude, it is urgent and imperative to take action on this immediately.