A Study on US Mass Shooting Using Data Analysis and Machine Learning

Andrew Fang
Wayland High School

ABSTRACT

Mass shootings have become one major problem in our country years ago. In 2021 alone, an astonishing number of more than 45,000 people were murdered in mass shootings (Reference 1). And I started to wonder, why are people doing this? If we can get some clues with existing data, it might help prevent future tragedies from happening. In many shooting incidents, the shooters seem to massacre without any reason. Many people wonder if it’s related to the murderer’s mental health. I started to gather some data and found two datasets about mass shootings and mental health on Kaggle (Reference 2 & 3), a public data-sharing website. First, I performed data analysis and statistical testing with the two datasets. To further investigate the relationship between mass shootings and mental health, I fitted a linear regression model with the merged datasets. I found out that there’s an obvious correlation between these two variables, which means mental illness was one of the direct reasons that caused mass shootings. In addition, I want to help people avoid mass shootings. So I made a linear regression model, which helps to predict how many total victims there will be based on input factors, like location, race, age, etc. Using this model, people can put more security in dangerous locations to best avoid mass shootings.

Introduction

The recent mass shooting in a Texas school horrified the whole world. The shooter brutally killed nineteen students and two teachers. For many years, gun violence has become increasingly commonplace throughout the U.S. It is particularly heart-wrenching that in a lot of cases, victims are children. They were killed for no reason. And I start to wonder why shooting tragedies keep happening in the U.S. There are many aspects to this question, so I did some research and listed a few reasons: (Reference 4)

First, guns are all over the place in our country. There are even more guns than people. According to statistics, there are more than 390 million guns in the U.S, while our population is only about 344 million. We have the highest civilian gun ownership compared to any other country.

Second, it is the division among people’s political views. Some fraction of people think our country is “under threat”, and violence is the only way to protect the nation.

Now they are more dangerous because the fast-growing social media platforms have become a powerful tool to spread violent opinions on public affairs quickly. It can also help terrorists secretly gather up and plan attacks. The contradictory political views are much deepened these years.

Lastly, it is what we are all interested in: are most mass shootings caused by people’s mental illness? In our society nowadays, a lot of people experience an unprecedented amount of pressure to be able to afford basic needs to live. Social media also revealed unrealistic lifestyles of the elite group, which causes even more anxiety in people. The percentage of people having mental health issues is alarmingly high. Could this be one of the reasons that drive people over the edge and commit heinous crimes? Since this topic is what I’m interested in, I decided to use machine learning and data science to find out if there’s a relationship between people’s mental illness and mass shootings.
Additionally, I wanted to set up a model to predict how many victims there will be in a shooting case, based on factors such as location, state, and race. So people can be more careful and have more security in dangerous places the model predicts. So, I’ll do a regression model with machine learning. A regression model is a function that describes the relationship between input values and gives the output best related to input factors. In other words, it predicts the output based on the input factors.

**Data and Method**

Since I want to figure out the connection between mass shootings and murderers’ mental health, I need the datasets for both to analyze if there’s a connection between them. So I found two datasets, one about Mass shootings’, including all details; another about people’s mental illness problems over time.

The mass shooting dataset has 323 rows and 21 columns, like date, target, caused, murderer’s age, etc. The timeframe of this dataset is from 1966-08-01 to 2017-11-5. Before I can analyze the data, some “cleaning up” is needed. For example, sometimes there is more than one murderer in a shooting case, and that caused there to be two numbers in the “murderer’s age” column.

**Figure 1.** Five rows of data with wrong age data (there are two values)

Therefore, I need to change it to only one number, to graph the age column and to prepare for the regression model. Besides that, I also need to rename some columns to make them more understandable.

Next, I randomly selected five rows of the dataset to show, and I noticed that there are many missing values in this dataset but described as “unclear” or “unknown”, which are categorized as null values by the machine. So I went over every column’s unique values and changed them to “NaN” which is Python’s null value.

There are also other problems in this dataset, like the “location” columns. It provided the city and state where the shooting happened, but machine learning can’t analyze this input. So, I need to make two new columns: city and state, to give the machine a clear input to analyze.

To examine and visualize the data distribution of numeric features, figures 2 and 3 are created.
At last, to prepare to put the data into machine learning, I need to convert object-type columns (Open/close location, Gender, Mental, Race) to numbers by one hot encoding and frequency encoding, a data type that changes strings to numbers so the machine understands.

For the rest of the texts/descriptive columns, word clouds were made to showcase the top-occurring words in each column.

**Figure 2.** Table of variable statistics, including numeric columns of the dataset.

**Figure 3.** Table of histograms, provides data distribution.
Mass Shooting Target Word Cloud

![Target Word Cloud](image1)

**Figure 4.** Target word cloud, bigger text means the word has a higher occurring frequency in the data.

Location Word Cloud

![Location Word Cloud](image2)

**Figure 5.** Location word cloud, the bigger text means the word has a higher occurring frequency in the data.
Figure 6. Incident area word cloud, the bigger text means the word has a higher occurring frequency in the data.

Similar steps were taken for the google trends data which allows us to see what people are searching for at a local level. It covers data from 2004 to 2017. I created the state column from the location column to prepare for merging with the mass shooting dataset.

After merging the two datasets by state, columns that have less than 1500 rows of non-null values were dropped. New columns of yearly total victims were made to be comparable with the yearly google search data. The correlation matrix of the correlation coefficients between numeric variables is visualized through a heatmap (as shown in figure 7). It shows that there’s a linear relationship between 2005 total victims and subsequent years’ depression search, and between 2010 total victims and some of the years’ depression search.
Heatmap visualizes the correlation matrix between the depression search records and the total number of victims of mass shootings from 2004 to 2017. The correlation coefficients stand for how linearly related each column is to each other. For example, the coefficient of depression from 2010 depression and 2010 total victims is 0.73, which is a pretty high number. That means there is a 0.73 positive relationship between 2010’s depression and 2010’s total victims.

To investigate the relationship between categorical and continuous variables, an analysis of variance (ANOVA) was performed, as shown in figure 8. Since all the p values are smaller than 0.05, there is a statistically significant relationship between gender, race, state, mental illness, and total victims.
Linear Regression (Reference 5 & 6)

To predict future mass shooting victims based on current trends, I tried lasso regression and random forest regression models. I used lasso regression instead of linear regression because there are a few input variables that are correlated with each other. The regularization parameter in lasso regression can avoid the multicollinearity problem among input variables. The coefficient of determination (R squared score) of lasso regression and random forest regressor are 0.997 and 0.960 respectively. The closer to 1 the R squared score is, the better fit the regression model is. So lasso regression performed better than random forest regression and can be used to predict future mass shooting victims in the US.

Conclusion

Diving into the analysis of mass shootings using Python language and machine learning, I was shocked, both by the technology and the result. I’ve learned programming languages like java and python for a while, but this is the first time I use it to do a project like this, and I’m impressed by how efficient it is. Using libraries like seaborn and sklearn, I can analyze the data without manually doing so. The correlation heatmap revealed a positive linear relationship between some years of mass shooting victims and the depression search data. And the lasso regression model was a great fit for the combined dataset. The insights hopefully will help the government to limit gun ownership to people with mental issues. Also, the learning teaches us to be cautious about certain locations to avoid incidents.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

https://www.gunviolencearchive.org/past-tolls (1)
https://www.kaggle.com/datasets/zusmani/us-mass-shootings-last-50-years (2)
https://www.kaggle.com/datasets/GoogleNewsLab/health-searches-us-county (3)
https://en.wikipedia.org/wiki/Regression_analysis (5)
https://hbr.org/2015/11/a-refresher-on-regression-analysis (6)