Analysing Gun violence in the United States

Anthony Robert Feliciano, Aboli Rawanhate

Department of Engineering and Applied Science, SUNY University at Buffalo

arfelici@buffalo.edu

aboliraw@buffalo.edu

Buffalo, NY 14260

I. Introduction

In the United States gun violence and gun restriction is a fiercely debated topic amongst many public representatives and citizens alike. The reason being that gun violence is very prevalent in certain parts of the country. Some have attributed this to the lack of gun restrictions in certain parts of the United States. This has led to a large diversity of different gun laws and restrictions being implement in different states.

The question that naturally arises is how effective are these gun restrictions at lowering the rates of gun violence and which ones are most effective? Also what factors lead to being a victim or perpetrator of gun violence? This is what we set out to learn when starting this project.

II. Implementation Methods

Overview

As instructed in class we split this project into five different phases:

1. Raw data collection and Data Processing
2. Exploratory Data Analysis
3. Applying Statistical Models and Machine Learning algorithms to the Data
4. Documentation of code base and Data Visualization
5. Product Building

In the following sections we will describe in detail our methods and practices for completing each phase.

1) Raw data collection and Data Processing

Before seeking out data we formulated our problem statement. Our problem statement being, What factors lead to increased amount of gun violence incidents?

For our raw data collection we mainly made use of the following three datasets:

- Gun Violence Data, a comprehensive record of over 260k US gun violence incidents(Reference 1)
- Firearms Provisions in US States(Reference 2)
- U.S. Census State and County Population Estimates 2000-2019 (in EN)(Reference 3)

The first dataset contains detailed information about gun violence incidents from 2013-2018. Such as the number of participants and demographic information related to said participants.

The second dataset defines 137 different categories of gun restrictions and has when they were implemented. We use this dataset to measure the effects of certain gun legislation.

The third dataset simply contains the total population of a particular state at a given year from 2000 -2019

Merging these three datasets we not only have detailed information about each incident and where it took place. But we also have information about each state and it's restrictions.

The datasets intersect from years 2013-2017 giving us four years total of gun detailed gun violence data.

We cleaned that data by removing irrelevant columns as well as normalizing the incident dataset to gain detailed information about each participant per incident.

We also had to clear out nan values and make assumptions about certain areas of the dataset.

One of the areas was the type of weapon used in each incident, since many rows were left empty in this category we made an assumption that every missing weapon value would be filled in by the average gun used in the county and state.

After cleaning and processing the data we formed our hypothesis.

Our Hypothesis based on prior assumptions are as follows:

- The location will affect the amount of gun violence incidents
- Young males will be most likely to be the perpetrators and victims of gun violence
- The state population will have a significant effect on the amount of gun violence incidents
- States with many gun restrictions will have a lower amount of gun violence incidents

2) ***Exploratory Data Analysis***

In the exploratory data analysis stage we prepared inputs for multiple machine learning algorithms.To do so we had to extract certain data from our dataset.

We made use of the following EDA operations:

- Univariate analysis
- Bivariate analysis
- Multivariate analysis
- Pearson’s Correlation Coefficient
- Principal Component analysis

Using Univariate we compared correlation between many features and observed their effect on the total number of gun violence incidents in our dataset.

Principally we measured correlation between population,total amount of gun laws and certain using scatter plots and visualized them as seen in the figures below.

Our findings are as follows:

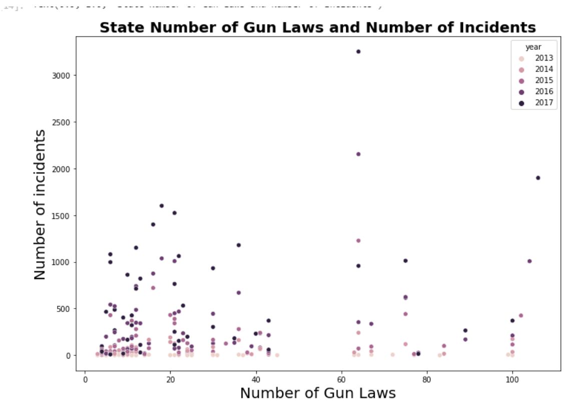


Figure 1.1 Correlation between number of gun laws and total number of incidents

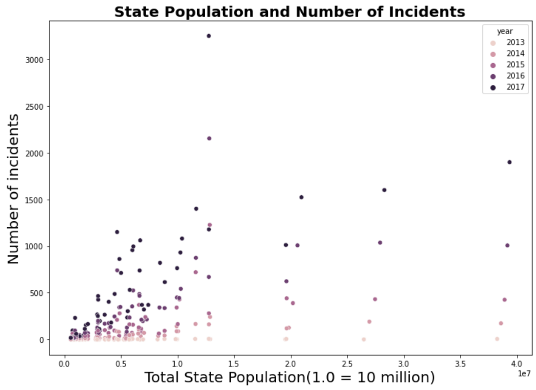


Figure 1.2 Correlation between total state population and total number of incidents

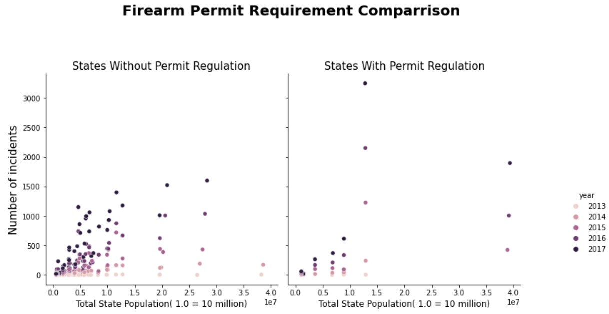


Figure 1.3 Comparison and Classification of States that require firearm permits to own a firearm

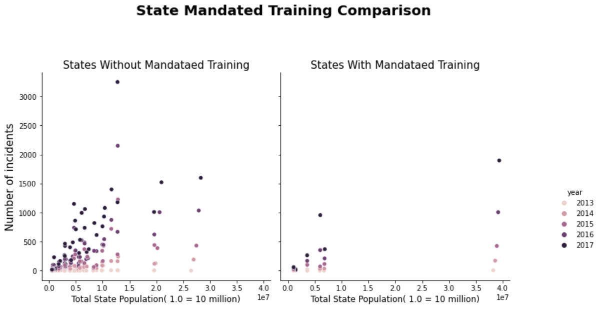


Figure 1.4 Comparison and Classification of States that require firearm safety classes to own a firearm

We also labeled certain incidents as mass shootings if there were more than three victims killed and analyzed this data as well. As shown in Figure 1.3.

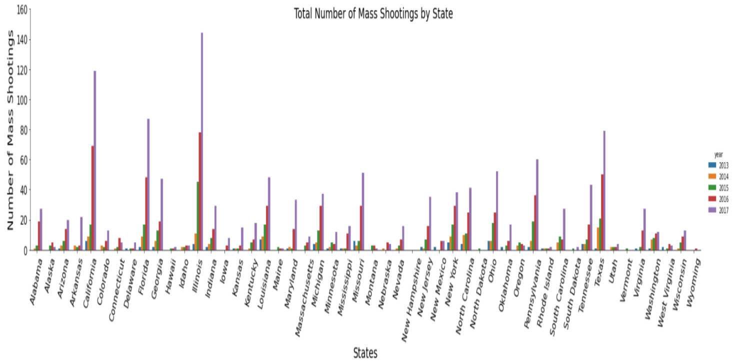


Figure 1.5 Amount of mass shooting incidents by state

To determine the most meaningful inputs to our algorithms for the machine learning phase of the project we used Pearson's Correlation Coefficient in order to create a heat map of all of our selected features

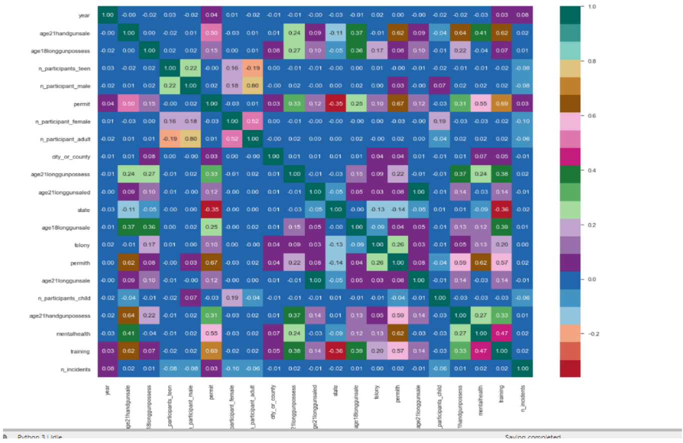


Figure 1.6 Heat map of all features

We plotted the heat map to check for highly correlated features from the figure above to choose as our inputs.

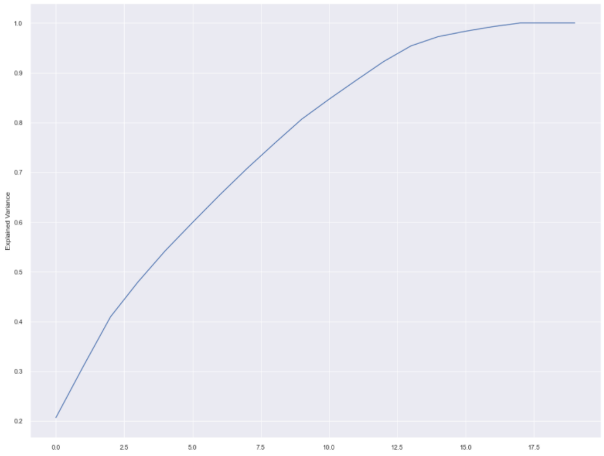


Figure 1.7 Elbow Plot

From the graph, we observed that the first 14 principal components keep about 95.61% of the variability in the dataset while reducing 6 (21–14) features in the dataset. That's great. The remaining 6 features only contain less than 5% of the variability in data. So we considered 14 principal components.

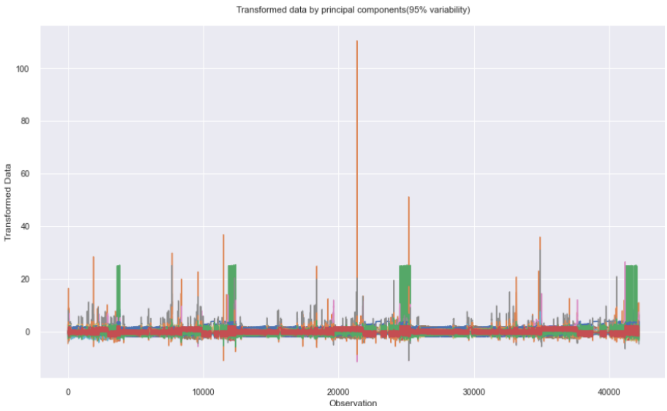
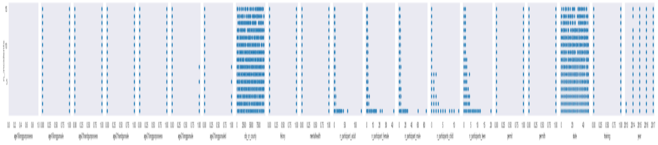


Figure 1.8 Transformed data by Principal Components(95% variability)

3) Applying Statistical Models and Machine Learning algorithms to the Data

I. Linear or Nonlinear



We plotted every feature against the number of incidents to check that data is linearly separable or not. From the graph, we observed that the data is non-linear.

II. Best-Fitting Regression Model

We validated data to test the effectiveness of a generated model on real world data using Cross validation.

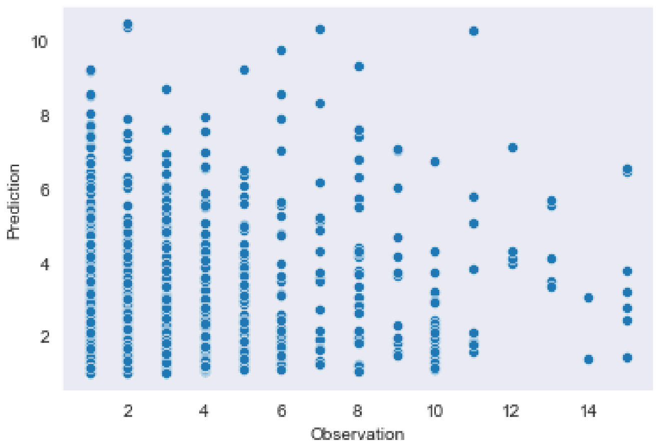
We applied Linear Regression after transforming nonlinear data to linear using yeo-johnson method so that probability distribution of the feature is more Gaussian. R squared is 4% for this algorithm. That means data is not so close to the regression line.

We applied Ridge regression and Lasso Regression to adjust variance and bias in the data. But it performed badly on the data. We measured error using mean squared error and r squared.

SVM Regressor performs better on nonlinear data as it has a kernel concept in the backend for linear transformation to higher dimension space. It gives Root Mean squared error = 1.76.

We applied a Neural Network algorithm, as our data have non-linear relationships. Neural Networks are good with learning functions which don't have very well defined shapes i.e neural network can learn any function or shape/graph. So neural networks can be used to learn the function behind the data and may be able to predict the number of incidents in a better way. Root Mean Squared Error : 1.80, it performed better like SVM. But we still didn't get our best fit.

Random Forest Regressor proved to be the best fit for our data. As it is an ensemble method there is very little chance of overfitting. Data is normalized so that all columns are kept at the same scale to any kind of unwanted bias in the model. So, Root Mean squared error for this model is 1.66 which is very less compared to other models. Also after plotting prediction vs actual, we got better results.

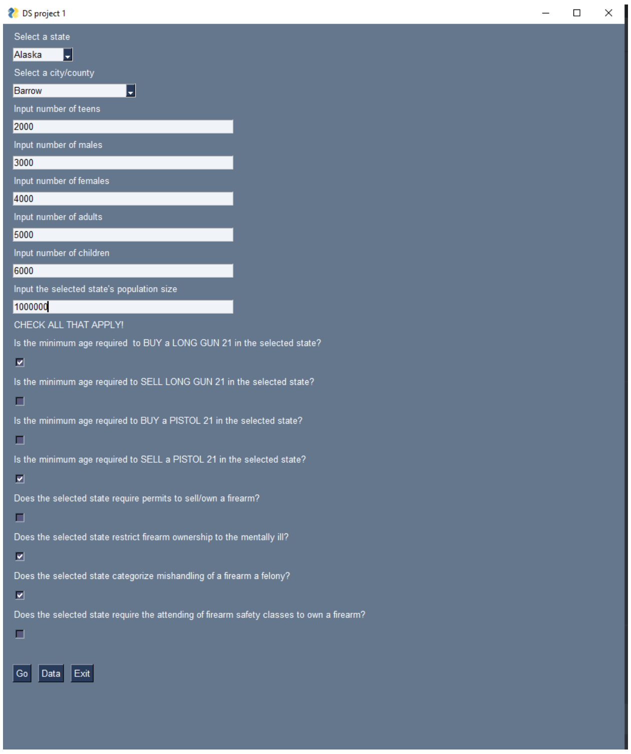


III. Product Instructions

- Fill in the questionnaire and make sure to answer all the questions.

- The first section of the questionnaire you input numbers into the boxes.

- The second boxes you only check the boxes that apply to the state you have selected.



Select a state

Alaska

Select a city/county

Barrow

Input number of teens

2000

Input number of males

3000

Input number of females

4000

Input number of adults

5000

Input number of children

6000

Input the selected state's population size

1000000

CHECK ALL THAT APPLY!

Is the minimum age required to BUY a LONG GUN 21 in the selected state?

☑

Is the minimum age required to SELL LONG GUN 21 in the selected state?

☐

Is the minimum age required to BUY a PISTOL 21 in the selected state?

☐

Is the minimum age required to SELL a PISTOL 21 in the selected state?

☑

Does the selected state require permits to sell/own a firearm?

☐

Does the selected state restrict firearm ownership to the mentally ill?

☑

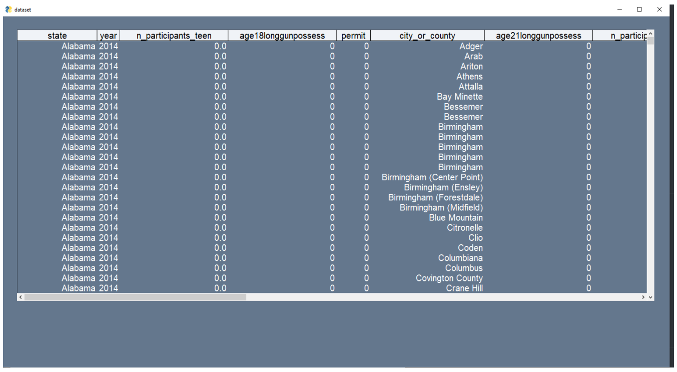
Does the selected state categorize mishandling of a firearm a felony?

☑

Does the selected state require the attending of firearm safety classes to own a firearm?

☐

To see the data we use to make these calculations you may click the data button.



state	year	n_participants_teen	age18longgunpossess	permit	city_or_county	age21longgunpossess	n_particip
Alabama	2014	0.0	0	0	Adger	0	
Alabama	2014	0.0	0	0	Arab	0	
Alabama	2014	0.0	0	0	Ariton	0	
Alabama	2014	0.0	0	0	Athens	0	
Alabama	2014	0.0	0	0	Attalla	0	
Alabama	2014	0.0	0	0	Bay Minette	0	
Alabama	2014	0.0	0	0	Bessemer	0	
Alabama	2014	0.0	0	0	Bessemer	0	
Alabama	2014	0.0	0	0	Birmingham	0	
Alabama	2014	0.0	0	0	Birmingham	0	
Alabama	2014	0.0	0	0	Birmingham	0	
Alabama	2014	0.0	0	0	Birmingham	0	
Alabama	2014	0.0	0	0	Birmingham	0	
Alabama	2014	0.0	0	0	Birmingham (Center Point)	0	
Alabama	2014	0.0	0	0	Birmingham (Ensley)	0	
Alabama	2014	0.0	0	0	Birmingham (Forestdale)	0	
Alabama	2014	0.0	0	0	Birmingham (Midfield)	0	
Alabama	2014	0.0	0	0	Blue Mountain	0	
Alabama	2014	0.0	0	0	Citronelle	0	
Alabama	2014	0.0	0	0	Clio	0	
Alabama	2014	0.0	0	0	Coden	0	
Alabama	2014	0.0	0	0	Columbiana	0	
Alabama	2014	0.0	0	0	Columbus	0	
Alabama	2014	0.0	0	0	Covington County	0	
Alabama	2014	0.0	0	0	Crane Hill	0	

- Press the Go button to calculate the number of incidents.

4) Analysis and Discussion:

We are accepting state, year, population metrics and law information as input from UI and predicting the number of gun violence incidents using Random Forest Regressor.

To improve the predictive accuracy and control over-fitting, we selected this algorithm. It is giving a more accurate and stable prediction.

Conclusion :

Gun Violence is a serious issue in the United States, it needs to be attended and resolved. We have observed that Population distribution, gun laws play an important role in gun violence incidents. As per our observation, suspects of most of the incidents are male so the state/country having male ratio maximum and lenient laws are susceptible to more Gun Violence incidents.

References

[1] https://www.kaggle.com/jameslko/gun-violence-data?select=gun-violence-data_01-2013_03-2018.csv

[2] https://www.kaggle.com/jboysen/state-firearms?select=raw_data.csv

[3] https://www.osti.gov/dataexplorer/biblio/dataset/1617641

[4]https://towardsdatascience.com/cross-validation-a-beginners-guide-5b8ca04962cd

[5] https://machinelearningmastery.com/regression-metrics-for-machine-learning/