Predicting Gun Violence and the Effectiveness of Laws

by

Xin Bian[1], Haomin Liu[2]

Supervised by

Prof. Jiebo Luo[3]

[1] Department of Mechanical Engineering, [2] Materials Science Program, [3] Department of Computer Science
University of Rochester
Rochester, New York

December 14th, 2017

1. Introduction

Gun violence in the United States is a substantial public health concern. Every day in USA, more than 200 people murdered or assaulted with a firearm[1]. Number of mass shooting in the US are on the rise, shown in figure 1. Over the past couple years, many studies have been working on explaining, preventing, and predicting gunshot violence at the individual level, community level, and federal level.

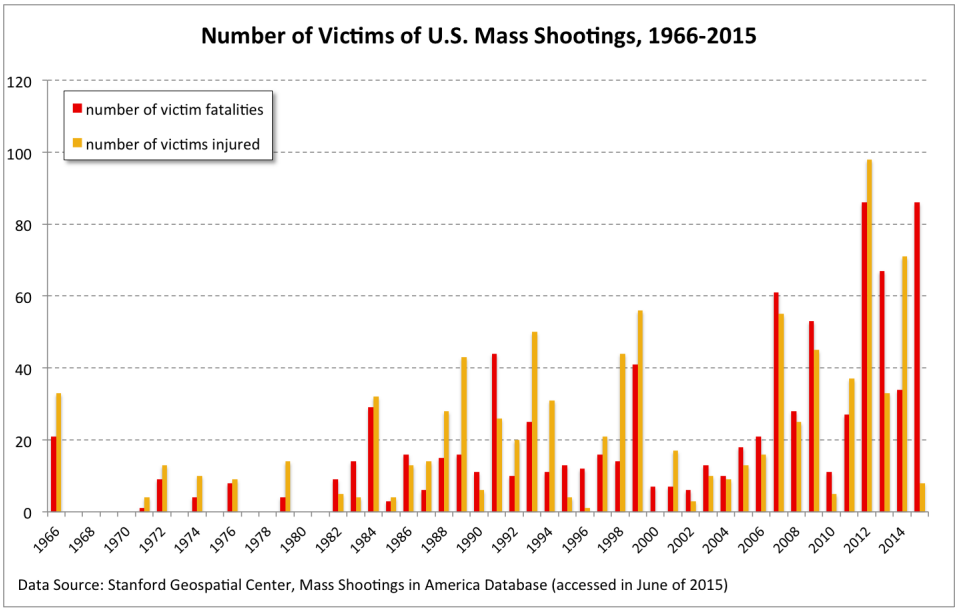

Figure 1. Number of Mass Shooting in US[1]

Green, et al[2] proposed an algorithm that tries to predict who is most likely to be involved in a shooting, either as perpetrator or victim. This algorithm assigns scores to people based on their criminal records, for example, arrests or shootings as well as any known gang affiliations and other variables. However, it is found that the assigned risk score is sometimes at odds and against some common perceptions.

Another study by Kalesan, et al [3] compared 25 different state gun laws and concluded that implementing three state laws at federal level could reduce the rate of US gun deaths by more than 90%. It is questionable that implementing three relatively modest gun restrictions could have a huge impact on gun death.

Fleegler, et al.[4] created a 'legislative strength score' using state-level firearm legislations of 5 categories of laws. They measured the association between forearm-related fatalities and the strength score. However, as noted in this article, the legislative strength score has not been validated, and some other factors such as nonfatal firearm injury, exploitation of loopholes are not considered.

Lee, et al. [5] summarized 34 articles investigating the relationship between firearm laws and firearm homicides from 1970 to 2016. Certain stronger gun policies are associated with reducing firearm homicide rates. Also, they pointed out that legislation is not the only way to help the country decrease firearm tragedies.

As best as we know, the literatures studying gun-control laws didn't consider enough variables that might affect gun violence. For example, an region's average income, crime rate, unemployment rate, education level, race, might associate with firearm crimes. There are so many factors need to be considered. Moreover, all papers studied firearm laws impact on state-level. However, within a state, different regions can be very different with respect to crime rate, economics, education, etc. For example, New York City is very different compared to cities in Upstate New York. New York City might have more legal and illegal immigrants, stronger economics, higher crime rates. Any conclusions drawn from the state level rather than city or country level might be not valid or persuasive.

In this project, we are aiming to analyze the problem using the firearm death data along with factors including unemployment rate, education level, crime rate, poverty rate, median household income and the strictness of law regulations at county level. The techniques used are Naive Bayes, Random Forest, and Neural Network. We first build regression models to predict firearm death rate, then try to find the relation, if any, between the strictness of law regulations and firearm death rate.

2. Methodology

The goal is to design a model which can be used to predict firearm death rate and to study the influence of laws on the firearm death rate. The data sets are collected in U.S at county level.

Approach would involve the following steps

1. Collecting county level data.
2. Data cleaning and selection of attributes
3. Algorithm implementation, such as k-NN, k-means, Naive Bayes, Random Forest, Neural Network, for predicting the firearm death rate

Algorithm details

1. Tools: Weka, Tableau, Exploratory, scikit-learn
2. Languages: python 2.7
3. Algorithms: PCA, k-NN, k-Means, Naïve Bayes, Neural Network, Random Forest
4. Verification: 10-fold cross validation

3. Data Preprocessing

3.1 Data Source

The firearm death data are from US Centers for Disease Control and Prevention[6]. The data set has fire accidents recorded by government and local body. Data are based on death certificates for U.S. residents. The data of mortality caused by firearm from 1999 to 2015 is used.

The data of poverty rate, median household income, unemployment rate, education level of 2015 are from US Department of Agriculture Economic research Service[7].

State Gun Control Laws is obtained from website of Brady Campaign to Prevent Gun Violence[8]. It provides data about gun control laws by state and a grading system for gun control laws (the year of 2013). We use the curved score for laws to measure the strictness of law regulations in a state.

3.2 Preliminary Analysis

Zero Values Removal:

We use a geographic approach to study mass shooting data in the United States, as shown in figure 2. It was found that the southern part of US had witnessed the maximum number for fire

accidents for that during these years while the majority of other counties are having a very low firearm death rate, which is in the range of 0 to 0.2. Furthermore, about 1500 counties in the dataset have the firearm rate of 0, as shown in figure 3 in blue, and it would be noise and outliers in prediction models. Thus, we decided to remove the zero values. Green region in figure 3 shows the data distribution after removing.

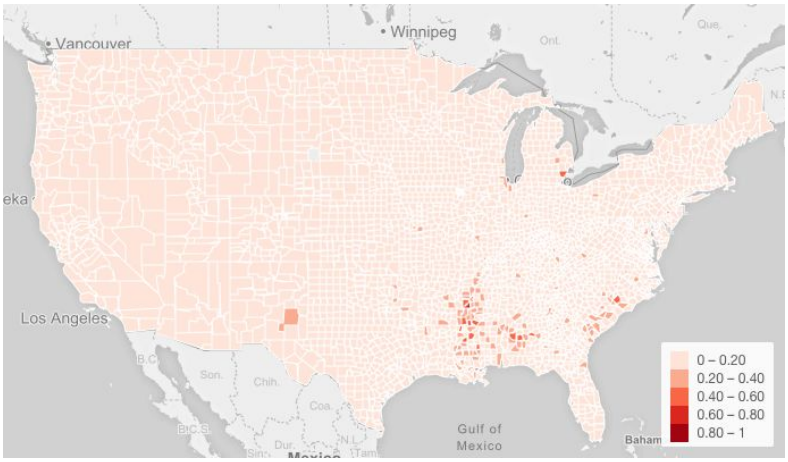

Figure 2. Graph for US Firearm Crime

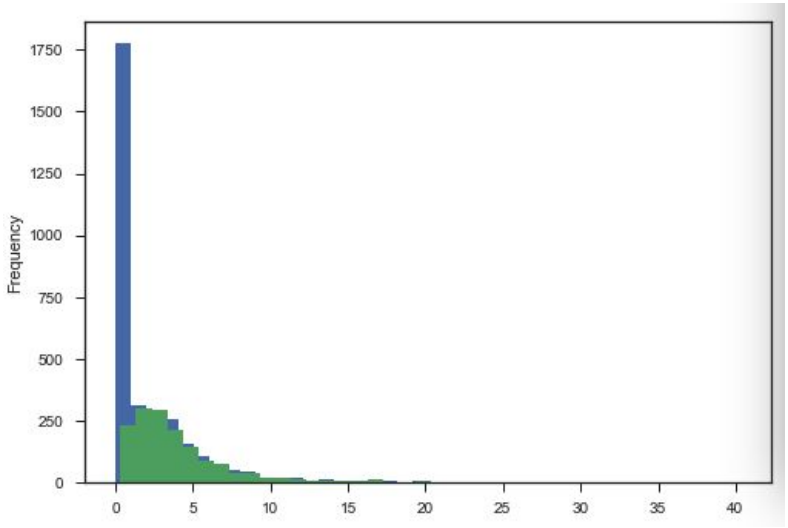

Figure 3. Distribution of Firearm Death Rate Data.

(Blue is raw data. Green is after cleaning)

Attributes Selection:

The data sets have attributes of poverty, education, income, law, unemployment and crime rate. Initial survey and research suggested, in some extent, the firearm rate is a subset of the crime rate. A simple linear regression analysis is applied to study the relation between crime rate and firearm rate. Results is shown in figure 4, even though the correlation is not well determined by this method we decide to remove crime rate attribute. In the future we are going to find a better way to capture the correlation between crime rate and firearm rate.

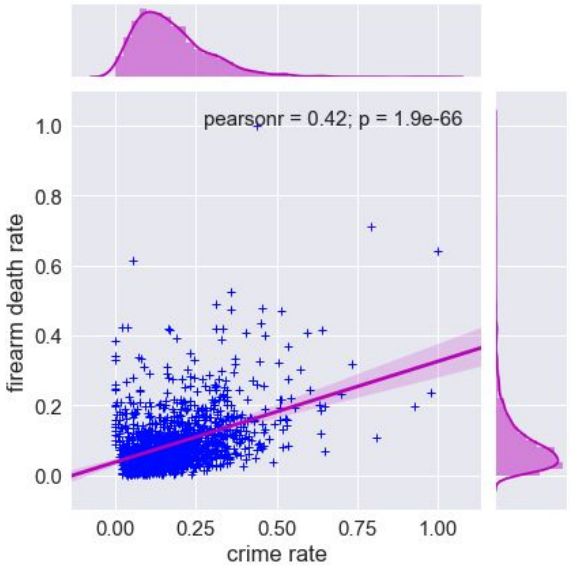

Figure 4. Relation between crime rate and firearm rate

.

On selecting attributes, we choose the data columns of median household income, unemployment rate, education level, poverty rate, curved score for laws, and firearm death rate. We consolidated the desired dataset to form different file, such as excel file, csv file, and arff file in order to prepare for the following implementation. For further pre-processing, we apply min-max normalization to rescale the numerical data. Final data set has total of 6 attributes and about 1500 tuples of different counties. Table 1 shows the attributes for final date set.

Table 1. Data set description

Attributes	**Description**
Income	2015 median household income

Unemployment	2015 unemployment rate
Education	2015 percent of adults completing some college or associate's degree
Poverty	2015 estimated poverty rate
Law	Restriction of law about gun control, higher value represents more restricted
Fire Rate	1999-2015 firearm death rate

4. Results

4.1 Predicting firearm death rate

The main goal is to develop a model to predict firearm death rate if given the data of county attributes. In this part, we discuss the implementation of different methods.

4.1.1 Naïve Bayes

Naïve Bayes is a classification algorithm and belongs to supervised learning category. In case of supervised learning, the label / classes for the data is already known. It is probabilistic algorithm which is used to classify / assign class label for new instances of a data set based on prior probability of already seen instances. Naïve Bayes calculates the prior probability of each class independent of other variables. Then it calculates the likelihood of a new instance belonging to a particular class. Finally, posterior probability is calculated by taking product of prior probability and Likelihood. The class with maximum posterior probability is assigned to the unseen instance.

In our case, the first phase of implementation had task of applying Naïve Bayes to predict firearm rate based on the county condition for a given area. The class was numeric. To perform the analysis task, Weka, a Data mining and analysis tool was used. The prepared data set was loaded in Weka. The visualization of raw data showed that data values were distributed over a large number. So the data was discretized and all the attributed were put into five bins. After pre-processing, Naïve Bayes was selected from the several data mining algorithm available. The default setting was chosen and instead of percentage split, 10-fold cross validation was used for building the classifier.

Naïve Bayes calculates the prior probability of known classes in as given: Prior probability of class level n = Number of instance of level n / Total number of data. Later Naïve Bayes calculates the Likelihood of a new instance to be part of Class n. Both the posterior probabilities are compares and the unseen instance is assigned a class with maximum posterior probability. The following figure shows the prior probability output from Weka.

Attribute	Class 1 (0.96)	2 (0.04)	3 (0)	4 (0)	5 (0)
Unemployment					
1	1752.0	10.0	2.0	1.0	1.0
2	994.0	84.0	10.0	2.0	2.0
3	52.0	14.0	4.0	2.0	1.0
4	1.0	1.0	1.0	1.0	1.0
5	2.0	1.0	1.0	1.0	1.0
[total]	2801.0	110.0	18.0	7.0	6.0
Household Income					
1	390.0	69.0	10.0	2.0	1.0
2	1549.0	33.0	5.0	2.0	2.0
3	693.0	6.0	1.0	1.0	1.0
4	133.0	1.0	1.0	1.0	1.0
5	36.0	1.0	1.0	1.0	1.0
[total]	2801.0	110.0	18.0	7.0	6.0
Poverty					
1	774.0	1.0	1.0	1.0	1.0
2	1405.0	20.0	1.0	1.0	1.0
3	518.0	57.0	2.0	1.0	2.0
4	87.0	24.0	8.0	3.0	1.0
5	17.0	8.0	6.0	1.0	1.0
[total]	2801.0	110.0	18.0	7.0	6.0
Education					
1	69.0	4.0	1.0	1.0	1.0
2	729.0	50.0	6.0	1.0	2.0
3	1329.0	50.0	9.0	3.0	1.0
4	621.0	5.0	1.0	1.0	1.0
5	53.0	1.0	1.0	1.0	1.0
[total]	2801.0	110.0	18.0	7.0	6.0
Law					
1	1565.0	77.0	7.0	3.0	2.0
2	691.0	26.0	6.0	1.0	1.0
3	436.0	2.0	3.0	1.0	1.0
4	1.0	1.0	1.0	1.0	1.0
5	108.0	4.0	1.0	1.0	1.0
[total]	2801.0	110.0	18.0	7.0	6.0

Figure 5. Prior probability calculated by Naive Bayes

The accuracy of Naïve Bayes classifier is 88.927%, which is very high. However, we notice our data is very screw. Most of the data lie in class level one. And we have few data lie in Unemployment level 3 to 4, Household Income level 5, Poverty level 4 to 5, Education level 1 and 5, and Law level 4. So It is likely that that our model produced by Naïve Bayes classifier is inaccurate at predicting the fire rate if given county condition is in above mentioned levels because we don't have enough training data for a more accurate model.

Moreover, we applied correlation analysis on the attributes. Results are visualized in figure 6. As we can see, only law and education, law and unemployment have a relatively low correlation. The rest attributes are all having correlation over 0.2, indicating the independent assumption is not well satisfied.

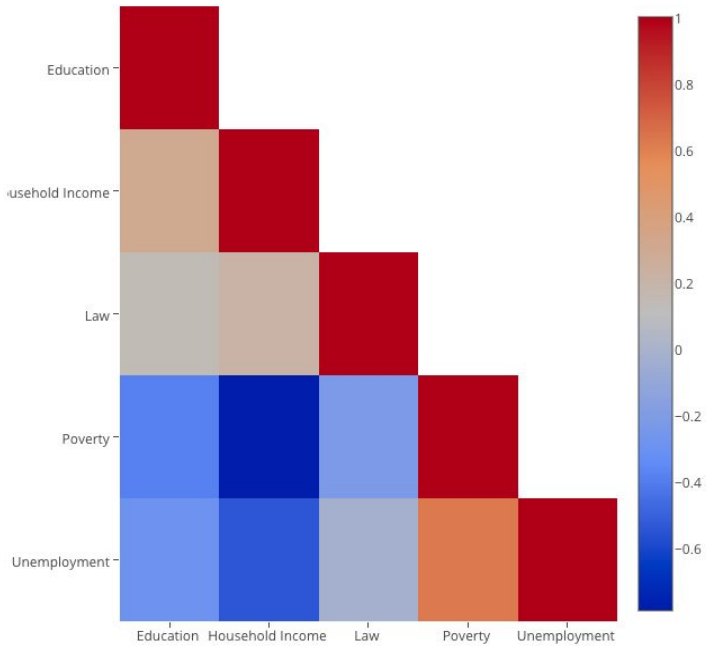

Figure 6. Visualization of correlation matrix

4.1.2. Random Forest and Neural Network

In this section, we build two regression models, random forest and neural network to predict firearm death rate. The RandomForestRegressor and MLPRgressor of scikit-learn are used to build the model.

For random forest, the default parameters are used. 5 attributes: Income, Unemployment, Education, Poverty and Law are used to predict Fire Rate.

For neural network, relu is used as activation function. The neural network has two hidden layers. Each hidden layer has 100 neurons. The input layer has 5 attributes: Income, Unemployment, Education, Poverty and Law, while the output is Fire Rate.

We conducted 10-fold cross validation on the two models by measuring the mean absolute error. The two models have similar errors. Random forest and Neural Network have

mean absolute error of 0.0488 and 0.0425, respectively. This is about 5% error or about 95% accuracy on average.

We can check the distribution of errors. The probability density function of percentage error is shown in Figure 7. The two models have very similar distributions. For most cases, the percentage error is below 15%.

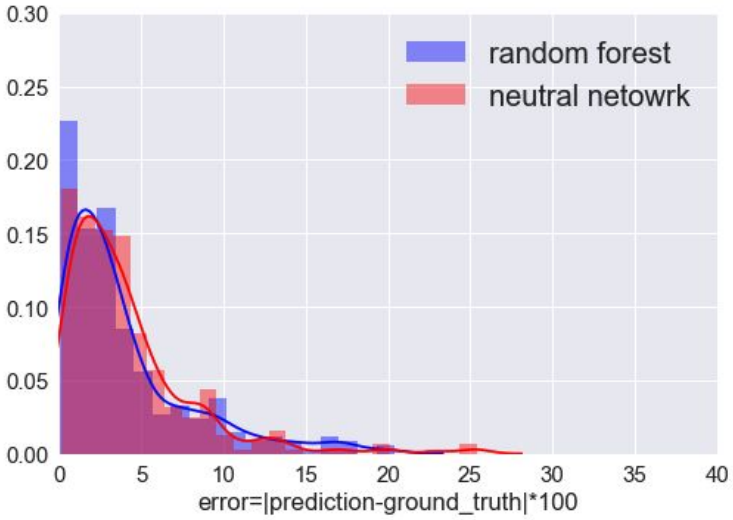

Figure 7. Probability density function of percentage errors

Moreover, the percentage error of Neural Network and Random Forest are visualized in a map, which are shown in Figure 8 and Figure 9. The prediction is conducted on a test data set. Darker regions mean larger error percentage. The percentage errors in the two maps show very similar results.

Table 2 and Table 3 show the 10 countries that have largest percentage errors in the two models. For Neural Network, we can see the counties that have large percentage error also have large fire rate. This is reasonable, since the data of firearm death rate is highly skewed, the tuples that have large fire rate are very few. The model doesn't have enough data to train at this high fire rate level, so its performance is low in this region. However, this is not the case in the Random Forest model. Even though high Fire Rate counties, such as Lake County, Essex County, Wayne County also appear here, several counties with very low Fire Rate appear. This is not what we expected. The Random Forest model has poor performance for high Fire Rate counties and some low Fire Rate Counties.

Table 2. Neural Network top 10 counties of percentage error

County	**Fire Rate**	**Percentage Error**
Wayne County, MI	0.41	26.57
Lake County, IN	0.35	26.90
East Baton Rouge Parish, LA	0.33	23.37
Essex County, NJ	0.29	22.31
Shelby County, TN	0.32	19.98
Montogomery County, AL	0.30	17.28
Washington County, MS	0.41	15.50
Mobile County, AL	0.25	15.33
Attala County, MS	0.28	14.15
Mahoning County, OH	0.22	13.36

Table 3 Random Forest top 10 counties of percentage error

County	**Fire Rate**	**Percentage Error**
Oktibbeha County, MS	0.085	26.97
Lake County, IN	0.35	21.99
Essex County, NJ	0.29	21.28
Winston County, MS	0.090	20.59
East Baton Rouge Parish, LA	0.33	19.02
Wayne County, MI	0.40	18.79
Kings County, NY	0.13	17.36
Shelby County, TN	0.32	16.64
Mahoning County, OH	0.22	16.49
Beauregard Parish, LA	0.055	16.42

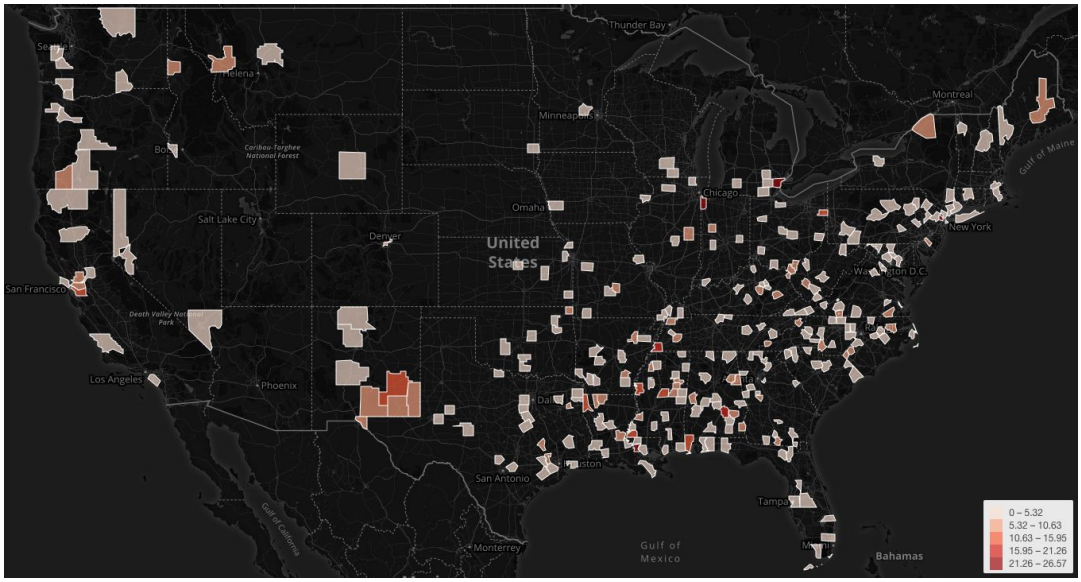

Figure 8. Percentage errors of Neural Network model

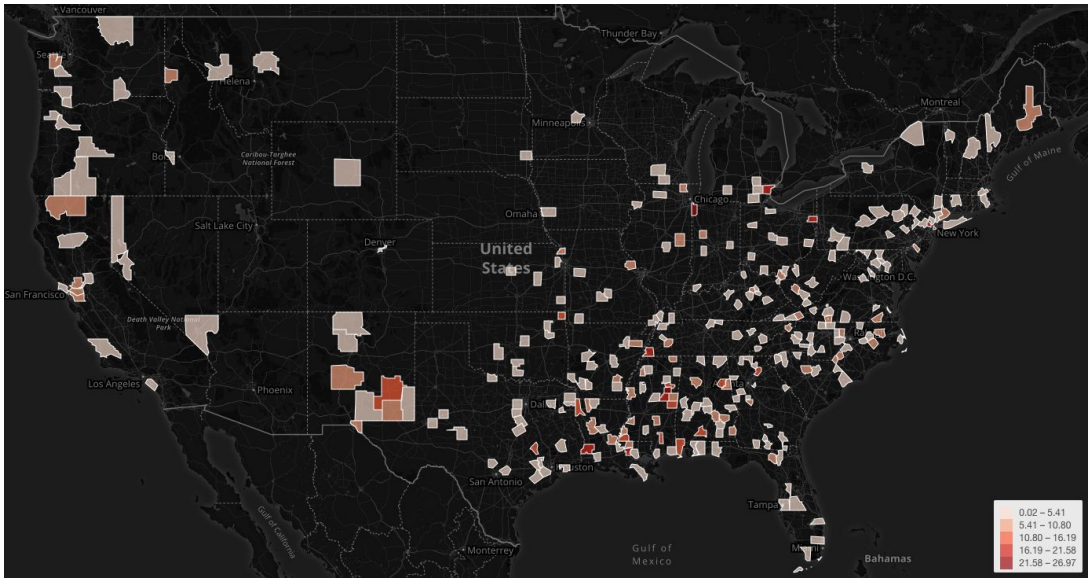

Figure 9. Percentage errors of Random Forest model

4.2 The Effectiveness of Law

Another goal in the project is to find out the effectiveness of law on reducing the firearm death rate. As above mentioned, we consider 5 factors in total: Income, Unemployment, Education,

Poverty, and Law. In order to only study the law, we need to get rid of the effects from other four factors: Income, Unemployment, Education, and Poverty. The effectiveness of law can be explored by finding similar counties or by the above regression models.

4.2.1 Analysis by Selecting Neighbors

First, we adopted principal component analysis (PCA). In this way, we were able to transform those four components into just two principal components. Figure 10 shows the result after PCA. PC1 and PC2 are two principal components.

Figure 10. Plot of data after PCA

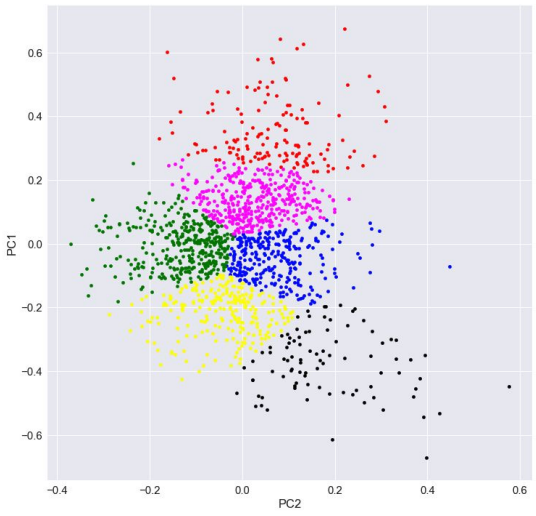

Figure 11. Plot of data after clustering

Then we apply k-means clustering method. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Here, we choose the number of clusters, k, as 5. Figure 6 shows the results after k-means clustering. Different colors represent different clusters. As we can see, the outcome is not satisfactory since most clusters have a very open boundary, meaning the elements (counties) within each cluster can still be very different.

We use another way to find similar counties. We can pick a data point from figure 5 and select its neighbors by K-NN algorithm. For example, Monroe county, NY is selected and the

nearest 250 neighbors of Monroe county is found. Using this method, we assume that we obtained 250 similar data points with respective to Income, Unemployment, Education, and Poverty.

With 250 similar data points in Income, Unemployment rate, Education, and Poverty, we move to the stage of exploring the relation between law and firearm death rate. The results are shown in figure 12. As we can see, the curved score for laws and firearm death rate have negative correlation with correlation coefficient about -0.27. The pink line shows linear regression results. It should be noted that linear regression doesn't accurately capture the relation between the two. Thus, at this stage, we would like to leave the regression for future improvement and just focusing the observation of the data. Based on our observation, we found that if we choose a column of the data point with same score of law, for example, 0.1, normalized firearm death rate is ranging from 0 to 0.3. It indicates two possibilities: First one is there still other factors that we failed to consider that affect the firearm death rate. Second one is that effect of law is not significant. Both of the possibilities are worth further studying in the future.

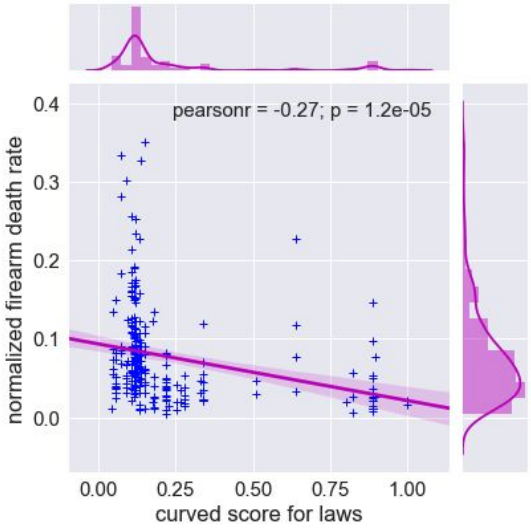

Figure 12. Relation between law and firearm death rate by K-NN method.

4.2.2 Analysis by Regression Models

The regression models can be used to predict the effects of laws. We use the real data for Income, Unemployment, Education, Poverty of each county and vary the curved score for law from 0 to 1 to predict the firearm death rate. Then we calculate the mean of firearm death rate for all available counties. The results are shown in Figure 13. By increasing the curved score for laws from 0 to 1, the average firearm death rate can be lowered about 60%. The two models have almost the same results. This is not surprising, since the prediction is an average of about 1500 counties.

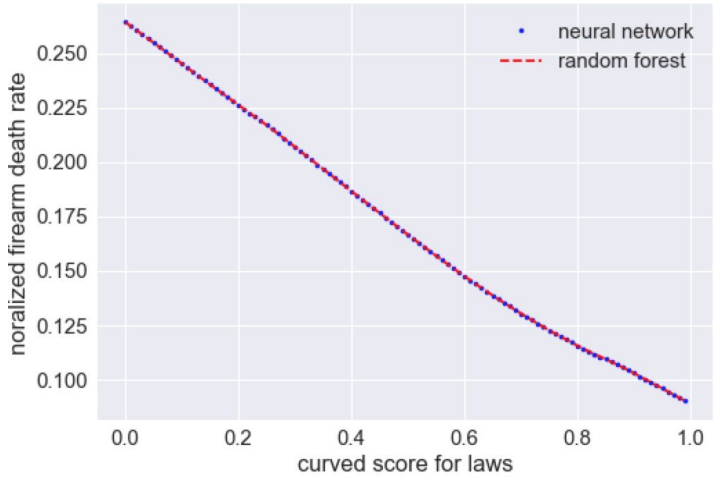

Figure 13. Relation between law and firearm death rate by regression models.

5. Conclusions

In this project, three algorithms were applied to predict the firearm rate at county level.

Naive Bayes gives an outcome of 88.927% accuracy which can be misunderstood as a good result initially. The correlation between different attributes is very strong, which breaks out the independent assumption of Naïve Bayes. The model fails under this scenario.

The other two regression models, Neural Network and Random Forest, have average mean absolute error of 0.0488 and 0.0425, respectively, corresponding to percentage error about 5%. The two models don't work well for counties have high firearm death rate because of

the skewed data set. What's more, the result shows that Random Forest model has large errors for some low firearm death rate counties.

The effectiveness of law on reducing firearm death rate is examined by two methods. For K-NN, the 250 neighbors of Monroe County, NY are selected, but we didn't find an appropriate method to model the relation between laws and firearm death rate. We then analyze the relationship using regression models, the two models show the averaged firearm death rate can be reduced by 60% if the normalized scores for laws is increased from 0 to 1.

6. Future Work

This project can be extended to perform better by including more attributes, such as averaging age of residents, population mobility, time averaged attributes, police officer per capita, etc. In this way, the model might be better trained and results might be better.

Further, we can check the relationship between law regulation and firearm death rate by picking other counties and selecting its neighbors in K-NN method. We can also examine the relation between firearm death rate with median household income, poverty rate, or education level.

Reference

[1] Mass shootings in the US are on the rise. What makes American men so dangerous? - Sociological Images: 2017. *https://thesocietypages.org/socimages/2015/12/31/mass-shootings-in-the-u-s-what-makes-so-many-american-men-dangerous/.* Accessed: 2017- 12- 14.

[2] Diebold, S. 2017. Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence in Chicago, 2006 to 2014. *The Journal of Emergency Medicine*. 53, 2 (2017), 281.

[3] Kalesan, B., Mobily, M., Keiser, O., Fagan, J. and Galea, S. 2016. Firearm legislation and firearm mortality in the USA: a cross-sectional, state-level study. *The Lancet.* 387, 10030 (2016), 1847-1855.

[4] Fleegler, E., Lee, L., Monuteaux, M., Hemenway, D. and Mannix, R. 2013. Firearm Legislation and Firearm-Related Fatalities in the United States. *JAMA Internal Medicine*. 173, 9 (2013), 732.

[5] Lee, L., Fleegler, E., Farrell, C., Avakame, E., Srinivasan, S., Hemenway, D. and Monuteaux, M. 2017. Firearm Laws and Firearm Homicides. *JAMA Internal Medicine.* 177, 1 (2017), 106.

[6] WISQARS (Web-based Injury Statistics Query and Reporting System)|Injury Center|CDC: 2017. *https://www.cdc.gov/injury/wisqars.* Accessed: 2017- 12- 14.

[7] USDA ERS - County-level Data Sets: 2017. *https://www.ers.usda.gov/data-products/county-level-data-sets/*. Accessed: 2017- 12- 14.

[8] Brady Campaign to Prevent Gun Violence : 2017. *https://www.bradycampaign.org/*. Accessed: 2017- 12- 14.