Aryan Deorah and Chris Goodhart

**Title:** Using Regression Machine Learning Models to Predict Gun Violence Death Rates Based on Gun Laws

**Problem:** The statistical relationship between gun laws and gun violence has not been analyzed yet using machine learning.

**Hypothesis:** If machine learning is applied to this subject, it will more accurately predict gun violence death rates than other statistical tests. Machine learning is much better than other statistical tests at processing many categories of data, and since the topic area has many categories of data, it would make sense that machine learning is more effective at predicting gun violence death rates. This research is done assuming that anything we are not controlling or accounting for in the research does not affect gun violence.

**Independent Variable:** The independent variable is the type of statistical test or model used to analyze the relevant data.

**Dependent Variable:** The dependent variable is the accuracy to which gun violence death rates are predicted.

**Constants:** There are no constants in social statistics.

**Control:** There will not be a control because there is no baseline statistical test that applies to this scenario.

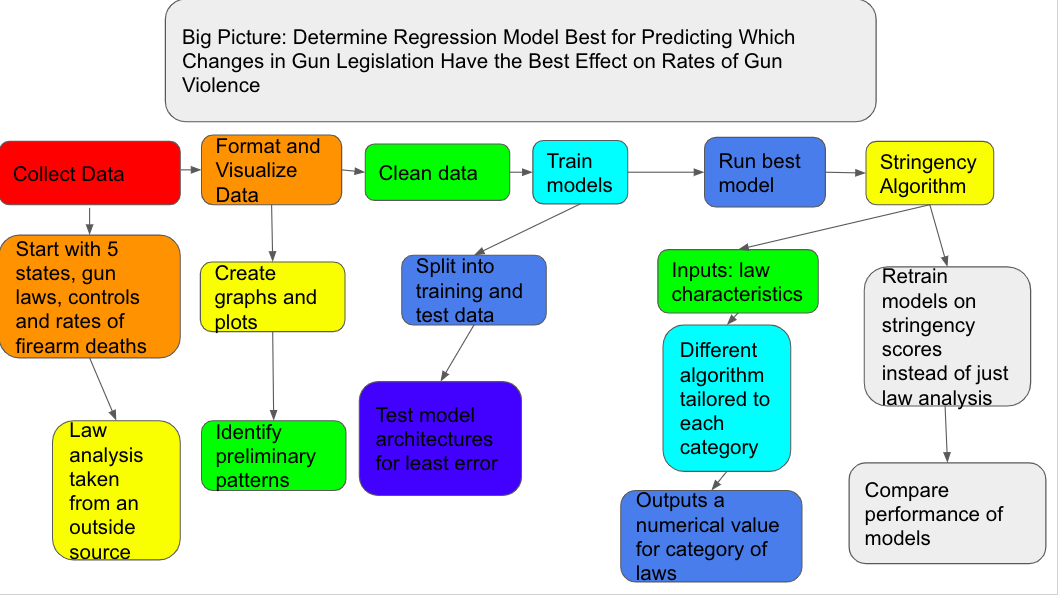
**Repeated Trials:** There are no trials that we are conducting because we are receiving our data from an outside source.

**Materials:**

* A computer
* Anaconda software downloaded onto computer
* Python software downloaded onto computer
* Weka software downloaded onto computer

**Procedure:**

1. Data on different types of gun violence (homicide, suicide, and accidents), already made gun law profiles for various states, and data on the various controls (unemployment rates, median incomes, violent crime rates, overall crime rates, per-capita gun ownership rates, race, and population density) will be collected from reliable government and/or non-profit sources.
2. Using the non-profit state firearm laws database, various laws will be sorted through to see which are the strongest and would supposedly have the greatest effect on gun violence. Once a handful of the strongest laws have been selected, relevant data regarding when and where a law would have been or is in place will be collected.
3. Next, the collected datasets will be combined and the data will go through the processes of data cleaning and feature engineering. This process may include determining where missing values are in the datasets, and whether to impute a value or drop a given column or row. This may also include changing non-numerical values, also known as objects, into numerical values. In addition to this, the gun violence death rate values will be controlled and modified according to the controls mentioned in step 1.
4. Exploratory data analysis on the data will be conducted using various data visualizations and softwares. Various graphs and plots, including but not limited to box plots, scatter plots, choropleth maps, violin plots, swarm plots, linear regression plots, KDE plots, and standard distributions will be used to analyze the data, find preliminary patterns, and help others understand the trends found. This will be done through various Python libraries, such as Seaborn, Pandas, Matplotlib, and Plotly, as well as a separate software called Weka.
5. The data will be split into test and training data as the first step of the model creation process. The models will be fitted and created based off of the training data and then their accuracy will be tested by determining how well their predicted values fit the test values.
6. A function for evaluating the effectiveness of various models predicting the gun violence death rates in a given area will be created in order to determine which model is best to use in the research.
7. Several relevant regression models will be selected and then evaluated based on the function mentioned in step six. The best one will be selected and fitted to the training data.
8. The model will be run and the errors will be analyzed and visualized. After this, we will repeat the process while varying various model characteristics and data.
9. For the next phase of the research, a stringency algorithm for the gun law profiles will be created in order to turn information about gun laws from a given state into a value to describe the strength of the gun laws for different categories (ammunition regulation, assault weapons and large-capacity magazines, buyer regulations, child access prevention, concealed carry permitting, dealer regulations, domestic violence, gun trafficking, immunity, possession regulations, preemption, prohibitions for high-risk gun possession, stand your ground, and background checks) of gun laws. Though stringency algorithms will be tailored to their given categories of gun laws, ideal inputs for the algorithms will be whether the laws exist in the first place and whether certain topics or phrases are used in the gun laws. This stringency algorithm will be based off of already existing stringency algorithms from research journals and non-profit organizations.
10. Using the algorithm from the previous step, the information from the law data set will be turned into stringency scores and steps 3 through 8 would be repeated using the stringency scores for categories instead of information on only the strongest laws.



**References:**

Dube, A., Dube, O., & García-Ponce, O. (2013). Cross-Border Spillover: U.S. Gun Laws and

Violence in Mexico. *American Political Science Review,* *107*(03), 397-417.

doi:10.1017/s0003055413000178

National Center for Health Statistics. (2018, January 10). Retrieved May 15, 2018, from

https://www.cdc.gov/nchs/pressroom/sosmap/firearm\_mortality/firearm.htm

Siegel, M., Ross, C. S., & King, C. (2013). The Relationship Between Gun Ownership and

Firearm Homicide Rates in the United States, 1981–2010. American Journal of Public Health, 103(11), 2098-2105. doi:10.2105/ajph.2013.301409

Aryan Deorah and Chris Goodhart

**Statistics:**

We plan to collect data on gun violence and gun laws in specific states across the United States. We plan to compare the stringency scores of different categories of laws and law analysis in a given state with the different types of gun violence death rates in that state. We plan to use multiple controls and these controls will be positive or negative based on what our research indicates. These controls are poverty and socioeconomic status, measured by unemployment rate and median income, general crime rates, measured by the violent crime rate and overall crime rate, the gun ownership rates, measured by per-capita gun ownership, and the population density, measured by people per a given area. We do not have any replicates for values because out data is received from external sources. We plan to randomize our sampling of the data by picking five states at first to analyze individually and then compare to each other. Below is an example table of inputs into a machine learning model for the research.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cause of Death** | **Year** | **Deaths** | **Population** | **Crude Rate** | **Age-Adjusted Rate** | **Stringency Score** | **Control** |
| Firearm | 2006 | 30896 | 298379912 | 10.35 | 10.2212447989244 |  |  |
| Firearm | 2007 | 31224 | 301231207 | 10.37 | 10.2377863536084 |  |  |
| Firearm | 2008 | 31593 | 304093966 | 10.39 | 10.2317483827028 |  |  |
| Firearm | 2009 | 31347 | 306771529 | 10.22 | 10.0492903471947 |  |  |
| Firearm | 2010 | 31672 | 308747508 | 10.26 | 10.0708021304884 |  |  |
| Firearm | 2011 | 32351 | 311663358 | 10.38 | 10.1612720202969 |  |  |
| Firearm | 2012 | 33563 | 313998379 | 10.69 | 10.4472458549102 |  |  |
| Firearm | 2013 | 33636 | 316204908 | 10.64 | 10.3766800564569 |  |  |
| Firearm | 2014 | 33594 | 318563456 | 10.55 | 10.2636648373972 |  |  |
| Firearm | 2015 | 36252 | 320896618 | 11.3 | 11.0283905731139 |  |  |
| Firearm | 2016 | 38658 | 323127513 | 11.96 | 11.726341196348 |  |  |

Regression machine learning models will be used because they are able to output continuous values and predict values or quantities of something, which will be most relevant to the project.