Project Title: Fatal Shootings by Police Officers in the United States: An in-depth analysis to understand the causes of fatal shootings by police and develop insights surrounding these tragic events

Authors : Connor Geiran, Keith Hetrick

Problem Statement

According to data collected by the Washington Post, American police have fatally shot 5,652 individuals since 2015. These shootings have galvanized the political discourse over the past several years. Much of the public discourse and media attention has focused on racism as a contributing factor to these fatal shootings.

Any death by the hands of police is a human tragedy. Often times, however, there are complicated factors at play leading up to a shooting by police. The Washington Post's dataset includes factors about whether the deceased citizen was armed or was displaying signs of mental illness. Subjects of fatal police shootings are often armed (56 percent) or have signs of mental health issues (23 percent). 64 percent of fatal shooting subjects were reported to have attacked the police officer before the shooting occurred.

American policing is mainly administered at the city or county level, rather than the state or federal level. Most police report directly to locally elected politicians like mayors or sheriffs, and not to their state's governor or the U.S. President. This decentralization means that local police in different counties and localities are truly unique from each other, unlike, say, branches of the military. If, for example, there were frequent shootings by military police on a U.S. Army base in Quantico, Virginia, it would be hard to reason that the socio-economic factors of Quantico, Virginia were an influential factor since the U.S. Army is organized on the federal level and recruits military police and servicemen from around the United States. However, if local police in Fairfax, Virginia were frequently fatally shooting local citizens, the thinking would likely be different. Local police forces are, for the most part, comprised of local citizens responsible for policing other local citizens. This suggests that socio-economic and public health indicators of a given locality could influence the behaviors and actions of the local police and their neighbors.

This decentralized structure leads to natural questions: Are there factors that explain why some localities have shootings every year while other localities with similar populations do not? What are the most influential aspects of these communities that may be key indicators or a blueprint for a negative outcome regarding fatal police shootings? This paper aims to discover the key factors contributing to fatal police shootings, which will in turn shed light on why these shootings occur more frequently in some localities versus others.

Data Sources

1. Fatal police shootings – the Washington Post collected detailed, location-based data on fatal police shootings occurring between 2015 to 2020. This dataset provides the dependent variable for our multi-pronged analytic approach.
2. County Health Rankings (CHS) - CHS data contains hundreds of factors including, but not limited to, income, demographics, health, crime, and education.

3. US General Elections – the MIT Election Data and Science lab has compiled a dataset with voting return information for federal elections by FIPS code (county) which includes results for the 2012 and 2016 presidential elections.

Hypothesis

Racism, violence, poverty, crime, lack of education, and mental illness are complicated problems that have been measured on a local level. The hypothesis of this paper is that fatal shootings by police are symptoms of larger socio-economic problems in localized communities.

Methodology

A key goal of this paper is to develop an understanding of county level socio-economic and public health information that can be identified as influential elements to fatal police shootings. The first steps were data collection, transformation, and defining what constitutes an adverse locality (a locality with an unusually high number and frequency of fatal police shootings) according to historical fatal police shooting data. Then, the team utilized predictive analytics to create a model that demonstrated predictive power to classify adverse localities and subsequently analyzed the importance of the variables to the selected model. Finally, the team utilized network analysis to create a network of nodes/variables and edges/connections within the data to measure the influence of the nodes to the adverse communities. This comprehensive approach allows readers to gain an understanding of which specific aspect of a given variable is influential to communities that have historically experienced many fatal police shooting events.

Data Collecting and Preprocessing

1. Connecting the Datasets

The initial challenge of this project was to find a way to connect local socio-economic data to fatal police shooting data. Publicly available data called the County Health Rankings (CHS) provided data on hundreds of socio-economic and public health factors by Federal Information Processing Standard (FIPS) code. FIPS codes identify counties and incorporated cities and are the ideal medium to compare local communities. The Washington Post's Fatal Police Shootings dataset included the coordinates where each shooting took place but not the FIPS code. The FCC's Census API was utilized to convert these coordinates into FIPS codes. These datasets were then joined on the FIPS code.

2. Data Imputation

Several important public health variables in the CHS data had missing values. Features that were missing more than one-third of the total values were discarded from analysis. Features missing fewer than one-third of the values were imputed using the python function KNNImputer, where the missing value for a given feature was calculated by finding the average of the nearest two neighbors to the missing data point, or assumed to be zero based on the variable in question.

3. Size of locality

Local public health problems and law enforcement strategies vary widely based on population size. Thus, the counties and cities were segmented by population size to better understand fatal police shootings in three categories of American regions:

- Rural/Suburban – between 0-154,999 people
- Urban – between 155,000-1,000,000 people
- Metro – over 1,000,000 people

The team determined that urban counties (and incorporated cities) would be the focus of the analysis for two reasons. First, urban communities alone make up 49.4 percent of the population and 42.2 percent of all fatal police shootings. Secondly, urban communities capture a wider range of American localities (e.g., communities with both relatively smaller and larger populations) as opposed to metro or rural areas that would display more similar community sizes and yield more expected results for those areas. All the results in this report are from urban communities with populations between 155,000-1,000,000 people.

4. Required Calculations to classify good/non-adverse and bad/adverse counties

After thorough exploration, the team used two main calculated variables to classify the adverse and non-adverse counties in the data:

- Recurring annual shootings – This is a calculated field capturing the recurrence of fatal police shooting events on a yearly basis and serves as a means to demonstrate a consistent occurrence of fatal police shootings and not purely an anomalous occurrence.
- Fatal police shooting rate – This is a calculated field captured by summing the shootings for each county over six years and dividing by the population of each county. The result of the division is then multiplied by 100,000 to achieve the fatal police shooting rate per 100,000 people.

5. Determining the specifics of the bad/adverse class

The next step was to determine which localities demonstrated greater adversity related to fatal police shootings and classify those counties as an adverse county. The team determined that urban localities should be classified as adverse if they met two conditions (1) the locality experienced a fatal police shooting in at least four years (out of six) and (2) the locality experienced a fatal police shooting fatality rate greater than 2 per 100,000 population. The four-year parameter meant that shootings were recurring , not just an outlier event, and the rate signified that shooting occurrences were much higher than the median fatal police shooting rate in urban areas (1.44). Of the 361 urban counties, 126 counties were deemed adverse, which means they had four or more years with a fatal police shooting and a fatal police shooting rate greater than 2 per 100,000 residents. Please note that any reference to "adverse", "adverse fatal police shooting event", "fatal police shooting(s)", or phrases of the like throughout the paper intends to capture how the team has defined adversity/bad outcomes for counties within our data as described in this section.

6. Creating multiple datasets

The original dataset was all continuous variables that measured rates and percentages of various features. This data, made up of continuous fields, was modified to include categorical fields, specifically quantile calculations to bucket the different elements in the data to better understand key features identified in the analysis. The quantiles were leveraged to transform the data into a graph network to support the analysis and interpretability of the results.

Results and Evaluation

1. Experimentation & classification models

In order to develop greater confidence in the results of our analysis, multiple different classification models were evaluated and the predictive power of each to classify the data were compared. As part of the experimentation, the team assessed the variables in the data to get an understanding of how they may relate to each other. This included an examination of collinearity as well as exploratory analysis of the variables produced from Lasso regression. Next, hyperparameter tuning was completed for various models to optimize performance. Last, to achieve better performance scores, the team utilized Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalance in the minority class (adverse cases) by randomly increasing adverse class examples through replication. This process created a more even distribution of the two classes in the data.

After conducting experimentation on all the models, the focus shifted to finding the model with the highest Sensitivity Rate (True Positive Rate). This metric was chosen as it focuses on the model's ability to correctly classify the positive outcome, which in this case are the adverse counties. By using this as our key metric along with good scores for other metrics (Specificity, Precision, Accuracy) it was ensured that the focus of the model selected was its ability to classify the adverse cases correctly and give credence to the subsequent analysis of important variables to predicting the target variable. After extensive data transformation, variable assessments, and exploring various analytics techniques, Random Forest was selected as the champion model with a Sensitivity score of 94%. (See Figure 1 below for a table of all the models tested and related performance metrics obtained during model experimentation and development).

	Accuracy	Sensitivity	Specificity	Precision
Nearest Neighbors	0.648936	0.725490	0.558140	0.660714
Linear SVM	0.755319	0.862745	0.627907	0.733333
RBF SVM	0.457447	0.019608	0.976744	0.500000
Gaussian Process	0.712766	0.843137	0.558140	0.693548
Decision Tree	0.659574	0.803922	0.488372	0.650794
Random Forest	0.797872	0.941176	0.627907	0.750000
Neural Net	0.542553	1.000000	0.000000	0.542553
AdaBoost	0.755319	0.803922	0.697674	0.759259
Naive Bayes	0.723404	0.823529	0.604651	0.711864
QDA	0.787234	0.764706	0.813953	0.829787

Figure 1. Table of performance metrics from model experimentation

The Random Forest model's ability to classify the positive cases correctly (see Figure 2 for graphic of the random forest model visualized using 2 principle components) with a high Sensitivity for this data sample, along with the some of the key elements of the algorithm, gave the team reassurance to select it as the champion predictive model for our analysis.

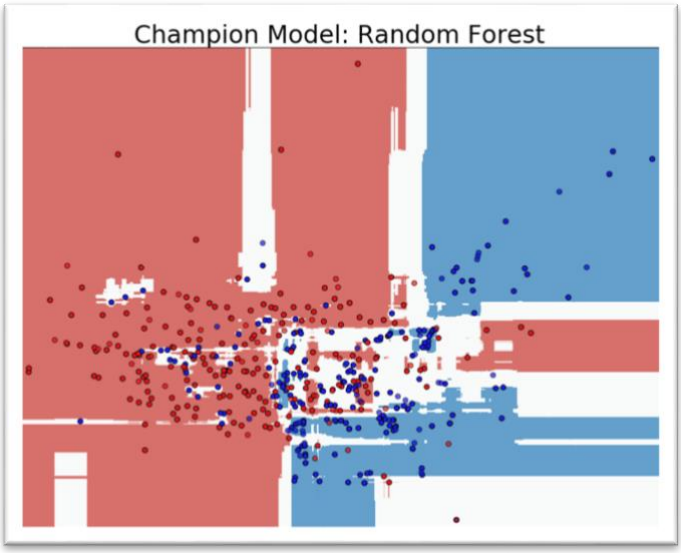

Figure 2. Champion Model Random Forest visualized using PCA top 2 principal components

The Random Forest is a robust model to use for our analysis because of its ability to handle high dimensional data and its use of an ensemble approach to model hundreds of decision trees to make predictions. For its creation of trees, the algorithm uses bagging, which aims to reduce the complexity of models that overfit the training data and reduce variance. Random Forest also uses feature randomness for each individual tree to try to create an uncorrelated forest of trees. The evaluation does not rely on one model, but each tree makes a prediction and the results are aggregated to classify the data based on the larger proportion of the predictions by each individual tree. (See Figure 3 for example depiction of a Random Forest)

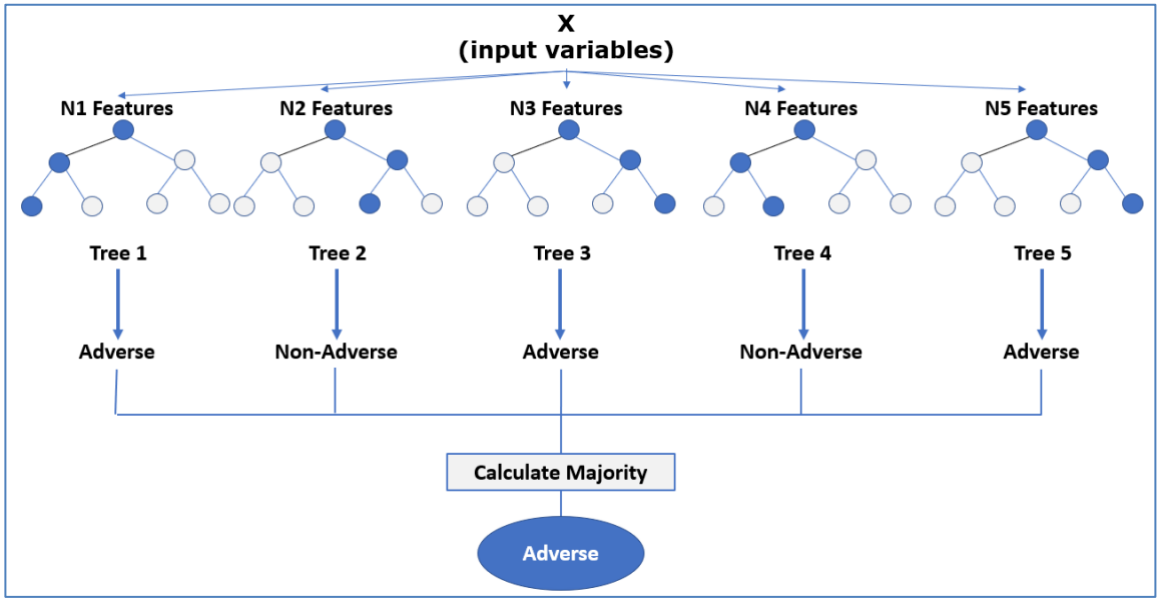

Figure 3. Example of ensemble of decision trees in a random forest and its final classification

2. Important Variables to the Random Forest model

Figure 4 and 5 display the important variables to the Random Forest model in descending order. The model derived that these variables are most important in explaining the target variable. The variable importance of Random Forest is calculated with Gini impurity which is a metric used to

determine how to split the data into smaller groups of similar values. Gini impurity measures how often a randomly chosen field in the data is incorrect which is essentially a measure of the probability of a new record being incorrectly classified at a given node in a decision tree within the forest. So, while training a tree, the model computes the impact that each feature contributes to decreasing the weighted impurity.

Random Forest Features Importance

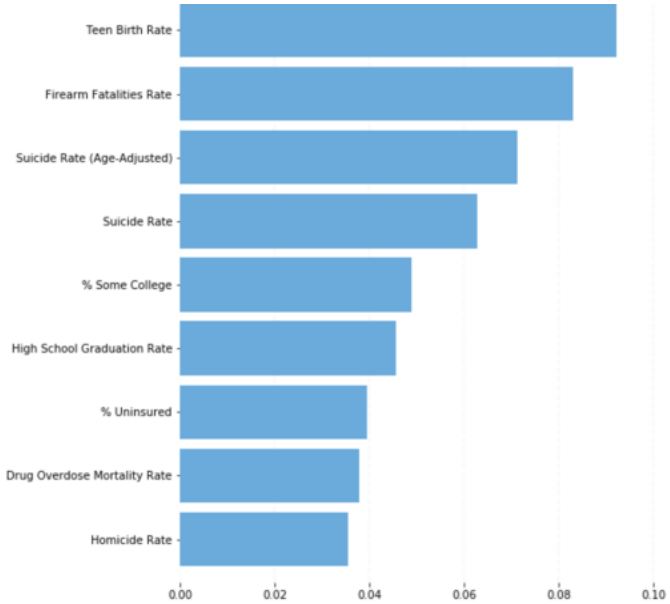

Figure 4. Top 9 Important Random Forest Features (Descending)

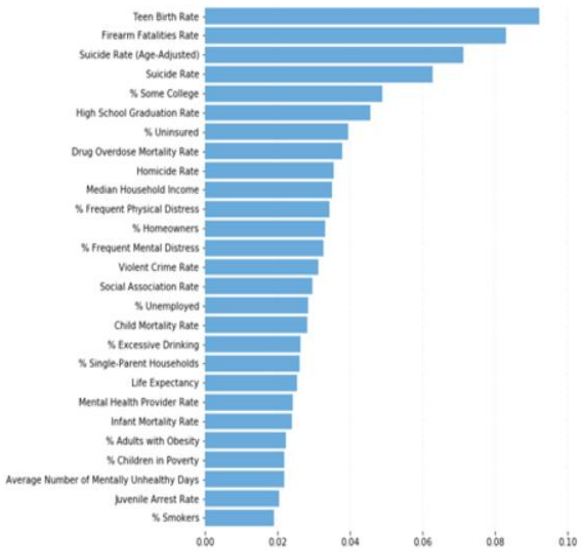

Figure 5. All Random Forest Features in Order of Importance (Descending)

3. Network Analysis

To further explore and understand the relationships of key features in the data, the team employed network analysis and calculation of centrality scores (Eigenvector centrality) to measure the influence of the variables/nodes to the adverse data population. Eigenvector centrality is the measure of influence of a node in a network and assigns a score to each node based on its connections to other highly connected nodes. If a node is connected to many important nodes, it will also be an important node. If a node is only connected to non-important nodes, then it will not be important.

Eigenvector Centrality (Eigencentrality):

$$x_i = \frac{1}{\lambda}\sum_{j \in M(i)} x_j$$

where $j \in M(i)$ means that the sum is over all j such that the nodes i, j are connected

To accomplish this, the data was first transformed into graph database schema (nodes: variables & edges: connections between the nodes) and developed into a network model. Second, the Eigencentrality measurements were calculated for each node within the dataset (both adverse and non-adverse counties) and within just the population of adverse counties (only adverse county data). Third, the centrality scores of the adverse nodes in the full data were calculated and all the nodes in the adverse data subset were compared to identify the nodes that demonstrated the

largest percentage increase in Eigencentrality as a measure for the influence of the node to the adverse outcome population. By doing this analysis, the team gained an understanding beyond only the well-connected nodes to identify the nodes that clearly demonstrate influence to the adverse population by comparing the its centrality score between the two networks. The results of this analysis combined with the variable importance obtained from the Random Forest model help us understand and more precisely interpret the key features prevalent to adverse counties in the dataset.

4. Network Analysis Influential Variables to the Adverse Class

Figure 6 displays the highest nodes from the adverse data population that result in the largest increase in Eigencentrality from the full network. The numbers corresponding to each of the nodes pertains to the quantiles for that variable in relation to the population (i.e. *Node (1)* indicates a low number and *Node (4)* indicates a high number)

Urban	
Node	**% Increase**
Mental Health Providers (1)	2.67
Firearm Fatalities (4)	2.10
Children in Poverty (4)	2.09
Median Household Income (1)	2.04
Some College (1)	1.94
Frequent Mental Distress (4)	1.90
Physically Unhealthy Days (4)	1.83
Mentally Unhealthy Days (4)	1.81
Teen Births (3)	1.77
Median Household Income (2)	1.75
Teen Births (4)	1.68
Smokers (4)	1.68
Drug Overdose Mortality (4)	1.68
Unemployed (4)	1.63
Homicide (4)	1.62
Suicide (4)	1.58
Some College (2)	1.51
High School (1)	1.50
Single Parent Home (1)	1.48
Uninsured (4)	1.44

Figure 6. Table of influential nodes to adverse data population obtained through comparison of Eigencentrality measurements between the adverse network and the full network (both adverse and non-adverse)

Final Results

Both the Random Forest and the Network Analysis indicate that the variable Firearm Fatalities is the second most important variable to their respective models. In fact, the network model results suggest that adverse fatal police shootings are occurring in localities experiencing a high number of firearm fatalities in the general population. This is an important finding as it may help to explain why police officers in some localities shoot to kill while police in other localities use non-lethal force. It seems possible that individual police officers serving in violent localities fear becoming a victim of a firearm fatality themselves. This fear may encourage police officers in

localities with high Firearm Fatalities to use lethal means towards a hostile citizen. This could be a significant factor in explaining fatal police shootings in certain urban localities.

The Network Analysis results show that a low rate of Mental Health Providers by population is the model's most influential feature. This means that the node for the lowest quantile of rate of Mental Health Providers demonstrated the largest percentage increase in Eigencentrality between the full and only adverse network. The Network Analysis' fifth most influential node, a high amount of Frequent Mental Distress in a region, and its seventh most influential node, a high number of Mentally Unhealthy Days for a region, similarly associates localities with numerous fatal police shootings to communities with many mental health issues. In the Random Forest model, the variables "Suicide Rate (Age-Adjusted)" and "Suicide Rate" were the third and fourth most influential variables respectively. In conjunction with the large increase in Eigencentrality in areas with higher numbers of suicides by population and related negative mental health indicators, it can be inferred that these elements proved to be a very influential factor within adverse counties. It is interesting to note that "Suicide by Cop" is a known suicide method where a person behaves in a threatening manner towards police with the intention of being fatally shot by a police officer. In one study[1], researchers identified over 250 cases of "Suicide by Cop" in a sample of 700 police shootings. In general, our findings tell us that mental health issues and/or lack of mental health support in adverse communities constitute localized public health crises in need of more awareness, understanding, and support.

A number of other influential nodes to adverse communities help to interpret the aspects of the important variables to the Random Forest model's ability to classify adverse communities. Particularly, high teen birth rate, lower rate of education, and fewer insured people demonstrate significance to adverse communities. From these findings, it can be inferred that adverse communities tend to experience a lack of education as well as a lack of health-related resources (which could include poor health awareness, unhealthy behaviors, and/or poor healthcare availability). Generally, these findings support the need to focus on elements of public health and general well-being for all communities, and particularly communities demonstrating deficiencies and experiencing adverse fatal police shooting events.

The connection of adverse fatal police shootings events with localities experiencing negative mental and physical health indicators and high firearm fatality indicators suggests that fatal police shootings are more of a symptom rather than the root cause of localized public health crises. Much more research is required to identify and understand the causes of firearm fatalities, poor public health, and causes of adverse police shooting events in particular communities. It is the hope of the authors of this report that the interconnected nature of these issues is acknowledged so that compressive action can help prevent future tragedies.

---

[1] Mohandie, Kris; Meloy, J. Reid; Collins, Peter I. (March 2009). "Suicide by Cop Among Officer-Involved Shooting Cases". Journal of Forensic Sciences. 54 (2): 456–462.