 CS109A: Final Project Report

Group #39 - Police Violence in the US

Team Members: Chika Okafor, Vasco Meerman, Matthew Parker, and David Koupaei

Harvard University - Fall 2020

Table of Contents

[1. Initial Problem Statement 1](#_Toc58093402)

[Motivations 2](#_Toc58093403)

[Problem Statement 3](#_Toc58093404)

[Data Source 3](#_Toc58093405)

[2. Visualizations and Data Summary 3](#_Toc58093406)

[Location Based Analysis 3](#_Toc58093407)

[Race Based Distributions 4](#_Toc58093408)

[Time Based Distributions 5](#_Toc58093409)

[A Baseline Model 5](#_Toc58093410)

[3. Further Modeling Approach 7](#_Toc58093411)

[Data Pre-Processing 7](#_Toc58093412)

[Modeling Plan 7](#_Toc58093413)

[4. Results 8](#_Toc58093414)

[Primary Model: Prediction of African American Race 8](#_Toc58093415)

[5. Conclusions 13](#_Toc58093416)

[High Level Results 13](#_Toc58093417)

[Extending Predictions 13](#_Toc58093418)

[6. Social Impact Study 14](#_Toc58093419)

[Reflections after Reviewing this Data 14](#_Toc58093420)

[7. References 14](#_Toc58093421)

# Initial Problem Statement

### Motivations

The main focus of this project is to take a deep dive into police violence datasets to answer key questions about racial equality and bias in the use of lethal force. In the past decade, social media platforms and other communication tools have enabled movements and widespread participation in this discussion, including Black Lives Matter. Particularly in the United States, there is political devisiveness about these topics, so this data analysis will attempt to transcend popular opinions to determine the extent to which bias exists in deadly police shooting events.

Excerpt from the project problem guidelines:

*“The experiences of many Americans - including Breonna Taylor, George Floyd, and Rodney King among, tragically, many others - speak to an epidemic of excessive and unjustified use of force by law enforcement officers whose sworn duty is to protect the communities they serve. Severe abuses of the police’s monopoly on violence, like extrajudicial killings, occur despite a constitutional presumption of innocence.”*

|  |  |
| --- | --- |
| Black Lives Matter is a social movement started in 2013 after the officer involved in the Trayvon Martin murder was acquitted. The underlying goal is to seek racial justice by eliminating systemic white supremacy and empowering local communities inflicted by violence from Police. The ideas have transformed into a worldwide movement putting racial equality at the forefront of discussions in popular media, news, and communities around the world. | Text  Description automatically generated |

The popularity of this movement has shifted over time, and has clear inflection points each time there is an unjustified murder, particularly of African American victims (*Figure 1*).

Table

Description automatically generated

*Figure 1: BLM Hashtag Popularity over Time*

The widely accepted opinion is that despite heightened awareness of this issue, protests, and requests for police and justice reform, there have been no significant changes to the rate of incidents involving use of lethal force by police over the last 5-10 years. The goal of this project will be to understand those relationships and determine some potential underlying causes of inequality.

### Problem Statement

The following questions were adjusted after significant EDA efforts:

1. Is the population-weighted incidence rate of police shootings different for white vs. black victims (vs. other races)?
2. How has the population-weighted incidence rate of police shootings shifted over the period 2015 to 2020?
   1. By race of victim
   2. By geographic region

### Data Source

After researching many available crime datasets, the team quickly realized that almost all projects referenced one of two key datasets: [Mapping Police Violence](https://mappingpoliceviolence.org/) and the [Washinpost's repo](https://github.com/washingtonpost/data-police-shootings). The team is confident these are the most thorough and trustworthy datasets publicly available.

**Washington Post:** In 2015, The Post began tracking more than a dozen details about each killing — including the race of the deceased, the circumstances of the shooting, whether the person was armed and whether the person was experiencing a mental-health crisis — by culling local news reports, law enforcement websites and social media, and by monitoring independent databases such as Killed by Police and Fatal Encounters.

**Mapping Police Violence:** This information has been meticulously sourced from the three largest, most comprehensive and impartial crowdsourced databases on police killings in the country: FatalEncounters.org, the U.S. Police Shootings Database and KilledbyPolice.net.

**Selection:** Washington Post sources their data from the mapping police violence dataset, but they cleaned the data more completely making it more useful for many required comparisons in this investigation. Therefore, the best data source to answer the proposed analysis questions is the Washington Post dataset.

# Visualizations and Data Summary

This analysis is broken down into various investigations to answer the problem statement question. A lot of effort was given to normalizing the data by population and by racial diversity using external data sources to help understand the bias in the incidents more clearly.

### Location Based Analysis

Initially it appears there is a large disparity with incidents in particular states. However, after correcting for state populations, there is a much more standard distribution of violent incidents. *Figure 2* looks at states which have the highest rates of violent police shootings.

![Chart, histogram

Description automatically generated]()

*Figure 2: States with Highest Rates of Fatal Police Shootings*

### Race Based Distributions

Race has a similar issue where the initial estimates revealed that whites were most frequently involved in violent incidents, but after correcting for racial population distributions, there is more bias toward minorities, particularly black victims. Whites are overwhelmingly involved based on total incident count, but they are the 2nd least likely to have a violent incident by percentage when normalizing the population.

![Chart, bar chart

Description automatically generated]()

*Figure 3: Frequency of Fatal Police Shootings by Race*

### Time Based Distributions

The other key item of interest for our model predictions is whether the trends have shifted over time, shown in *Figure 4*. Considering 2020 is a partial year, there is a very even number of total incidents over the last 5 years, but some trends arise when filtering by race, particularly a decrease for whites and an increase in unknown data, generally flat trend otherwise.

|  |
| --- |
| *Figure 4: Count of Police Killings over Time Sorted by Race* |

Discuss how a larger majority of the “unknown” category is made up of Latino population. Dig into why white is dropping. Possibly join/average 2019/2020

### A Baseline Model

For both of our modified questions, we will restructure the Washington Post dataset, so that each row corresponds with a unique state and year.  For each state-year, we will construct two outcome variables: (1) the total black civilian shootings per capita and (2) the total white civilian shootings per capita.

* **To answer Modified Question #1**, we will use a Paired T-Test, comparing the means for total black shootings per capita and total white shootings per capita
* **To answer Modified Question #2**, we will fit the following Linear Regression Model:
  + where denotes the total shootings per capita in state *s* during year *t* for race *r*. This expression will be evaluated at least for when race is white and when race is black
  + denotes the total shootings per capita in state *s* during year *t* for geographic region *r*
  + denotes fixed effects
  + denotes with time fixed effects
  + denotes the year. As 2015 is omitted from the regression specification, the coefficients are normalized such that the 2015 values are set to 0
  + are a series of state and year controls
  + The coefficient of interest is , which demonstrates the relative impact on the outcome variable for each subsequent corresponding year, compared to the normalized year of 2015.

**Results:** The table below graphically depicts estimates of the evolution in the police killing rate per capita over time, both by racial group and by region. These graphs follow the regression specifications under the Problem Statement #2. All estimates are calculated relative to the year 2015, which is normalized to 0. 2015 corresponds with the first full year after the heightened national dialogue on policing began (the lack of indictment for the police shooting of Michael Brown happened on Nov 24, 2014). All estimates are in log points. A log point estimate of *a* approximates a (100\**a*)% effect when the magnitude of *a* is small. In short, log points approximate percentage change for small changes in the independent variables. Vertical lines represent two standard errors.

The graph on the left in *Figure 5* provides evidence that the police killing rate per capita declined most over time for the white racial group, followed by the black racial group. In contrast, the police killing rate per capita *increased* for the “other” racial group. At the aggregate level, there was not a statistically significant shift in the overall shooting rate per capita over this time period. This suggests that the rate of police killings has remained relatively flat over this time period, but that the composition of the victims by racial groups has changed over time—away from white and black victims and toward the “other” racial categories (predominantly Latino). This shift might be explained by changes in police presence across communities, and is consistent with other findings that have found that the “Ferguson Effect” has led to lower police presence in predominantly black communities since the Ferguson uprising in 2014.

The graph on the right in *Figure 5* provides little evidence of regional differences in the evolution of the police killing rate over time. No geographic region has statistically significant differences from the baseline year of 2015, suggesting that the flat trend in the total levels (illustrated by the graph on the left) is matched by a flat trend at each geographical subregional level (illustrated by the graph on the right).

|  |  |
| --- | --- |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |

*Figure 5: Results of Baseline Model*

# Further Modeling Approach

### Data Pre-Processing

While the Washington Post data was generally well curated, there were a number of additional processing steps taken to assist in generating higher fidelity models.

**Cleaning Categorical Data**

Some of our predictors were categorized by text which had many unique values and was not grouped in any sensible way.

**Adding additional Predictors**

By using external data sources, we were able to add key predictors which should yield more insightful predictions

* Population Density: Used lat/long coordinates to identify location and subsequent population density
* Urban Classifier: Sorted population density by {urban, suburban, and rural}
* Political Party: Selected political affiliation of the state and year by election results {Red State, Blue State}
* Average Income: Used zipcode of incident to determine the median income in that area

**Encoding Categorical Predictors**

Key categorical parameters were One-Hot-Encoded using sk-learn to create Boolean predictors.

### Modeling Plan

The basic time dependent relationships with race ended up yielding no obvious correlations, so the analysis shifted to predicting other parameters which were key in determining the likelihood of police shooting occurrences.

The primary model developed to answer our key problem statements:

1. Prediction of African American race by all other predictors
   1. The goal here is to attempt to predict whether the victim of a fatal police shooting is likely to be African American based on all other predictors.

**Models attempted:**

* Logistic Regression
* Polynomial Feature Enhanced Logistic Regression
* Regularized Ridge and Lasso Regression with Cross Validation
* Random Forest
* Neural Network

Other models were investigated due to the broad nature of the datasets, but they will not be reviewed in full detail for this report.

1. Prediction of Mental Illness based on all other Washington Post columns
2. Prediction of Outcome of Judicial Process for all Fatal Incidents (was killing Justified?)
   1. Specifically are there any key predictors yielding differences in likelihood of the killing being considered justified by our judicial system.
3. Prediction of the number of fatal police shootings in 2020 using data from 2015-2019.

# Results

### Primary Model: Prediction of African American Race

###### Logistic Regression Model:

The first model attempted was a basic logistic regression model. This yielded decent results with **79.4% training accuracy and 78.0% testing accuracy**.

Add plot?

###### Lasso Regularized Logistic Regression Model:

The next was a modification to the basic logistic model by adding Lasso Regularization and Cross Validation techniques.

Key Predictors from Logistic Regression Model are shown in *Figure 6*. Some key takeaways from this feature importance analysis:

* City Density and Age are the top two predictors
* Louisiana, Maryland, California, Georgia, and Missouri are the top 5 states with most easily differentiate racial groups in police shooting events.
* Mental Illness is a better predictor than whether the suspect is armed with a gun

**Table

Description automatically generated**

*Figure 6: Logistic Relative Feature Importance*

*Figure 7* takes a closer look at our top predictors to see how the data are distributed.

Chart, scatter chart

Description automatically generated

*Figure 7: Review of Top 2 Classification Predictors*

There was a slight improvement in performance with **79.9% training accuracy and 78.6% testing accuracy**, using a hyperparameter C=1.0.

Chart, line chart

Description automatically generated

Figure : Lasso Regression Model Classification Accuracy Scores

###### Random Forest Model:

Based on the structure of the original dataset, we were hopeful that the decision tree classifier architecture would be more capable of achieving good prediction results.

Chart, line chart

Description automatically generated

*Figure 9: Cross Validation of Random Forest Tree Depth to Optimize Test Accuracy*

Results: By using cross validation to select the best tree depth (max\_depth = 20), the prediction capability improved over previous models at **88.0% training accuracy and 78.0% testing accuracy**. Note that while this is not the maximum test accuracy achieve, this depth is a good balance between high test accuracy without overfitting too much.

###### Neural Network Model:

The last model attempted was a neural network with the following parameters:

* 1 input layer
* 2 hidden layers with ‘relu’ activation function
* 1 output later with sigmoid activation function

The model was fit with 50 epochs using ‘binary\_crossentropy’ as the loss function and ‘adam’ as the optimizer function.

Results: *Figure 10* explores the loss function and prediction accuracy of the model, indicating that 50 epochs is sufficient to reach stability of the model. The overall prediction capability is average compared to other models at **76.2% training accuracy and 74.6% testing accuracy**.

Chart, line chart

Description automatically generated

*Figure 10: Loss Function and Accuracy of Neural Network*

**Review of Model Prediction Accuracy:**

After attempting each model, the Logistic Regression model utilizing Lasso regularization and Cross Validation was determined to be the most effective, by highest accuracy score in the test data, see *Figure 11*.

Chart

Description automatically generated

*Figure 11: Race Model Accuracy Scores*

Chart, line chart, scatter chart

Description automatically generated

Figure : Receiver Operating Characteristic for Race Models

# Conclusions

### High Level Results

The project initially intended to show clear trends between police shootings over time, including differences between races. However, the EDA and modeling process yielded no definitive time-based or race-based connection with the occurrence of a deadly police shooting.

Answers to the problem statement questions:

1. *Is the population-weighted incidence rate of police shootings different for white vs. black victims (vs. other races)?*
   1. Yes, there is a higher probability of fatal police shooting occurrence for African Americans than other races. Add statistics
2. *Has the population-weighted incidence rate of police shootings shifted over the period 2015 to 2020?*
   1. No, there is no statistically significant difference over time between the distribution of police violence when controlling for population based on racial group or other key predictors
   2. This was predominantly answered during the EDA process and initial baseline model building. After showing that no strong time based relationships, the team focused heavily on answering question 1 with more detailed predictive models to assess the various types of racial bias.

### Extending Predictions

While the results of the project didn’t align directly with our initial predictions, there were some other very interesting conclusions through modeling different types of predictions.

1. We are able to predict mental illness of a victim based on other data predictors (how could this be used in real-time?)
2. The mapping police violence data can produce an adequate model to predict justification of the police shooting (was the officer charged, acquitted, etc.) based on other raw and derived predictors related to population density.
   1. The two key predictors in these models is signs of mental illness and Year
      1. Expand on mental illness playing a role
      2. While the original baseline model doesn’t show a clear time-based trend in the number of killings, this analysis suggests there is a difference over time on how the cases are handled after the event. (discuss how this could related to change in dash-cams/availability of information)

# Social Impact Study

### Reflections after Reviewing this Data

At a high level, it’s very clear that a racial bias still exists in the United States based on police killings and our response to those events. Furthermore, the bias appears to be present in many of the other predictors like location, mental condition of the victim, age of the victim, and other characteristics of the event.

Despite the heightened awareness of issues like Black Lives Matter, there has been no appreciable change in the disproportionate police killings of African Americans.

Add more thoughts

# References

1. Black Lives Matter: <https://blacklivesmatter.com/>
2. Reference data for black lives matter hashtag usage on Twitter: <https://www.pewresearch.org/internet/2018/07/11/an-analysis-of-blacklivesmatter-and-other-twitter-hashtags-related-to-political-or-social-issues/>
3. Population Density Estimates: <https://simplemaps.com/data/us-cities>
4. Urban/Suburban Classifier: <https://medium.com/codait/got-zip-code-data-prep-it-for-analytics-7022b47652d9>
5. Race Classifications:
   1. Original Set: <https://github.com/washingtonpost/data-police-shootings>
   2. Additional Categorical from Wikipedia: <https://en.wikipedia.org/wiki/Race_and_ethnicity_in_the_United_States#Racial_categories>
   3. 2019 Estimates from Census Bureau: <https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>