Analyzing Fatality of US Police Shootings

ORIE 4741 Project Final Report

Joshua Fraser (jdf254), Ayman Naji (an443), Laura Gonzalez (lg458)

December 13th, 2020

I. Background:

The deaths of Breonna Taylor, George Floyd and other black victims at the hands of police brutality ignited national dissent over the unfounded use of violence by police on unarmed and unresisting black individuals that resulted in their untimely deaths. Over 4.6 million Americans channeled their anger into attending protests, while also voicing grievances to local and federal government, raising money for affected families, and deepening social movements advocating against police brutality and racially motivated violence against black people such as the Black Lives Matter Movement.

It has been argued that systemic, interpersonal, internalized and institutional racism lies behind the racial disparity among police shootings. Others deny the existence of a disparity or cite other factors as influencing the disparity. Data shows that over the past five years there has been no reduction in the racial disparity in fatal police shooting victims despite increased use of body cameras and closer media scrutiny, according to a new report by researchers at Yale and the University of Pennsylvania. Through databases compiled by The Washington Post, researchers have found that victims identifying as Black, Indigenous, or People of Color (BIPOC), whether armed or unarmed, had significantly higher death rates compared with whites.

To contribute to this conversation and narrative in a data-driven way, we aim to explore the many features of a police shooting to determine if certain features can give insight into the extent and nature of the racial disparity. Specifically, we will explore the predictive power of three different models using various demographic and situational features.

II. Data:

Data Description:

The dataset consists of 4,400 police shootings from 2010-2016. It contains descriptive data about the victims such as age, gender and race. In terms of how the shooting transpired, the dataset includes the date and city of the event, nature of event, as well as the race and gender of the responding police officer(s). Due to the variation in transparency and methods of record-keeping for the law enforcement agencies, the dataset has some missing entries and inconsistencies that would have enriched the analysis further. Additionally, the source had to make certain subjective assumptions in cases where data may have been lacking or unclear. For example, in cases where the number of individuals involved in the shooting was unknown, the dataset assumed there to be one victim. If an officer said a subject was reaching for his gun, is the subject “armed”? What if the gun was found at the scene after a shooting? Conclusions were made by Vice News in these cases. Nonetheless, this dataset is highly comprehensive as it includes features not found in other datasets, and is distinguished by the inclusion of non-fatal

shootings. This inclusion of non-fatal shootings data can allow us to predict whether a shooting was fatal or not based on the features in the dataset.

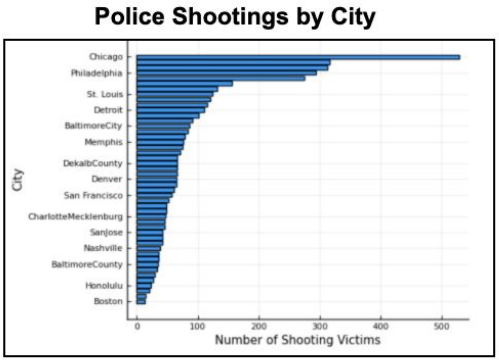
Categories of Data

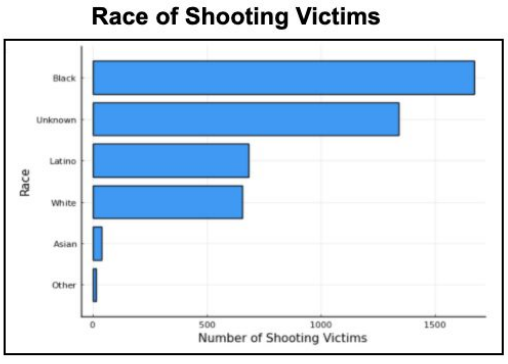
Date (String)**	NumberOfOfficers (Int64)*
NumberOfSubjects (Int64)	OfficerRace (String)*
Fatal (String)**	OfficerGender (String)*
SubjectArmed (String)**	Department (String)*
SubjectRace (String)**	FullNarrative (String)*
SubjectGender (String)**	City (String)*
SubjectAge (String)**	Notes (String)*
NatureOfStop (String)*	Year (Int64)***
NumberOfShots (String)**	

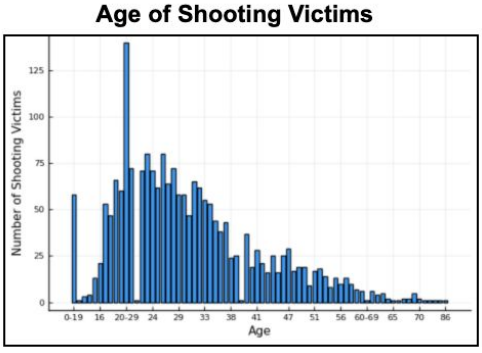
**Categories not used, **Encoded & Converted to Int64, ***Added after formatting*

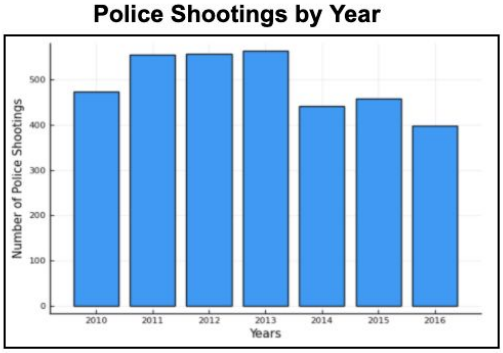
Data Visualization

In our dataset, Chicago had by far the most police shooting events with 12% of all police shootings in the dataset occuring in the city. The majority of police shooting victims were black with 38%. 62% of subjects involved in police shootings died and only 38% of all subjects involved in police shootings were armed. The age of subjects involved in police shootings varied but the majority of subjects were between 20-30 years old. The number of police shootings that occurred was fairly evenly distributed and the most police shootings occurred in 2013, while fewest occurred in 2016.









Data Cleaning

Many features in our dataset had missing or incomplete entries. We started by deleting features that had information that could not be used in our model such as “NatureOfStop”, “NumberOfOfficers”, and “Notes”. We then decided to focus on the features with the least amount of missing values and in a format that was easy to use in our models. We chose to focus on the features “Fatal”, “SubjectArmed”, “SubjectRace”, “SubjectGender”, “SubjectAge”, “NumberOfShots”, and “Year”. We chose these features because we found that they were the ones that had the greatest impact on the fatality of US police shootings.

Before reformatting the data, we had to delete missing/corrupted data entries to make it usable for our models.These features contained entries with “U” listed, which means the information is unable to be determined and “NA” listed, which means the data is missing. We corrected this by removing the entries. In beginning the reformatting process, we eliminated inconsistencies in the logging of dates such as “month/day/year” and “day/month/year” by extracting the month and year information and discarding the day since it was not functional for our models. Filler zeros in the month and day (i.e: 09 vs 9) were removed as well. Ultimately, the month and date column was removed due to the fact that it was resulting in overfitting for our loss functions. As for the SubjectAge feature, some entries recorded as ranges like “20-29” or “juvenile”, for cases where certain police departments did not know the subject’s precise age. We reformatted these entries by calculating the mean of any age ranges and defining juvenile to be 15-- which was a subjective assumption we made. For NumberOfShots, some entries represented the amount of shots as an inequality such as “>/=5” which we changed to be the integer in the inequality. After reformatting, data in type String was converted to type Int64 for the features utilized in the models. The unused features were left as strings.

The 7 features “Fatal”, “SubjectArmed”, “SubjectRace”, “SubjectGender”, “SubjectAge”, “NumberOfShots”, and “Year” that were the main focus of the project so we disgraded all other columns before our analysis. Across many columns there were also extra blank spaces in entries, which weren’t eliminated, but rather accounted for in the preprocessing. Once reformatting and data cleaning was completed, the dataset size diminished to 1028 data points. We were then able to establish a 80/20 ratio for the training and testing set respectively.

Preprocessing

For the SubjectArmed, SubjectRace, Fatal and SubjectGender features we utilized the one hot/many hot encoding methods. This method is defined as the process by which categorical variables are converted into a format that machine learning programs can use for prediction. For the SubjectArmed feature (“Was the subject armed?”) , "N" was encoded as 0, “Y” was encoded as 1, “U” meaning “unknown” was encoded as 2, and “NA” was encoded as 2. In the case of SubjectRace, we assigned integers that corresponded to the different races as follows: 1 is “White”, 2 is “Black”, 3 is “Asian”, 4 is “Latino”, and 0 is “Other”. Furthermore, for the Fatal feature, “N” was encoded as -1, “Y” as 1, and “U” as 0. The SubjectGender feature was encoded such that 0 is “U”, “NA” and “N/A”, 1 is “Male”, and 2 is “Female”.

III. Problem:

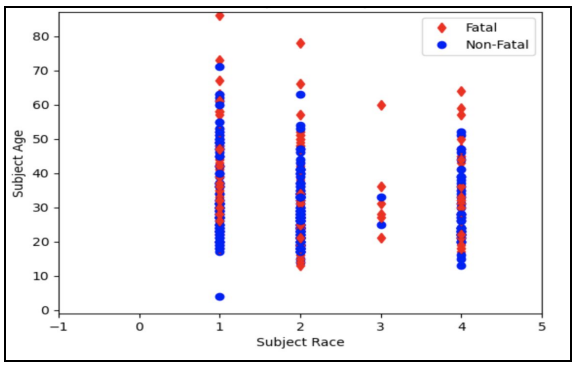
Our aim is to determine which of our features (after data cleaning and preprocessing) are the best predictor of whether a police shooting is fatal or not, in addition to examining the predictive power of polynomial regression models and loss functions. We expect to see that SubjectRace plays a significant role in the ability to predict the fatality of police shootings in a way that is reflective of commonly known statistics about police shootings.

IV. Analysis:

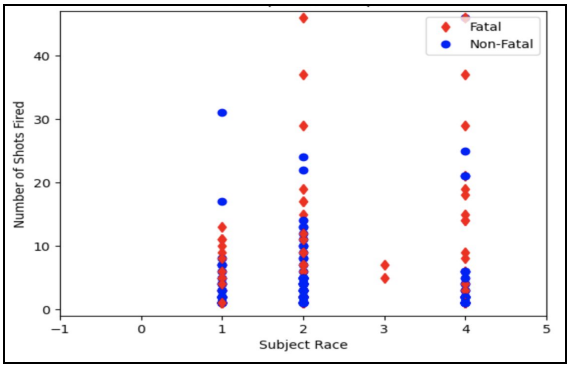
1. Perceptron Algorithm

The perceptron model is a linear classification algorithm that divides data into two subsets. The perceptron algorithm can only fully work if the data used is linearly separable. The first model we ran on our dataset was the perceptron model to show how the features in our dataset impacted fatality. We first ran this model to see the relationship between race, age, and fatality. We then ran the model to see the relationship between race, number of shots fired, and fatality. Our data is not ideal for this algorithm because our data is grouped by a set number for subject race in both graphs. This model is valuable because it will help determine if our data is linearly separable and knowing this helps determine the best method for predicting the fatality of police shootings.

Race vs. Age:



Race vs. Number of Shots Fired



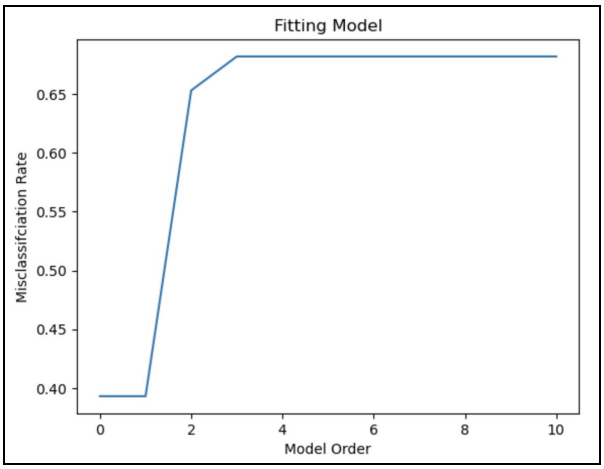
The perceptron shows that our data is not linearly separable because the fatal and non-fatal data points overlap. In order for our data to be linearly separable there is some hyperplane that separates the data into positive and negative examples. If our data was linearly separable then some specific age ranges or types of race would alway have been involved in fatal or non-fatal shootings. This proves that race, age, age or number of shots fired alone can determine whether a police shooting will be fatal or not. This model shows that in order to predict the outcome of police shootings we must perform other more complex models.

2. Polynomial Models

Our second approach was to fit a polynomial model across various features in our dataset to measure the risk of fatality in a police shooting. The first feature we decided to measure against fatality was the subject's race, to investigate whether there was any bias exhibited by the police

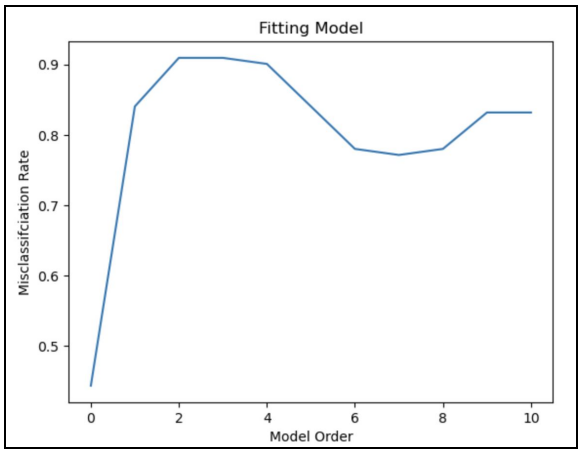
officers in each scenario. We stored our optimized features in a vandermonde matrix, and then fit models for orders 1 through 10, calculating the misclassification rate for each model and graphing it below.

Subject Race:



According to our results for subject race, the polynomial model that made the most sense with respect to our dataset was the second order model. A parabola best fit makes sense with respect to how we encoded our subject races (1=White, 2=Black, 3=Asian, 4=Latino), considering it would have a minimum point correlating to a subject's race being asian, which was the smallest demographic out of all races in our dataset. Given that the model with the lowest misclassification rate was the 0th order model within our output, we are unable to make adequate predictions on the risk of fatality with regards to a subject's race. This may be because our dataset lacked enough entries for each respective race in order to get accurate results in measuring this bias with respect to fatality.

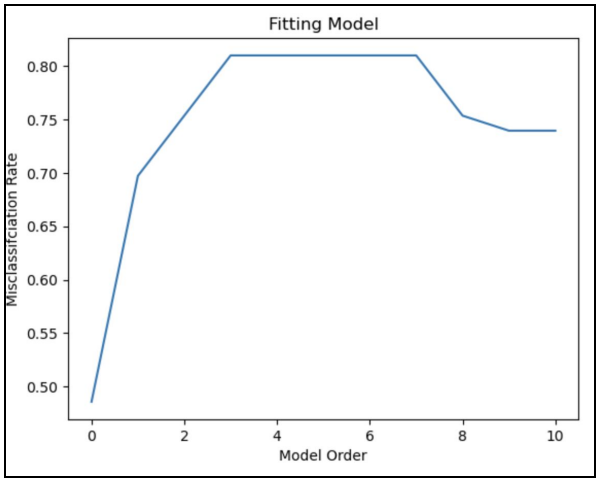
Subject Age:



The next feature we decided to measure against fatality was the subject's age. Considering the second best model was the 7th order, we assumed this was overfitting our data so it was not a

valuable indicator for our predictions. Since the 0th order was again the model with the smallest misclassification rate this tells us that the subject's age was also not a valuable predictor in the fatality of the police encounter.

Number of Shots Fired:



The last feature we measured was the number of shots fired against the probability of fatality. As we expected, our second best model was the 1st order, which states that the fatality of a shooting was potentially linear to the number of shots fired. However, since our results state again that the 0th order model was our best fit, it shows that the number of shots fired are also impractical with accurately predicting fatality. This may be because the number of shots fired are not equivalent to the number of shots that actually hit a subject, considering police officers tend to be inaccurate shooters so there may be a lot of missed bullets in most of these encounters.

Overall, our polynomial models are just initial assessments in our predictions for the fatality of police shootings, and do not provide us with definite conclusions since each model only accounts for one predictor variable (feature) with respect to fatality, We will now look into loss functions in order to account for multiple features as our predictor variables to get better predictions for fatalities.

3. Model Selection

Our final approach in analyzing the data was to use different loss models to classify whether or not a police shooting would be fatal based on our formatted dataset. Given the seven features we finalized as our input space, we experimented with different loss functions and regularizers, establishing parameters from our training set and then using them to calculate the misclassification rates on our test set.

The first model we ran was a least squares problem, a quadratic loss function with no regularizers, which produced a misclassification rate of 0.3566, We also ran a quadratic loss function with L_1 and quadratic regularizers, which gave us misclassification rates of 0.3411 and 0.3566 respectively. Next, we ran logistic and hinge loss functions (with no regularizers) since

they fit best for our classification problem, and reported misclassification rates of 0.3527 and 0.3643 respectively. After that, we used a Support Vector Machine model to avoid overfitting by running logistic and hinge losses with quadratic regularizers, and reported misclassification rates of 0.3372 and 0.3643 respectively. Finally, we ran L_1 regularizers on our data to introduce sparsity in our parameters, considering police officer(s) tend to react differently based on a given scenario for each shooting. Keeping this in mind, we wanted to see which one of these features played the most significant role in predicting police shooting fatality. After running logistic & hinge loss functions with L_1 regularizers we found that the only coefficients that were nonzero for both losses were the *SubjectAge* and *Year* parameters, and we included the model coefficients with respect to these two loss functions in the appendix below (see *Figure 1*). The misclassification rates for the logistic loss - L_1 regularizer and hinge loss - L_1 regularizer were 0.3527 and 0.3643 respectively, and we reported all our misclassification rates for each loss function and regularizer in the appendix as well (see *Figure 2*).

Note: After running the loss function models with the inclusion of the Month, Year, and Date1 features originally, we found that removing the Month and Date1 features prevented overfitting and therefore led to our decision to remove them from the set of relevant features aforementioned. This, however, was at the expense of the misclassification rate which rose (≈ 0.1 for every loss function).

V. Conclusion:

The goal of this project was to determine whether a US police shooting would end in a fatality based on our predetermined features, and which ones would have the greatest impact on this prediction. Using the perceptron algorithm we were able to determine that no one feature was able to predict the outcome of police shootings given our data is not linearly separable. Moving onto our polynomial models, we did not find a specific feature that was able to generate a function in predicting fatality amongst our dataset, which shows that these scenarios are too unpredictable to accurately classify a shooting's fatality based on a subject's race, age, or number of shots fired. Finally, our loss models were our best approach in analyzing fatality, considering it used every feature in our dataset combined in order to predict whether or not a police shooting was deemed fatal. We found that the *Logistic Loss with quadratic regularizer* produced the lowest misclassification rate **(0.3372)** compared to the rest of our loss functions, and that the ***subject's age*** and ***year*** of the shooting had the greatest effect in predicting the fatality of a police shooting as they produced significant model coefficients for certain loss models. This makes sense considering with respect to age, we found that according to Yale News, the average age for all victims is 34 with the average ages by race being 30 for Black people, 33 for Hispanics/Latinos, 31 for Native Americans, and 38 for White people. In addition, the risk of being killed by police peaks between the ages of 20 and 35 for men and women and for all racial and ethnic groups, as established by a Proceedings of the National Academy of Sciences of the United States of America (PNAS) study. It is evident that real world statistics corroborate with year showing one of the greatest effects in predicting fatality. Furthermore, we found that according to the Bureau of Justice Statistics, crimes tended to be higher in the summer than during other seasons of the year. Seasonal patterns existed in household larceny

and burglary victimization rates with the summer having the highest rate of household crimes. This pattern is reflected in police shooting fatalities because the higher the rate of crime, the more likely it is for a police shooting to transpire. This is in line with our ability to obtain significant model coefficients for the Logistic Loss & Quad Loss with L_1 regularizers for these two features. Our project is most likely not a weapon of math destruction because the outcome in question —fatality— is easily measured in the sense that there is no ambiguity in whether a subject was killed or not at the time in which a police shooting transpired. Predicting whether US police shootings are fatal would not endanger people, but rather help people by elucidating the general public on the heightened risk of younger individuals that may be classified as “at-risk” and therefore more likely to commit a crime. This information can also be valuable for institutions in our society such as schools, universities, and police departments in devising solutions for reducing police shootings that function on a systemic level. Our prediction would not create a feedback loop because we are collecting data from every reported police shooting that occurred between 2010-2016.

VI. Future Improvements:

In the future, we would like to further analyze the effect of location on the fatality of US police shootings. To perform this analysis we would most likely need additional location information than what our dataset provided. This relationship between these features would be valuable to explore because it was clear that some cities had significantly more police shootings than others. We would also like to explore the relationship between subject race and police officer race. There is the perception that white police officers are harsher toward black subjects and this would be an interesting theory to explore further. Our dataset was missing a significant amount of information for the race of the police officers involved in the shooting and thus we would need additional information to include this variable in our models. After listening to guest lecturer Rich Caruana explain EBM models and how they are equipped to handle bias in predictions, we think this model would have been valuable to explore in our project if we had more time and learned about EBM models earlier.

Moreover, the project could be improved by finding a dataset that has less subjective influences, missing data, and ambiguities in how the data was recorded on a department basis. We chose the Vice News dataset because it was one of the only available online that included both fatal and nonfatal shootings. This, however, meant that Vice News acted as an intermediary between the police department and the data users (us), which introduced subjective assumptions into the dataset that we manipulated. This issue is hard to avoid because most police departments do not make this information readily available to the public. Vice News explains the difficulties they faced when compiling the comprehensive data directly from the police departments, therefore lack of transparency by the source is a factor that could explain model inaccuracies. Additionally, there were many missing entries caused by the scant records kept by some police departments and ambiguities (such as the age being a range) that introduce inaccuracy and lowered precision in our models. In an ideal world, police departments would be fully transparent and place more exigency on logging extensive reports for the benefit of researchers, data scientists, media, and the general public.

VII. Appendix

Feature Names	Logistic Loss L_1 Reg	Quad Loss L_1 Reg
Subject Armed	0.0	0.0
Subject Race	0.0	0.0
Subject Gender	0.0	0.0
Subject Age	**0.0136**	**0.0101**
Number of Shots	0.0	0.0
Year	**-0.000346**	**-0.000187**

Figure 1: Model Coefficients by Feature

Model	Misclassification Rate
Quadratic Loss - No Reg (Least Squares)	0.3566
Quad Loss - L_1 Reg	0.3411
Quad Loss - Quad Reg	0.3566
Logistic Loss - No Reg	0.3527
Logistic Loss - L_1 Reg	0.3527
Logistic Loss - Quad Reg	**0.3372**
Hinge Loss - No Reg	0.3643
Hinge Loss - L_1 Reg	0.3643
Hinge Loss - Quad Reg	0.3643

Figure 2: Model Misclassification Rates

IX. Bibliography:

1. Arthur, Rob. “Get Data on Nonfatal and Fatal Police Shootings in the 50 Largest U.S. Police Departments.” VICE, 17 Dec. 2017, www.vice.com/en/article/a3jjpa/nonfatal-police-shootings-data

2. Belloni, Alexandre, et al. “An Efficient Rescaled Perceptron Algorithm for Conic Systems.” Mathematics of Operations Research, vol. 34, no. 3, 2009, pp. 621–641. JSTOR, www.jstor.org/stable/40538435

3. Home.” Black Lives Matter, 18 Nov. 2020, blacklivesmatter.com

4. Belli, Brita. “Racial Disparity in Police Shootings Unchanged over 5 Years.” YaleNews, 29 Oct. 2020, news.yale.edu/2020/10/27/racial-disparity-police-shootings-unchanged-over-5-years

5. McLaughlin, Eliott C. “How George Floyd's Death Ignited a Racial Reckoning That Shows No Signs of Slowing Down.” CNN, Cable News Network, 9 Aug. 2020, www.cnn.com/2020/08/09/us/george-floyd-protests-different-why/index.html

6. Edwards, Frank, et al. “Risk of Being Killed by Police Use of Force in the United States by Age, Race–Ethnicity, and Sex.” PNAS, National Academy of Sciences, 20 Aug. 2019, www.pnas.org/content/116/34/16793

7. Lauritsen, Janet L. “Seasonal Patterns In Criminal Victimization Trends.” Bureau of Justice Statistics (BJS), 17 June 2014, www.bjs.gov/index.cfm?ty=pbdetail