Fatal police shooting …

Is it a civil right violation of citizens by abusing power or an act of self-defense?

This project is about fatal police shootings in the United States between 2013 and 2018.

Our main focus here is on features of each shooting and the demographics of the cities and states that shootings took place, in order to find patterns or the possible relation between shootings and the ethnicity background, gender and the age bracket of the deceased.

FBI and the Center for Disease Control and Prevention log fatal police shootings, but they acknowledged that their data is incomplete. I hoped to work with official data, but in the absence of official records I went with the second choice which was unofficial resources. Dataset for this project was accumulated mostly from Washington Post database.

Washington Post as the main source of data reveals a selection bias for our dataset. Deeper look into audiences/ readers of Washington Post shows lack of diversity, as the great portion of the readers can be considered educated middle-class people in big or mid size cities. It’s also fair to assume that not all of audiences/ readers were aware of the campaign of collecting records for fatal police shootings. So clearly there were incidents that were never reported to Washington Post.

At last, we need to clarify what sort of incidents were covered in this report. This dataset only includes incidents that civilians were fatally shot by active duty police officers.

Challenges:

Missing Data:

Like any other real-world dataset, dealing with missing data was inevitable. My approach to missing data was different, depending on the category:

Missing Names:

Missing or invalid names were removed. The logic behind the decision was if the name wasn’t collected, other information such as Age and Race couldn’t be reliable.

Missing Age:

Filled missing age values of each ethnicity with the mean of the same ethnicity group, in order to prevent outliers in each ethnicity group from influencing and skewing the whole data.

Missing Race:

I researched internet to find articles or news about the shooting incident to identify the ethnicity of the deceased. For very few incidents I couldn’t find any useful clue in regard to their ethnicity, so I guessed based on the names.

Missing values for Threat Level, Body Camera and Flee:

Filling the missing values with Logistic Regression and KNN algorithms.

Outliers:

Outliers showed up in the Age column. Under 13 years old and over 70 years old initially was considered as outliers. But after a little research it turned out these are true outliers, so were not removed from the dataset.

Inconsistency:

Like any other real-world data dataset has inconsistencies in spelling city names caused problem in counting total number of shooting in each city and also merging tables. In order to mitigate inconsistencies, applied “r” library to remove unnecessary terms like: CDP, County, township …

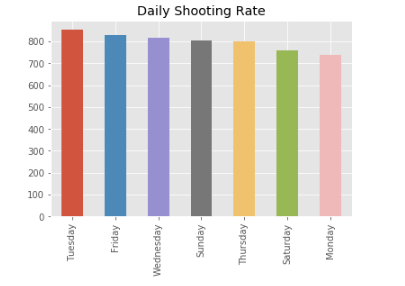
Whatever couldn’t be corrected with r library was corrected manually.

Observations:

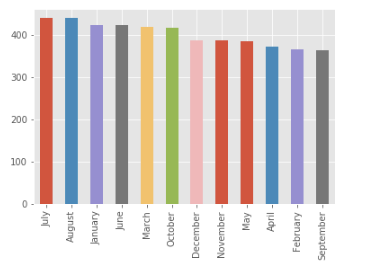
Let’s see what data tells us about:

* Timing:

“It’s all about timing”, but is it really? Initially, before starting the project I was expecting to see a different pattern of killings on the weekends, as usually weekends are more eventful. People usually drink more during the weekends, so we assumed there might be more shootings over the weekends than other weekdays. Well, interestingly our data did not support this theory as weekdays have higher rate of shootings compare to weekends.

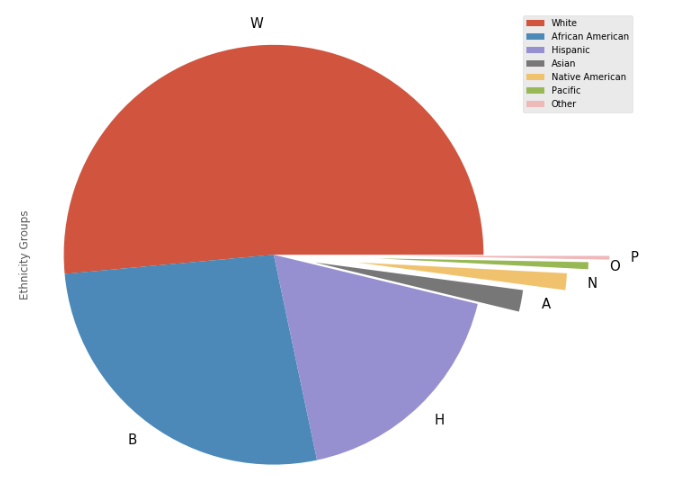


Other observation related to timing was the months of the year with lowest and highest rate of shootings. Initially, I was expecting to see the least rate of shooting in December as people are more in holiday spirit and more forgiving, but interestingly the month with the lowest rate of shooting turned out to be September and highest rate turned out to be on July.



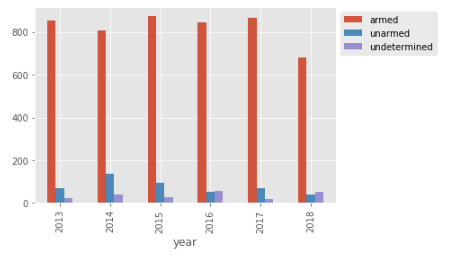
* Ethnicity Background:

Quick look at ethnicity background of deceased, shows highest portion of deceased are among Whites, African Americans and Hispanics.



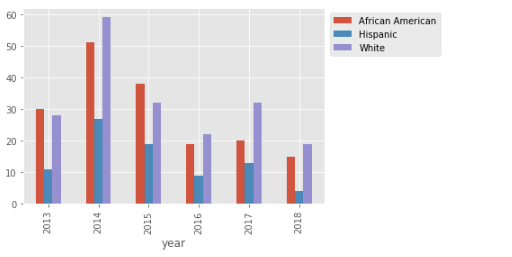
* Armed Vs. Unarmed:

As expected, being armed and carrying weapon increase the chance of getting fatally shot and killed by the officer.

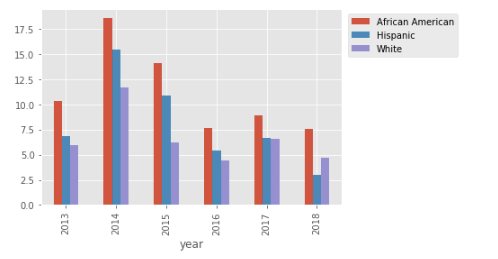


* Percentage Vs. Raw Data:

One of the interesting parts of the project for me was checking the ratio of Armed deceased to Unarmed deceased. Raw data shows with the exception of 2013 and 2015 in other years White ethnicity is dominating other ethnicities in Unarmed deceased:

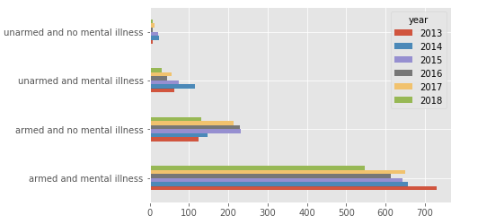


But if you take into account what portion of the deceased in each ethnicity was unarmed we see a different picture. The portion of unarmed white civilians who were shot and killed by on duty officers is less than African Americans and Hispanics.



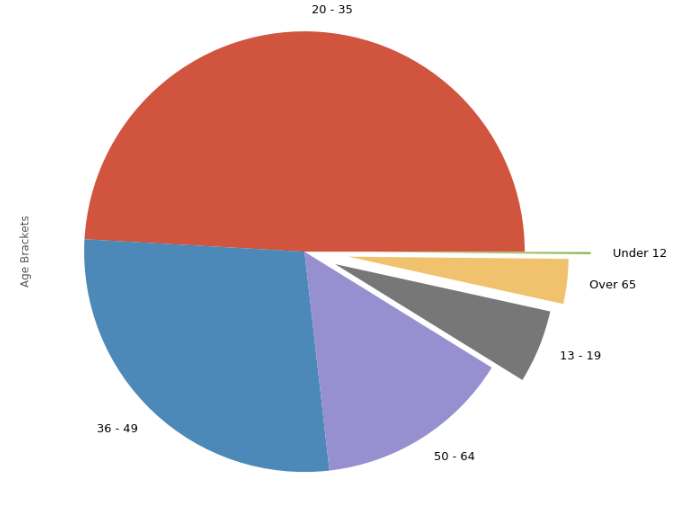
* Mental Illness:

Looking at Mental illness data, implies showing sign of mental illness by itself does not increase the risk of getting fatally shot by the police. As you see in below graph civilians who are being armed are still in bigger risk of getting shot.

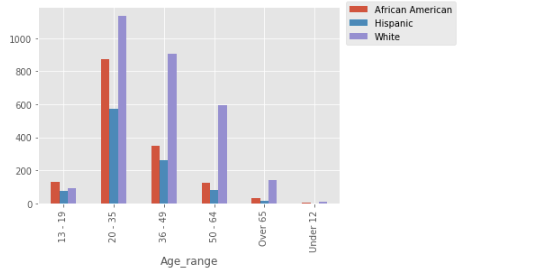


* Age Bracket:

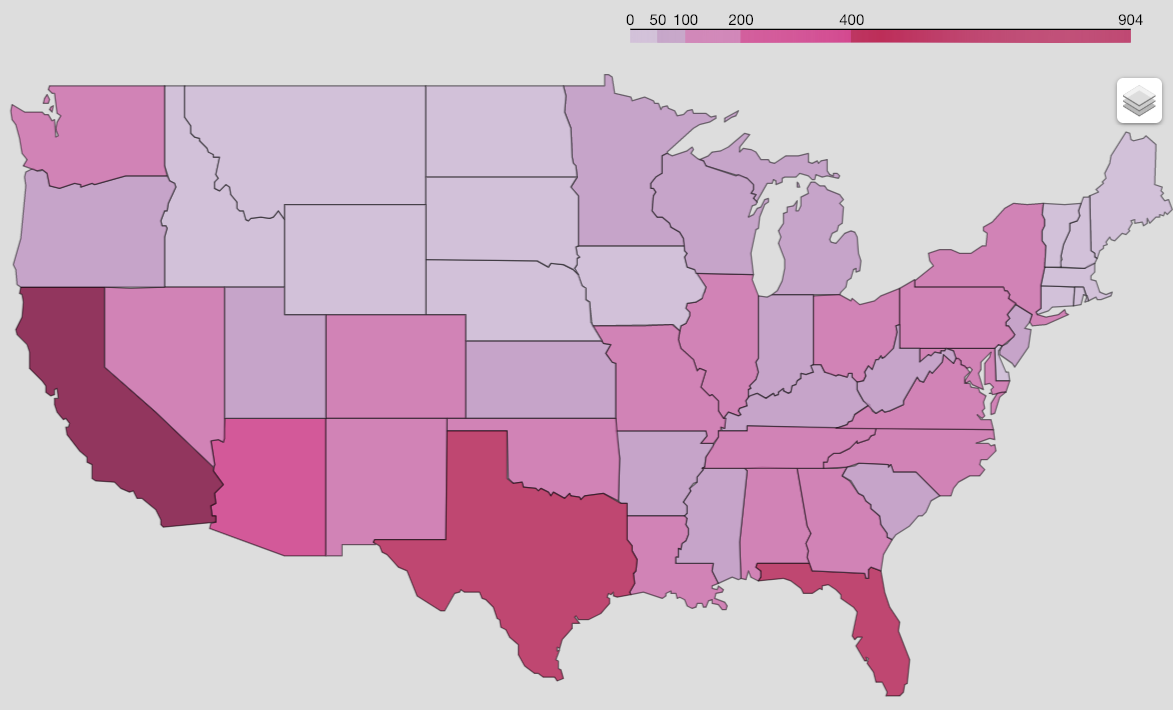
The highest portion of deceased were among young adults, age bracket of 20 to 35.



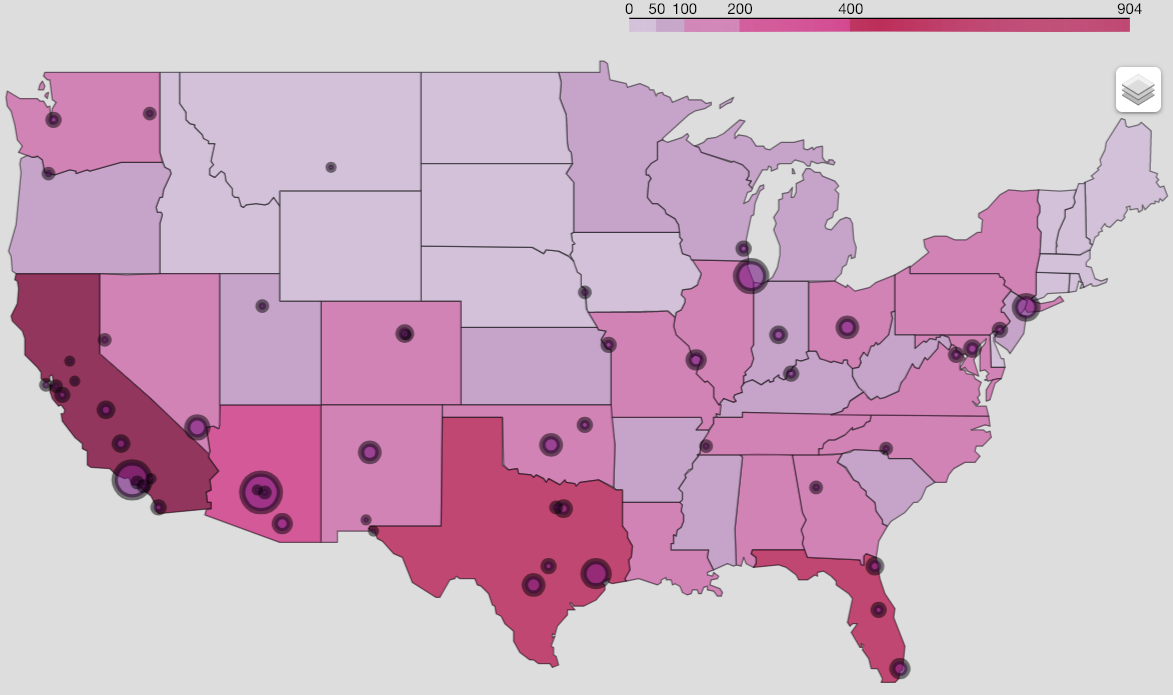
Looking closely at top three ethnicities, shows their highest number of deceased were in age bracket of 20-35.



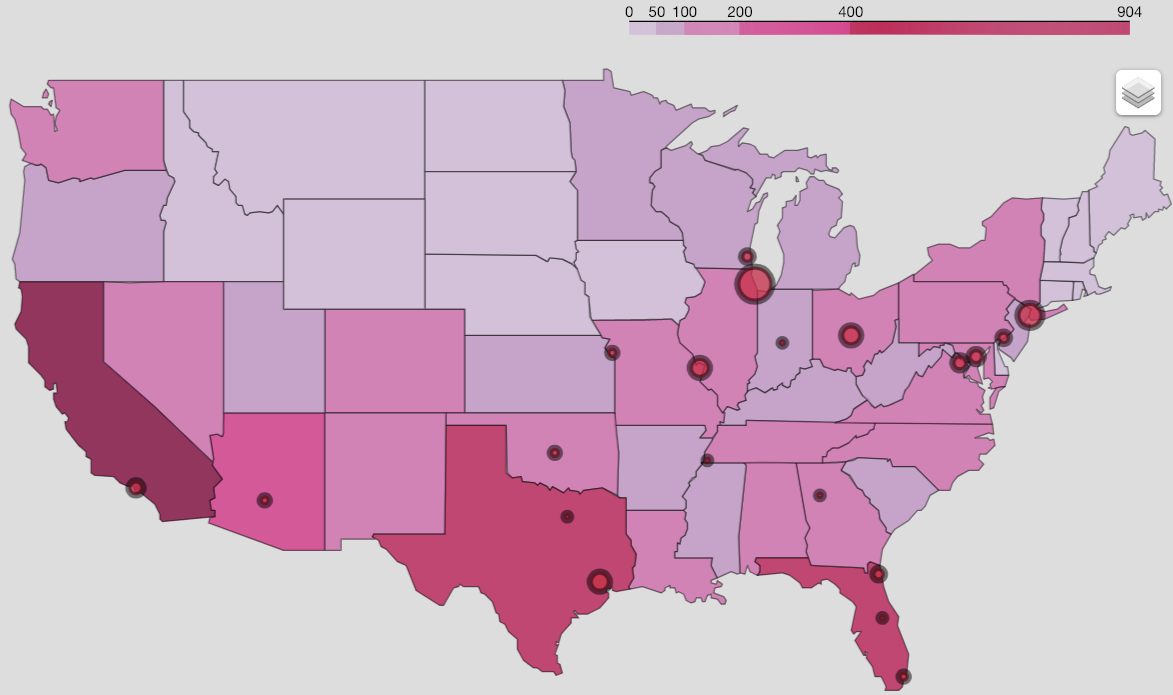
* Geographic distribution of fatal shootings:



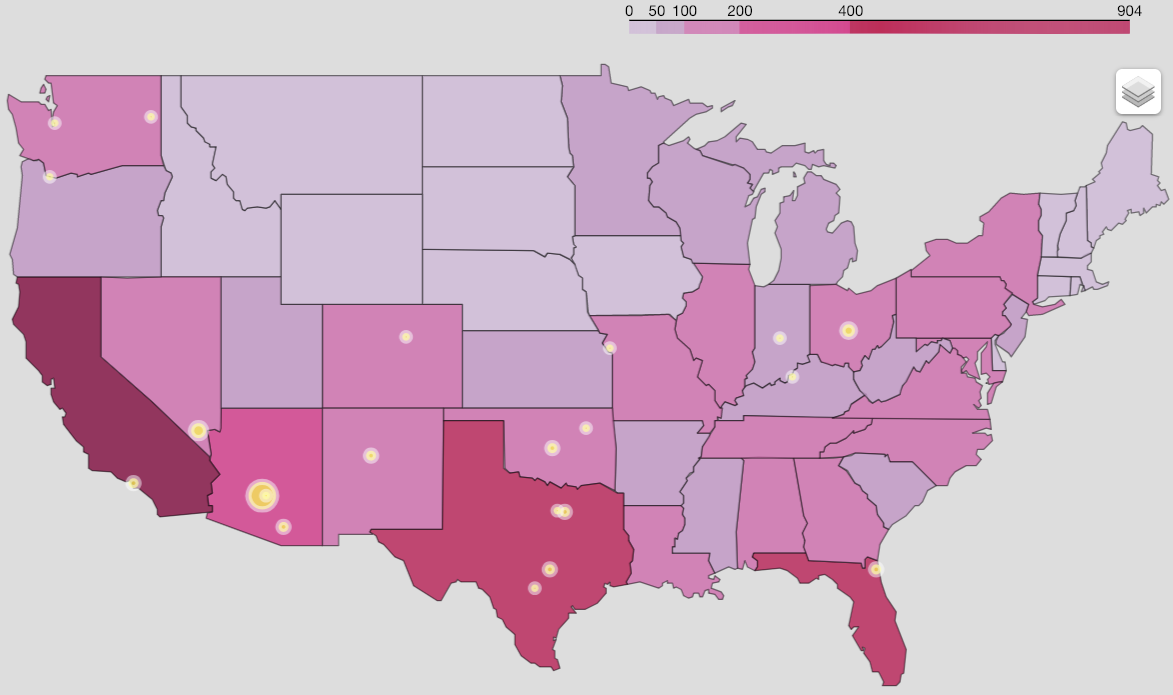
* Cities with overall highest number of fatal shootings:



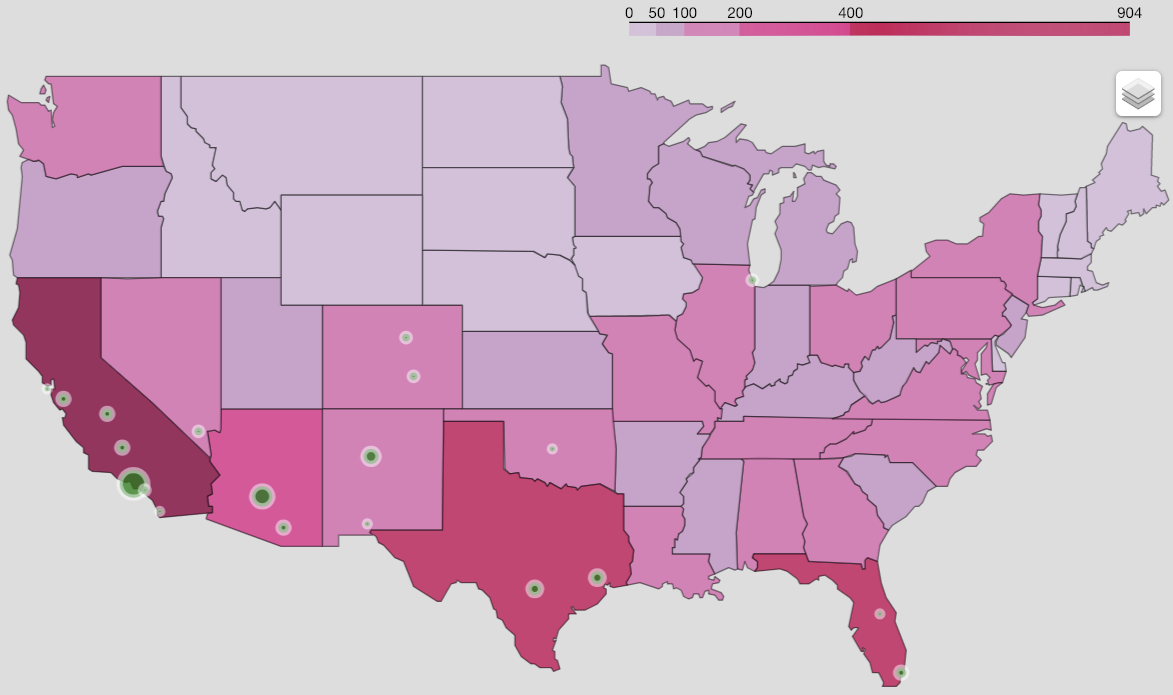
* Cities with the highest rate of shooting African Americans:



* Cities with the highest rate of shooting Whites:

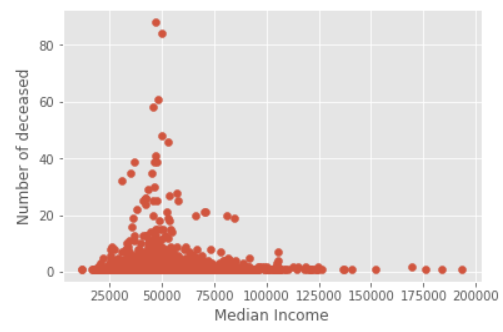


* Cities with the highest rate of shooting Hispanics:



* Economy:

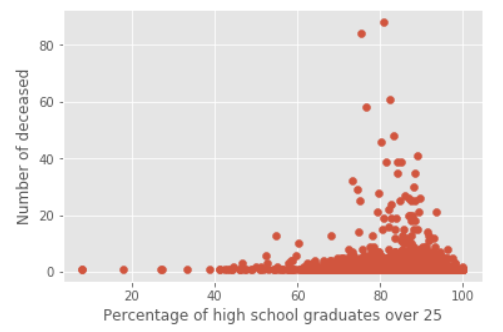
One of the main questions for this project was assessing the effect of economy on the rate of fatal shooting. The speculation of having higher rate of shootings in lower income cities or poor economies wasn't validated by this data. I wasn't able to find any direct relation between poverty percentage and median income of the cities with the number of fatal shootings that took place in them. Perhaps if we had more detailed data, such as zip code that shooting took place, we could go in depth and evaluate the economy situation of those zip codes. Big Metropolitans are economically diverse, and high-income neighborhoods can affect the total Median income of the cities. While some of these shootings could have taken place in lower income neighborhoods.





* Education:

Another main question was possible relationship between education level and number of the deceased in each city. Data did not support any relation between Percentage of high school graduates and number of the deceased.



Now it’s time to be more sophisticated!

Before studying the Inferential Statistics, I was a firm believer in numbers and what my set of observations represent to me. I trusted my eyes maybe a little bit too much. While scrubbing through my dataset to find possible current patterns in my dataset I overlooked the possibility that some of these patterns and results are unique to this dataset and possible changes in size or scale of the dataset may significantly affect the patterns and conclusions that I drew from the current dataset.

I also initially discarded the circumstances that could have effects on observations but were never taken into account or included in the original dataset.

Well, I learned my lesson the hard way (literally hard way, Inferential Statistics was a true challenge for me): Sometimes what you see is not enough.

Correlated incidents and patterns that we found in the current dataset could be due to chance. For Example, my dataset only includes data between 2013-2018, expanding the dataset to include before 2013 fatal Police shooting information could affect the patterns and correlations that we found in the current dataset.

When it comes to drawing conclusions, rejecting or accepting Null Hypothesis, it’s important to take into consideration the features or factors that could affect our conclusions but are possibly missing from the dataset.

Having said all these, let’s see what we found after looking at this dataset through the compound lenses of inferential statistics.

In studying the deceased information in order to determine if there are any possible correlations between the Ethnicity background of the deceased and their Gender, Sign of mental illness and the Age of deceased, we found out the P-Values for all mentioned categories are very small (less than 0.01) and that indicates there were statistically significant difference between Ethnicity background with Sign of mental illness or Gender or Age.

In the other word, the fact that we didn't find any correlations between these categories in our dataset wasn’t due to chance and the possibility of finding correlation between these categories in general is low, less than 1% and in some cases almost 0.

Was that all?

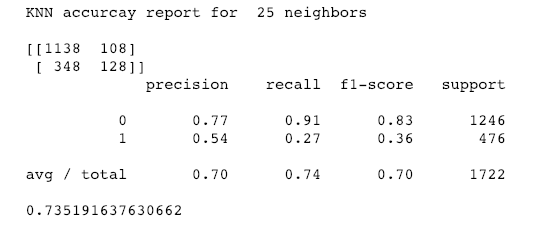
No! So far the focus was to see the picture that our data paint for us and listen to the story it tells us. But it doesn't stop there the main goal of Machine Learning this sexy, most wanted skill is to predict future based on past and help us see what's coming and be prepared for it.

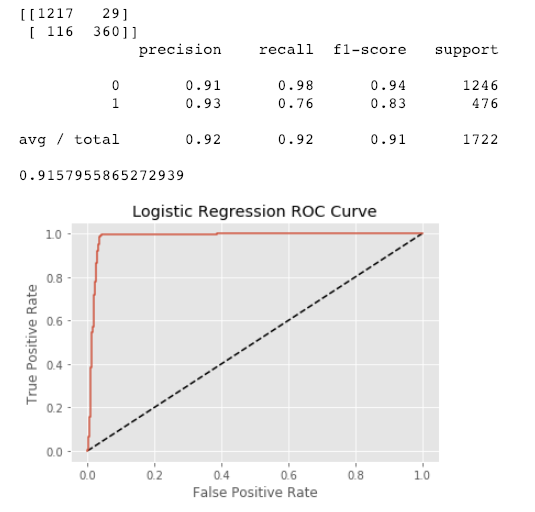
Based on the current dataset we can predict below:

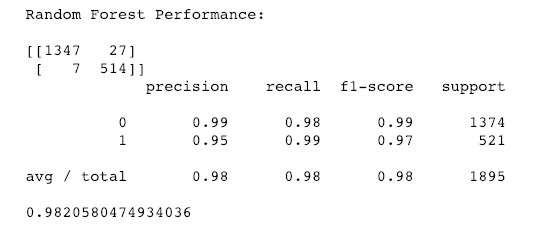
**1- Predicting the race of the deceased and determining whether it was black or not:**

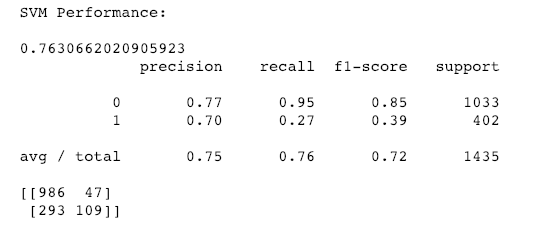
Problem is a classification problem, so we need to use classification algorithms such as KNN, Logistic Regression, SVM, Random Forest and Naive Bayes and compare their results to see which one provides a better result.

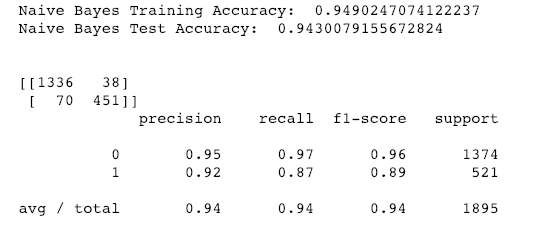
Since we have an imbalance classes, it’s necessary to check confusion matrix and classification reports in addition to accuracy to check Precision and Recalls finding how many of our positive and negative predictions are TRUE.

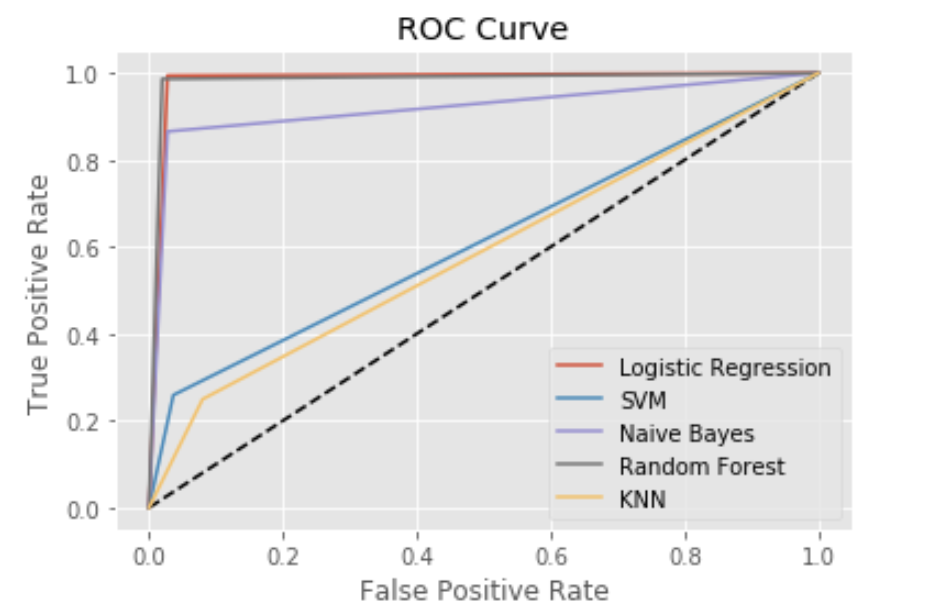








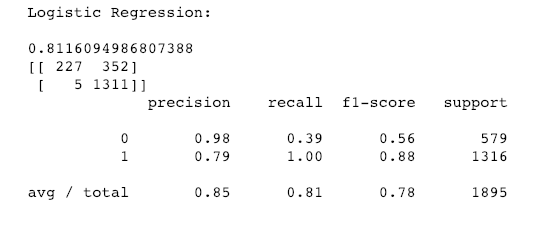




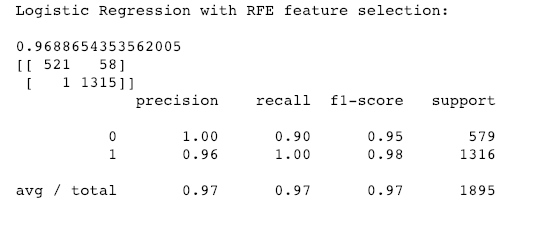
Comparing the results of all these models we see Random Forest and Logistic Regression created two of best models to predict if the deceased was African American or not. Given a new person with 98% accuracy the Random Forest model can predict if the deceased was African American.

**2- Predicting if the deceased was armed or not:**

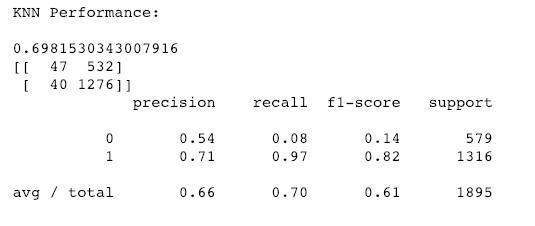
Here is another case of imbalance classes, it’s necessary to check confusion matrix and classification reports in addition to accuracy to check Precision and Recalls to find how many of our positive and negative predictions are TRUE.



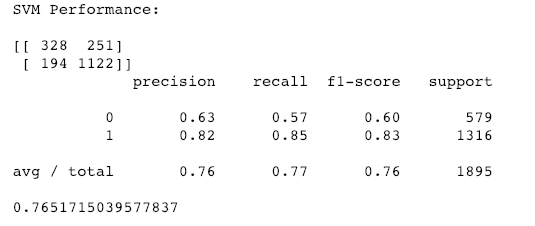
It seems the number of True Positives are low. One approach to improve the number of True Positives is to do feature selection. RFE is an automatic method to pick the features with the highest influence and filter out the noises. Through RFE, we picked the top 30 most influential features and re-ran the model to see if we had any improvements.



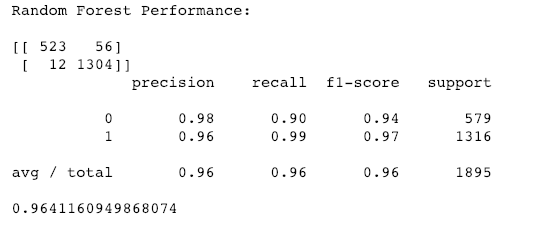
We see a clear improvement in our model after reducing features from 42 to 30.



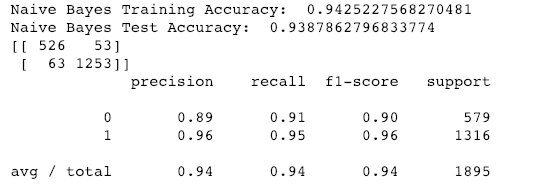
As expected KNN is not a good model to predict correct classes for imbalance datasets. Even though the accuracy rate is 0.70 but we can’t rely on the model as we see very low rate of True Positives.



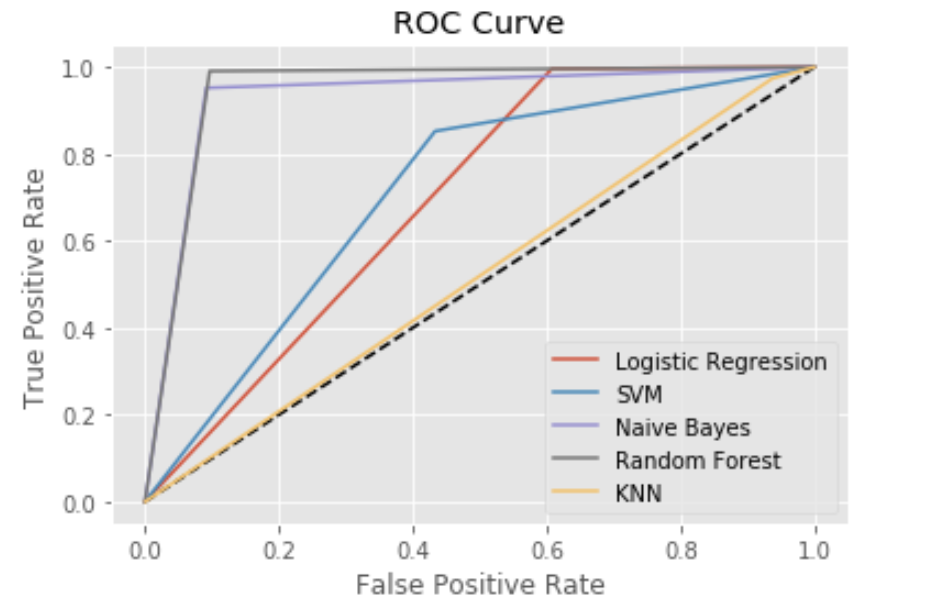
After tuning hyperparameters for SVM( C = 1000 and Gamma 0.001) we got 0.76 accuracy, but we have high number of false positives. Which is not very desirable or can be improved by using different algorithms.



Random Forest is one of the best algorithms for classification, not only we have very high accuracy, but number of False predictions (False Positive and False Negative) are very low.



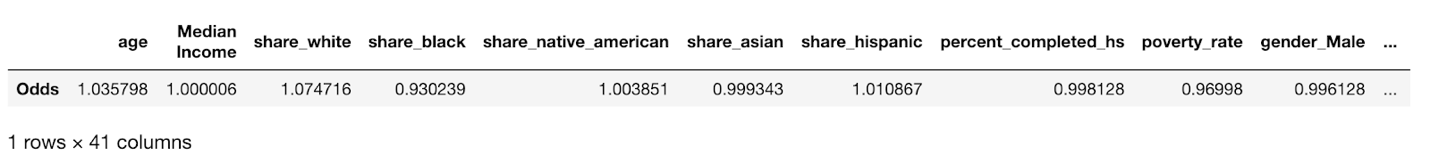
Naive Bayes is another great algorithm for classification with high accuracy and and low number of False predictions.



Comparing the results of all these models we see Random Forest and Naive Bayes creates two of best models to predict if the deceased was armed or not. Given a new person with 96% accuracy the Random Forest model can predict if the deceased was armed or not.

**3- Is there any relation between the probability of being armed and the age of the deceased?**

In this case we have multiple features and no interaction terms.

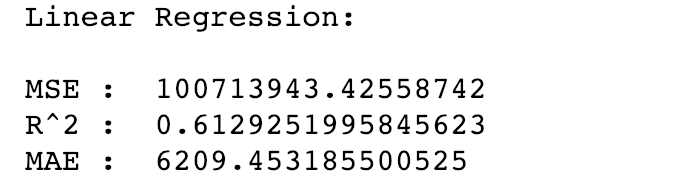


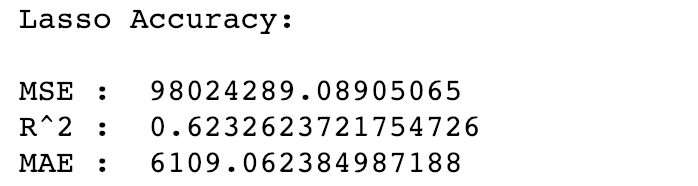
This fitted model says that, holding all the features except “age” at a fixed value the odds of finding the deceased armed is exp(1.035) = 2.82. In terms of percentage change, we can say that the odds of being armed increases 3% as the age increases.

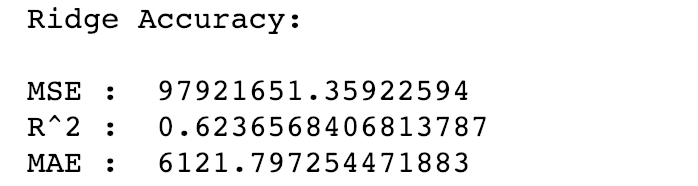
**4- Predicting the Median Income for the cities that shootings occurred:**

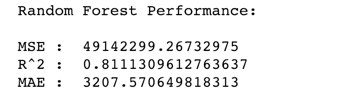
Large coefficients can lead to overfitting. In order to penalize large coefficients, we need to do some sort of regularizations: Lasso or Ridge Regression.

Here we are going to compare R-Square, MSE and MAE for Lasso, Ridge and Linear Regression to see which one has a better performance.



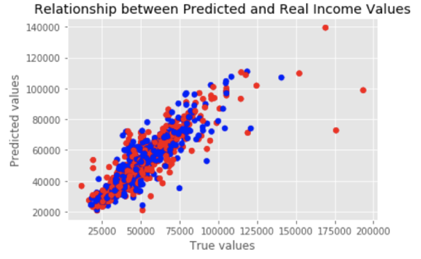






We see that Random Forest is bar far superior to other three Linear Regression models.

* Given a new incident, model can predict the approximate Median Income of the city that incident occurred with approximate accuracy of 81%.

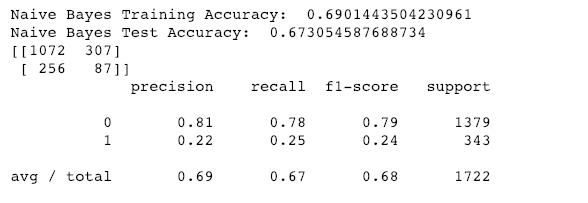


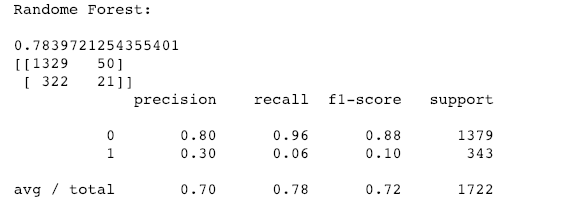
In general, there is a strong linear relationship between fitted predicted values and the original price. However, it's not perfect and we see some outliers in the graph.

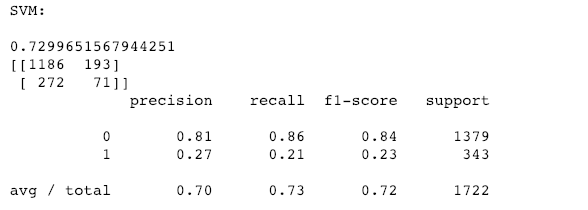
**5- Predicting if the deceased had shown any signs of mental illness or not:**

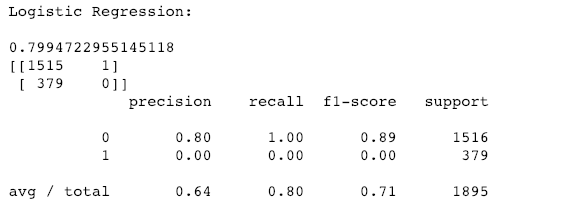
Problem is a classification problem, so we need to use classification algorithms such as KNN, Logistic Regression, SVM, Random Forest and Naive Bayes and compare their results to see which one provides a better result.

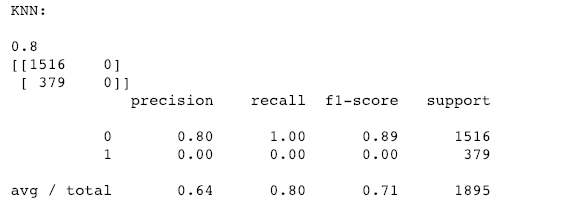
Since we have an imbalance classes, it’s necessary to check confusion matrix and classification reports in addition to accuracy to check Precision and Recalls finding how many of our positive and negative predictions are TRUE.





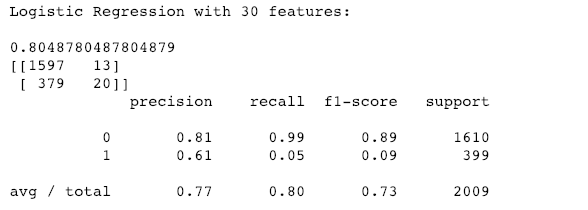


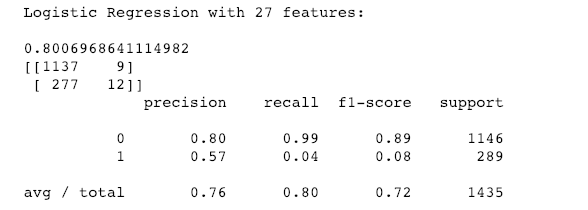


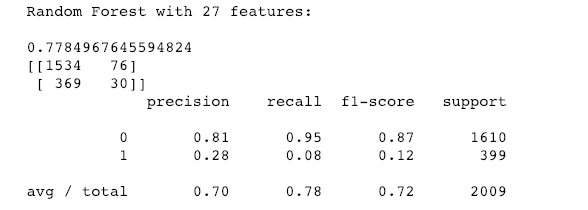


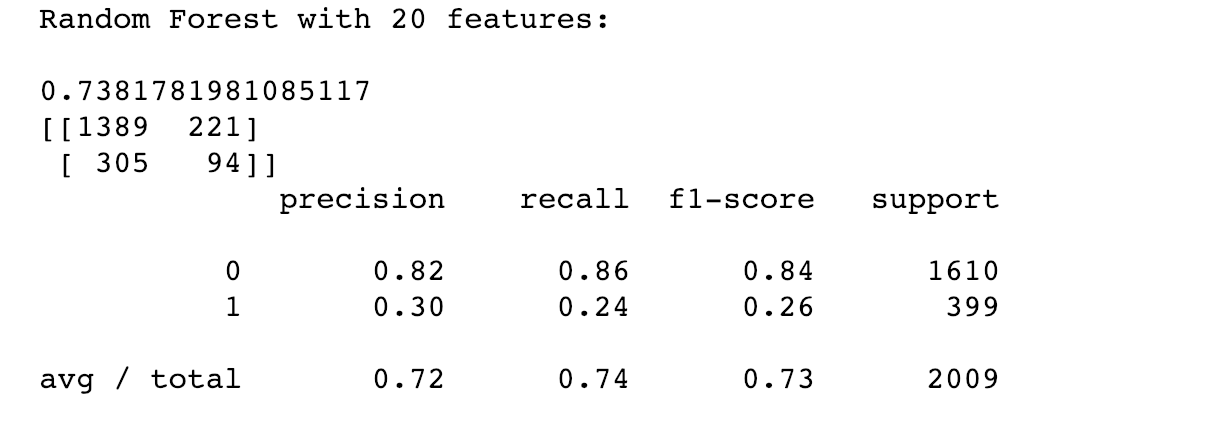
In all the above predictions we see recall is significantly low, that shows models have a tendency of classifying all the points as Positive and that create great deal of False Negatives. High number of False Negatives and False Positives can be a common case for highly imbalanced classes.

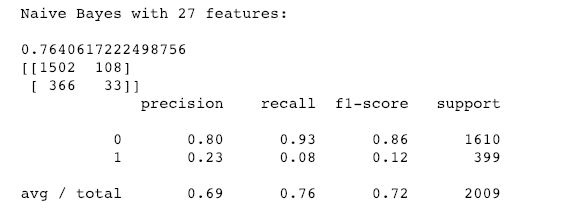
One common approach can be feature selection. Let's see how much we can improve the models by selecting more influential features and filtering out noises.

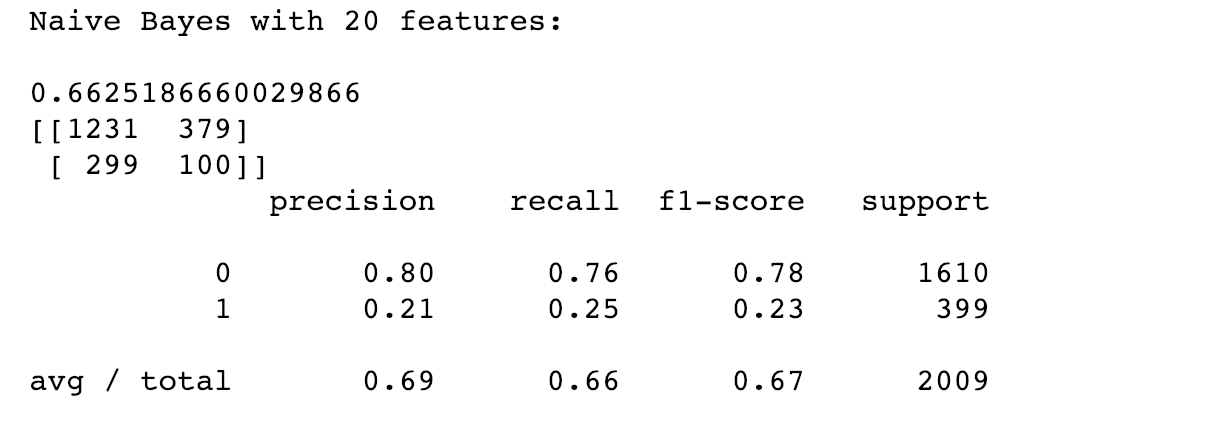


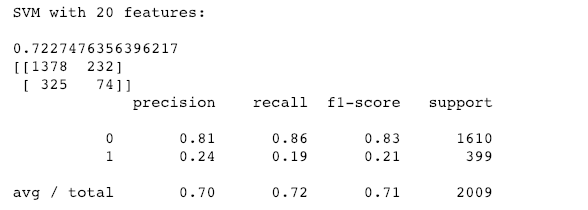








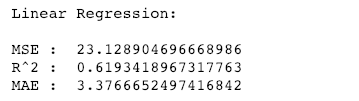


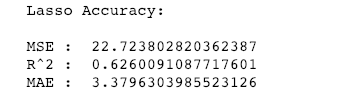


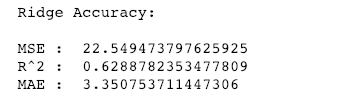
There was slight improvement in performance of Random Forest and Naive Bayes and SVM after feature selection, but still predictions are not good. The low performance of the model can be explained due to ambiguity of the phrase “sign of mental illness”. Mental illness can be investigated based on medical history, genetics, social and economic background, the environment that the person grew up, family, … Collected features about the deceased and the location of the shootings has very little to do with the causes of Mental illness or even detecting them.

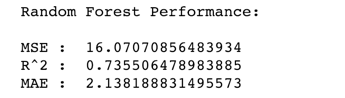
**6- Predicting if the poverty rate of the cities that shootings occured:**

Here we are going to compare R-Square, MSE and MAE for Lasso, Ridge and Linear Regression to see which one has a better performance.



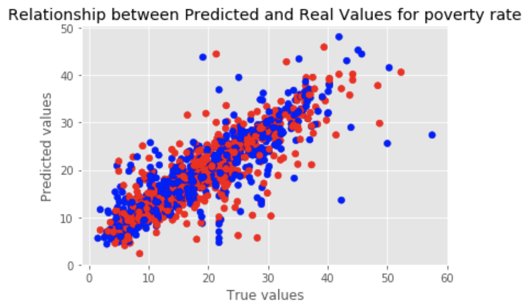






We see that Random Forest is bar far superior to other three Linear Regression models.

* Given a new incident, model can predict the approximate poverty rate of the city of the incident with approximate accuracy of 74%.



What was missing?

After wrestling with my dataset and going into the depth of problems, there were questions that tickled my brain and I was curious to know the answers but unfortunately my data set did not cover those areas. I would like to know the circumstances around the shootings better, for example, what percentage of these shootings were police self defense and what percentage weren’t real threats to police officers. It could be addressed by determining if the deceased was shot from behind or front, how many bullets were fired and what body parts did police officers aimed at.  There was a column in the dataset named as ‘armed’ which basically identifies if the deceased carried any sort of weapon or not. Well that’s not enough by itself to answer this question thoroughly. For example, if a person carry a gun in his/her trunk, and was fatally shot within or outside his car, with no access to his locked trunk it’s not right to assume the deceased was armed.

Another useful missing data could be the zip codes that incidents were taken place in. We could find a more accurate demographic info based on the zip codes.

I am also curious to see what percentage of fatal shootings occurred during the day time and what percentage during the night time and in the dark. This question could be answered if the exact time of the shootings were recorded.

Summary:

The goal of this project was to possibly find a recommendation to reduce the number of fatal shootings. We built and trained a model to predict the likelihood of deceased being armed. According to the model’s result African Americans has the least likelihood of being armed in their encounters with police officers. We can use this finding and make a recommendation that in future police officers use their teasers instead of weapons when they encounter African American civilians or suspects.