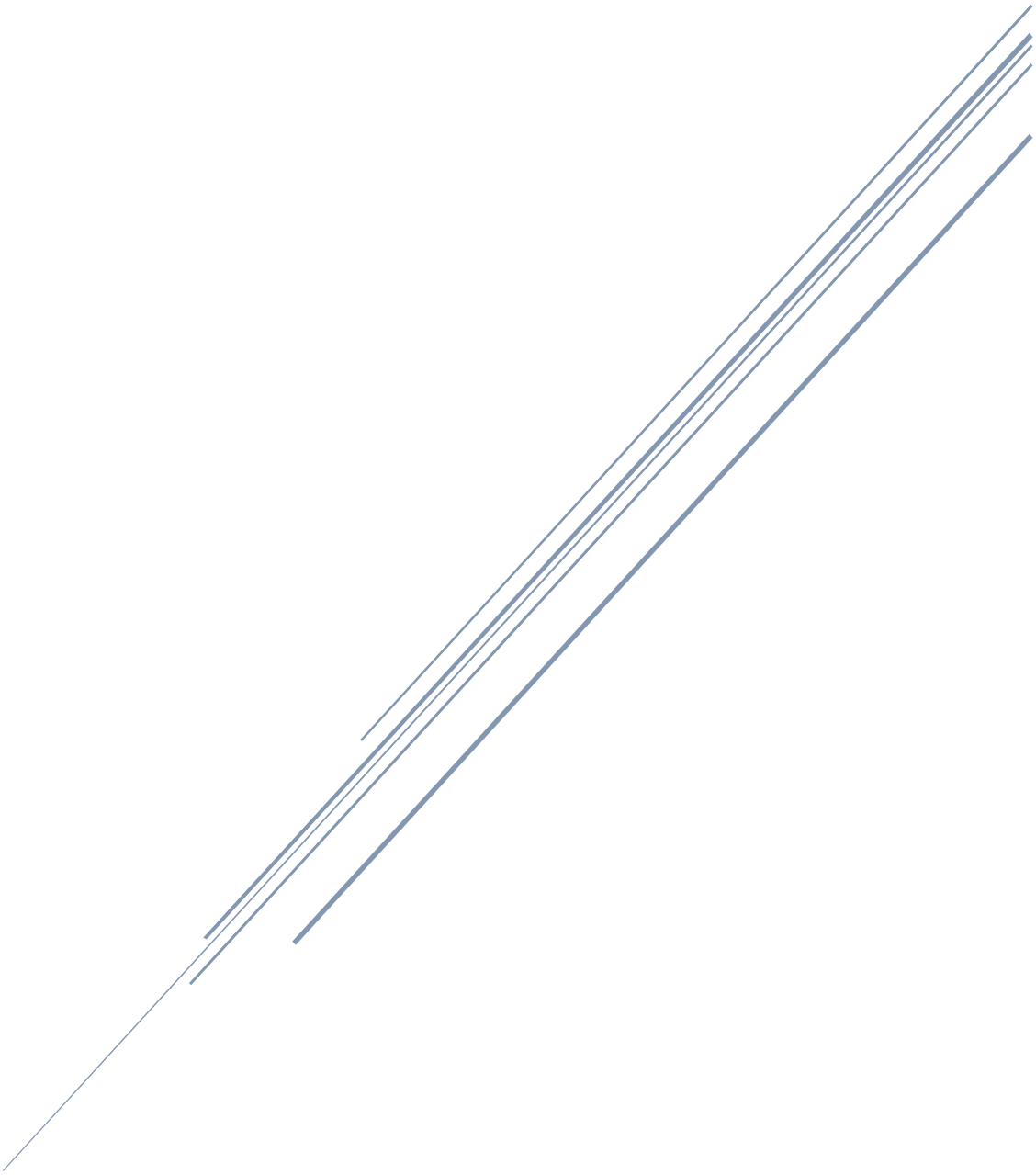
MACHINE LEARNING

Homicide Reports 1980 - 2014



David Coughlan
R00009964

Predicating the use of a firearm as a murder weapon on Homicide Reports 1980-2014

Abstract:

In this project, we use the homicide reports dataset containing data between the years of 1980 and 2014 to predict the likelihood of a firearm being used as a murder weapon. This paper considers the different ways to approach the data to explore accuracies using decision trees, svm, nearest neighbor, random forest, naive bayes and log r. The focus is then on label encoding and feature selection.

Table of Contents

Chapter 1: Introduction

1.1 Motivation

With shootings becoming a stable part of the average person's media diet in the past few decades. It has become more important to try to make predictions on existing data to help lessen the chances of them occurring or remaining unsolved. There is a regular occurrence where a mass shooting happens in America and shortly after the government discredits the right to buy firearms with minimal checks. The reduction of being able to openly buy firearms has had a huge effect over the past number of decades of homicide especially in America.

1.2 Problem

The aim of the paper is to develop a solution to a multi classification problem. Initially starting with a raw dataset of information and applying learning algorithms to help build the model. This project will be trying to accurately predict if a firearm was used as the murder weapon in a homicide

1.3 Approach to the problem

Initially the approach taken in the project is to perform pre-processing techniques. Some of those techniques are removing features from the dataset, imputing values that are missing and encoding labels.

After this pre-processing has been completed a series of algorithms are then run on the data each classifier using the cross-fold validation. From these results the best classifier can be chosen for the dataset and it can then be fine-tuned to attempt to raise its accuracy.

Chapter 2: Related Research

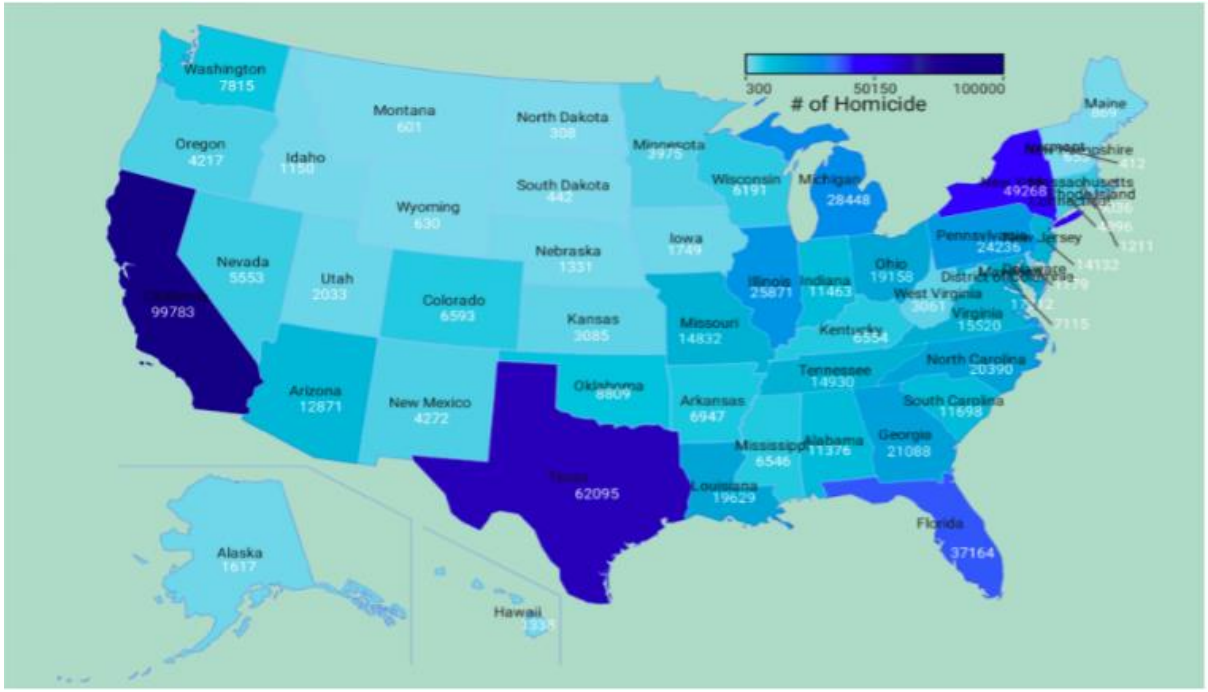


Figure 2.1: Homicides in each state. [1]

The dataset [2] originated from the Murder Accountability Project, a non-profit organization that is there to highlight the importance of investigations into homicides in the United States [1]. This organization seeks information from the federal, state and local authorities and is mainly run by ex-law enforcers, investigators and others who have a profession in criminology. The dataset contains murders from the FBI's homicide report and the Freedom of Information act data on over 22 thousand homicides that were not reported to the Justice Department. It consists of approx. 620 thousand valid homicides data where 440 thousand are solved and 180 thousand remain unsolved [1].

In the report I used for the basis of this project as a guideline but while also keeping my skewed task in mind. This dataset doesn't only deal with firearm murders but murders as a result of other causes or inflictions from other weapons such as knives, drowning, fire, drugs and more.

From another study done on the data [3] a general background can be given from the data there are a few key points:

- The rate of gun violence has in general gone down from 1980 to 2014. There was a peak in the early 90's in gun violence affecting states such as California, Texas and New York the most severely.
- The dispersion of gun related deaths has also gone down over time. This could be due to the increase in population to the areas allowing for less noise in the results as the true trends appear the higher the numbers are. This can be conveyed in the above figure in the contrast of California and Arizona where it has six times less in terms of population.
- From the data the least amount of deaths by firearm occur in February where they peak in July. Having taken the year 2000 as a sample. We can also see a second trend where it rises again in December and there is a dip just before this in November.

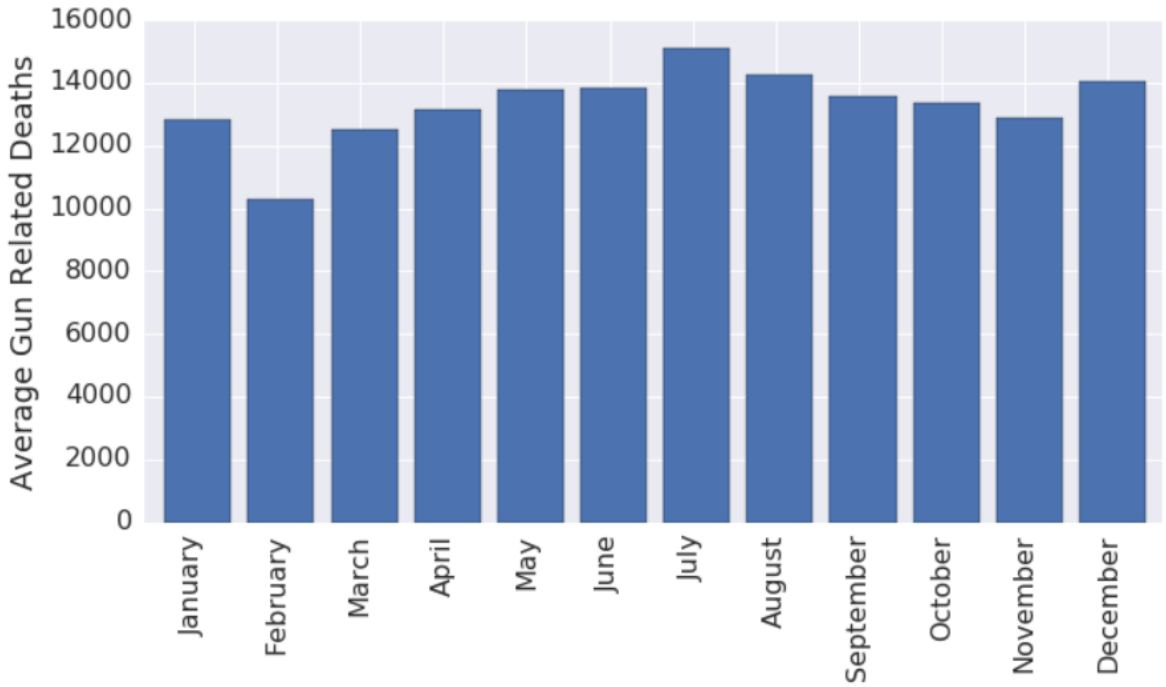


Figure 2.2 Firearm related murders in 2000 by month. [3]

The quite frightening correlation of solved murders for the same time line.

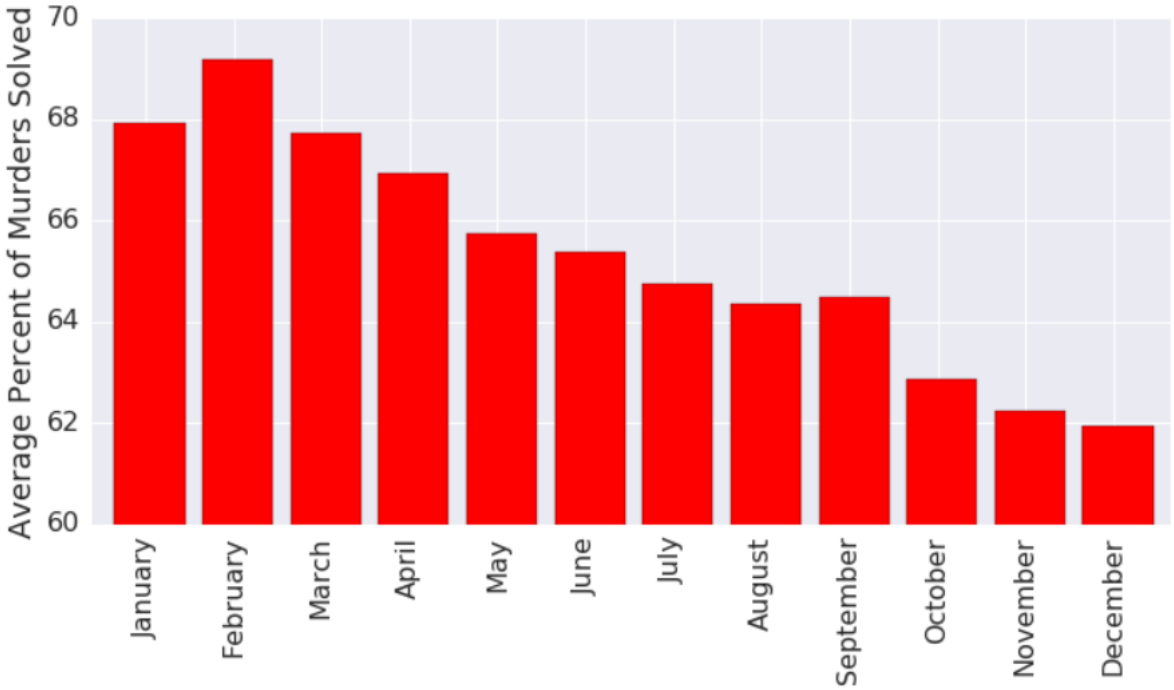


Figure 2.3 Solved murders by month in year 2000. [3]

This illustrates that the decline in solved murders as the year progresses. Also, another valid observation is that the month with the lowest murders by firearm also have the highest solved correlation.

In the research on this report they considered the dataset to be high imbalanced and therefore chose to tackle the problem with a soft margin SVM and over sampling method. Each task was considered individually choosing what features are correlated highly for the task.

Before training of the model occurred, the data had to be cleaned. The dataset obtained from Kaggle contained the following columns: Record ID, Agency Code, Agency Name, Agency Type, City, State, Year, Month, Incident, Crime Type, Crime Solved, Victim Sex, Victim Age, Victim Race, Victim Ethnicity, Perpetrator Sex, Perpetrator Age, Perpetrator Ethnicity, Relationship, Weapon, Victim Count, Perpetrator Count and Record Source.

Both from my own analysis and from the research report it is evident that features need to be dropped. [1] There are many that would differ in the separate goals however there are many common views on the features in question. Crime Solved feature has a lot of missing features so its deemed unsuitable. This would not have any correlation to our aim. There is a lot of untrustworthy data in regards of the cases thought to be solved where the perpetrator's age and victim's age could not be zero. Dropping the rows where this occurs. This helps to remove the nonsense incidents that's details prove unreliable. The next step is to encode the labels to have a numeric value. This is then applied to all races, genders, crime types and the month. I have also processed the feature weapon. I grouped all the firearm classifiers in the data. Any value with rifle, firearm, shotgun, handgun or gun are replaced with the classification of gun and anything that doesn't fall into these as other. These values are then label encoded giving any firearm murder weapon as a one and any others as a zero.

The crime type is of great importance to us. Since we are only trying to predict the use of a firearm in a murder we can use the crime type feature as a way of filtering the data. The crime type feature contains different degrees of crime type. Since we are only interested in murder or manslaughter we can ignore the manslaughter by negligence. This is done by grouping the dataset on that crime type immediately after loading.

Chapter 3: Algorithm/ Model Detail

On the data when plotted matching the gender with the murder weapon we get the following data.

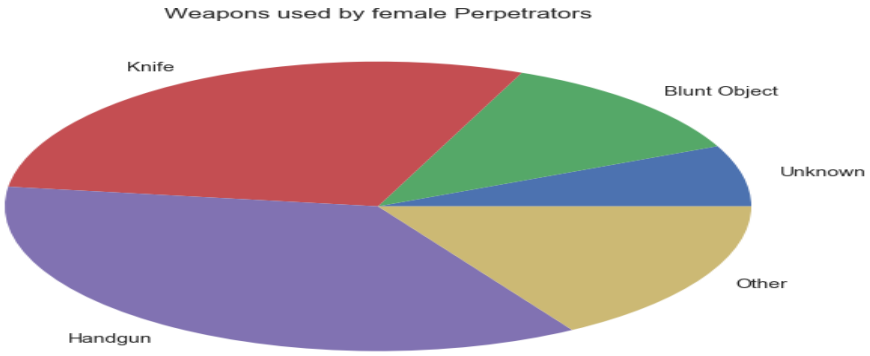


Fig3.1

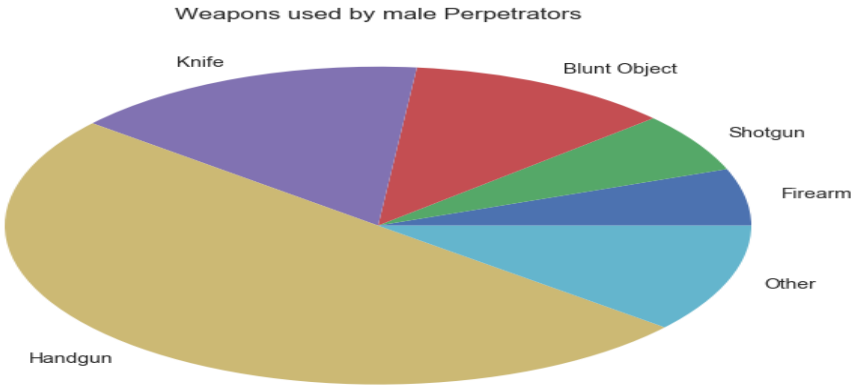


Fig 3.2

The first thing to be done with the data is to group it by crime type as we are only interested in the crime type of Murder or Manslaughter.

Considering we are looking for a firearm in general. I gathered all the subtypes of firearm in the data and replaced the weapon type as gun and anything that wasn't a firearm as other.

The initial algorithms that were used on the dataset were:

- Decision Tree
- K-Nearest Neighbour
- SVM
- Random Forest
- Naïve Bayes
- Logistic Regression

From initial results I decided to use the Random Forrest. [4] This algorithm is the most popular classification algorithm. The algorithm can be used for both classification and regression problems. The algorithm creates a forest of trees. The more trees present the higher the accuracy. It is an ensemble learning method for both regression and classification that operate by constructing multiple decision trees at the time of training and outputting the class that is the mode of the classification or the regression of each individual tree. Random Forest corrects the decision trees characteristic of over fitting the training set.

Decision trees are an often-chosen model when it comes to machine learning. Offering an almost off the shelf approach to the task. The algorithm is robust in its nature as it produces quantifiable results even with irrelevant features being present. It is never changing under scaling and other transformations of feature values.

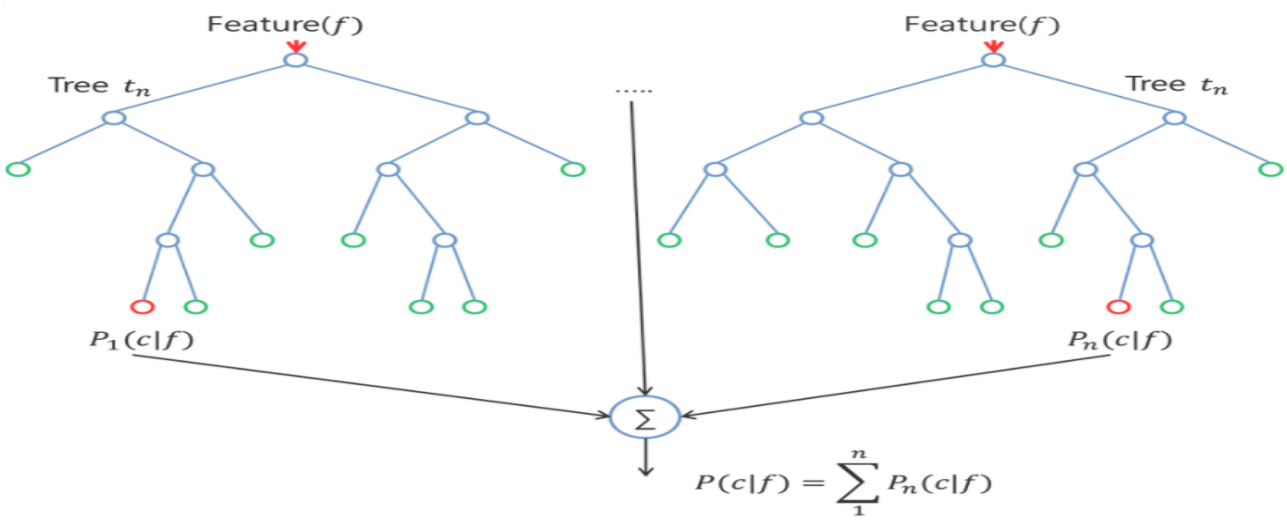
Random Forests use numerous classification trees. To classify a new object from an input vector the input vector is put down each tree in the forest. Each tree then gives a classification, often referred to as a vote, for that class. The forest then picks the classification that has the most votes from all the trees in its forest. [6]

The forest error rate depends on two factors. The correlation between two trees in the forest. An increase in correlation increases the error rate of the forest. The strength of each tree in the forest. Any trees with a low error rate is considered a strong classifier.

Random Forest has the following features:

1. Runs efficiently on large data sets.
2. Can handle thousands of input values without much value deletion.
3. It generates an internal unbiased estimate of the generalization error as the cycle of forest building increments.
4. It has an efficient method for estimating missing data and maintains accuracy when a large portion of the data is missing.

Figure3.3 Random Forest illustration



3.1 Random Forest Algorithm

The random forest algorithm [ ] (for both classification and regression) is as follows:

1. Draw nTree bootstrap samples from the original data.
2. For each of the bootstrap samples grow an unpruned classification/regression tree with a modification. At each node randomly sample mTry of the predictors and chose the best split from among these variables. This is an alteration from the usual occurrence of choosing the best split from the predictors.
3. Predict new data by aggregating the predictions of the nTree trees. (Majority votes for classification and average for the regression).
   An estimate of the error rate on the training data can be gotten by:
4. For each bootstrap iteration predict the OOB (Out of Bag) data (data not in the bootstrap sample) using the tree developed from the bootstrap sample.
5. Aggregate the OOB predictions. (Usually each data point would be OOB approx. 36% of the time so we aggregate these predictions). Calculate the error rate and call the OOB estimator of error rate.

Chapter 4: Empirical Evaluation

4.1 Initial Results

One of the tasks is to run tests on the data to see what the most accurate classifier is on the data. To do this we reduce the features of the dataset. Removing any features that were deemed not useful in predicting the use of a firearm in a murder. The dataset was so larger that a lot of the unknown data could be excluded instead of predicting a suitable value based on other entries in the feature.

I excluded SVM due to the lack of computation power available for the size of the dataset to make it a viable solution to work with given the shortage of time available. Running for three hours wasn't producing a result to I was left with no choice but to exclude it form my research.

The initial results were as follows:

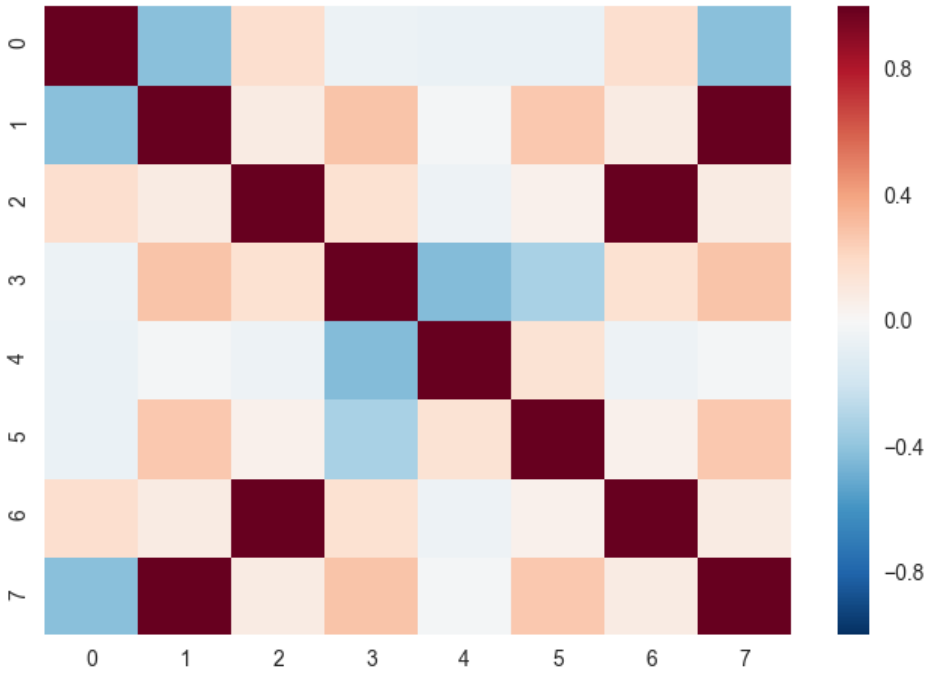
4.2 Classifiers run initially

Classifier Algorithm	Accuracy
Decision Tree	63.799%
K Neighbors	67.6204%
Random Forrest	69.628%
Naïve Bayes	67.709%
Logistic Regression	67.209%

From these results we can clearly see that the Random Forrest is the more promising approach. These results involved spitting the data set into two pairs of both training and testing sets selected at random to give the fairest level of accuracy.

For these initial results the accuracy is quite high. This is due to certain features being added back in that I had initially removed from my model. These features are all related to agency information. I noticed a large increase in the percentages from when they were omitted.

I added features back into my model based on my confusion matrix.



From this stand point I decided to try and further my accuracy and found the Random Forest Regression.

A Random Forest regression (Regression Forests) are an ensemble of different regression trees and are used for non-linear multiple regression. Each leaf contains a distribution for the continuous output variables.

Chapter 5: Conclusion

The results gathered in this project were very interesting. The first conclusion was where I had omitted features as well as other pre-processing techniques. Assumptions had been made about the correlation of some of the features and upon further development I was able to add back in some key features which made a huge impact on my results. This bundled with splitting the dataset into multiple sub sets for training and testing ensured the changes made. The use of hyper parameter optimization was them researched even though time restrictions stopped me from developing it as far as I would have liked.

A lesson to be taken away from this is to double check results for both with and without the changes made to the model. This will ensure that each step taken is a strategic move when developing a machine learning model.

5.2 Future Work

If time was more plentiful I would have liked to explore techniques for tweaking the model and algorithms more with knowledge acquired from more research papers being able to spend more time weighing up the best techniques for the problem at hand. For future works I would try an array of ensembles to raise the accuracy even further as well as hyper parameter optimization techniques. The use of weightings on certain features paired with an evenly more precisely cleaned dataset would be a good environment for further successes.

References

1. https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a010.pdf
2. https://www.kaggle.com/jyzaguirre/us-homicide-reports
3. http://ritvikmath.com/Gun_Violence_in_USA/
4. https://en.wikipedia.org/wiki/Random_forest
5. https://www.quora.com/How-does-random-forest-work-for-regression-1
6. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm