Murderer's characteristics prediction

Abstract

The aim of this project was to research and implement a solution that would help with homicides in the United States and hopefully assist with catching criminals. Homicide Reports 1980-2014 database was used and machine learning algorithms to help and predict perpetrator's features. Different methods were used to help come up with the most accurate model and numerous pre-processing techniques were applied to further aid these algorithms.

1. Introduction

According to FBI statistics, the violent crimes in the US rose by 4.1% in 2016 when compared to 2015. Murder and non-negligent manslaughter offences in 2016 have risen by 8.6% from 2015[1]. This is a significant increase in just over a year. The goal of this project is to try and predict perpetrators basic characteristics such as age, race and gender. With the help of machine learning process, such predictions can help model basic features of the perpetrator that the given agency is trying to catch. Same techniques and process can then be applied to a more extensive dataset which could contain a lot more information about the victims, perpetrators, crime scene, evidence etc.

The dataset consists of over 630K rows of homicide data, 440K which of are solved and rest are unsolved. Most of this project's time was spent on getting familiar with sci-kit and utilizing its tools for pre-processing the data in order to get "clean" data. Missing values, unnecessary or unrelated features to the test class and were dropped

2. Method

The dataset that has been used for this project was Homicide Reports 1980-2014. The reports contain 26 following features : 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 Race, Perpetrator Ethnicity, Relationship, Weapon, Victim Count, Perpetrator Count and Record Source.

This dataset contains very basic details of the homicides that occurred in the US and through the process of machine learning, basic perpetrators features can be predicted. The idea is to assess the accuracy of such basic prediction and eventually apply similar techniques to larger and more advanced datasets.

2.1 Model selection

Initial prediction of perpetrator's race was assessed using different algorithms such as:

- K-Nearest Neighbours
- SVM
- Naïve Bayes

- Decision Tree
- Logistic Regression
- Random Forest

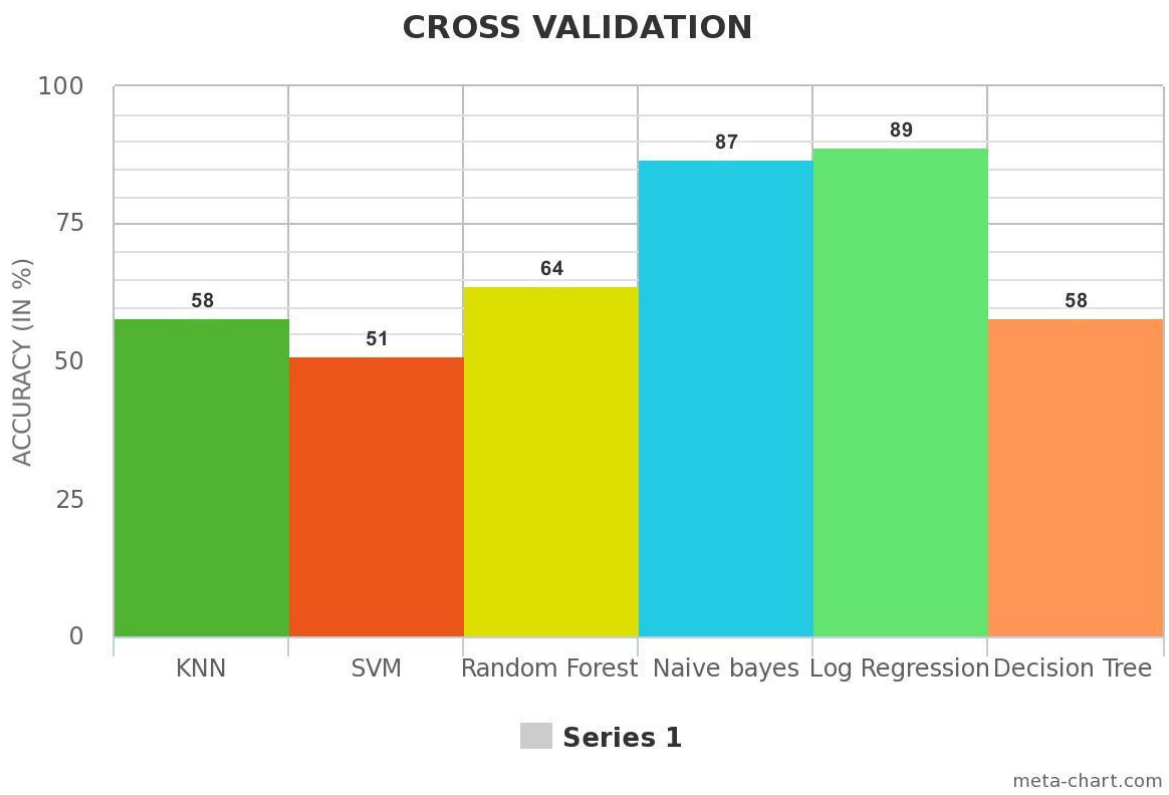

Figure 1. Captures the initial accuracy using different algorithms

Initial thoughts: the high accuracy looks to be high due to oversampling. There are a lot more white perpetrators than black and this could have skewed the accuracy highly in favour of Naïve Bayes and Logistic Regression algorithm.

3. Data pre-processing

In order to achieve the highest prediction possible, numerous pre-processing techniques had to be applied such as feature selection, cleaning of no value data, feature correlation, deleting rows with absurd non-sense values and so on.

3.1 Feature selection

Arguably, the more features I dropped, the less data I had to work with, therefore a worse chance of correct prediction. This would purely depend on the importance of the feature though. Some features could be dropped straight away as they actually had a lot of missing values such as the victim and perpetrator ethnicity features. Agency Code and Agency Name

– these two features classify the same thing – the agency, so the agency name was dropped and Agency Cod was used instead.

Using univariate selection I was able to get the score of importance of selected features. An absurdly high value for victim race feature as expected as most perpetrators are related to the victim, thus the race is almost always the same as well. Next most important feature was perpetrator's sex, with the F-score of 574 and victim count with the score of 438. Interestingly, one of the features, Month, had a really low score of 0.46 and victim's age had a f-score of 1.27 which is extremely low given that the age is realistically an important factor.

All of the features that were not a numerical type, had to be converted to an integer. Label encoder from sci-kit transforms all values to numerical data for example Knife was mapped to 0, Blunt Object was mapped to 1, Strangulation was mapped to 2 and so on.

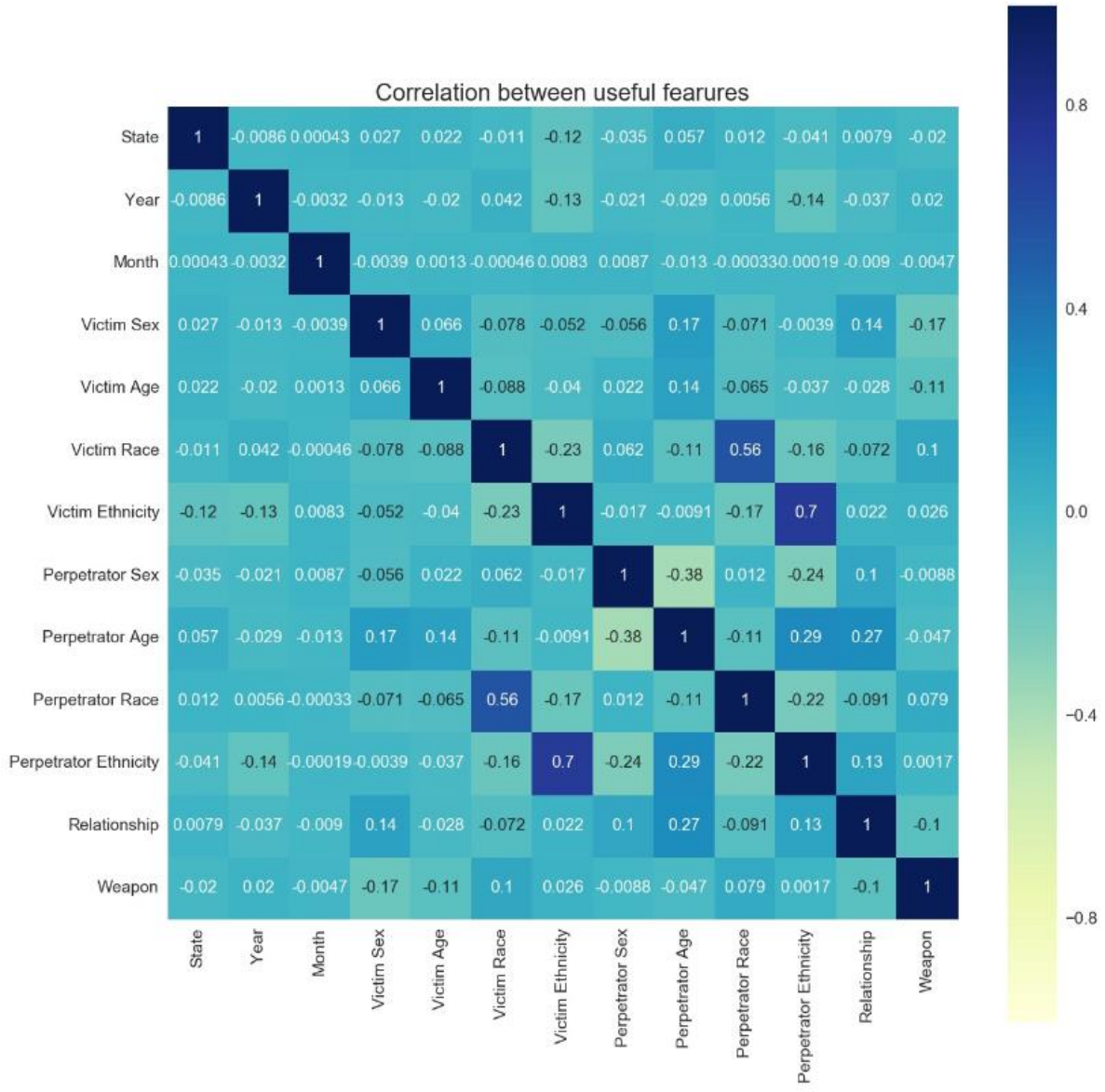

Figure 2. (Shen Liu) Captures the correlation between useful features

From figure 2, it is apparent that victim race has a strong correlation coefficient with perpetrator race, with the coefficient being 0.56. This is the strongest correlation as shown in the graph with the second strongest being Perpetrator Sex tied with State at a 0.012 correlation coefficient which is extremely low.

4. Related Research

Homicide clearance rates in police departments are decreasing. The clearance rate for homicide has dropped from 91% in 1965, to 64% in 2002 [3]. This can be attributed to more stranger to stranger homicides, resulting in crimes harder to solve than those with a close relationship. Other reasons could include inside politics and bureaucracy of the police departments, lack of resources and personnel turnover.

Unsolved homicide murders not only have a devastating impact on the relatives of the victim, but also the perpetrator is out on the streets, being able to murder another person again. This is a worrying trend in homicide investigations and hopefully with the help of machine learning, those numbers can rise back to what they were back in 1965. It's not going to be a short process but every tool developed to help the investigations is better than no tool.

The most ideal view of machine learning and homicide investigations would be to of course eventually pinpoint the exact characteristics of a murder with enough data given, but that is not something that would be possible at this day and age with the current technology. What looks like to be the future is the use of cameras. Image processing is always advancing forward and it is probably one of the best ways to go about catching criminals.

Liu et al predicted perpetrator's sex using three different models. Soft-Margin SVM, oversampling data and SVM and oversampling data and logistic regression. Seeing as the number of males to females has a ratio of 8:1, the algorithms favoured males most of the time, that's why the accuracy was skewed to 90% all the time.

Liu first reshuffled the data, using 200K to train, 100K to validate and another 200K to test. For the purpose of pre-processing, where the age had a value of less than 0 and more than a 100, the rows were dropped. After the pre-processing, about 350K of data was left for training, testing and validation.

Not only accuracy was considered, as True Positive Rate, True Negative Rate and Balanced Error Rate had to be evaluated as well for the truer prediction accuracy as the dataset was imbalanced in terms of perpetrator sex feature.

Liu chose a feature array with the following columns : [1, Victim_Ethnicity, Victim_Sex, Victim_Age, Victim_Race, State]. Perpetrator's sex could never be guessed if the perpetrator was female and a baseline module was used, but using true negative has helped balance the model prediction. Oversampling performed better than margin SVM method in terms of true negative results.

Similarly, perpetrator age and race were predicted using logistic regression, SVM and the baseline model. The accuracy results are extremely close and often not further off than 1%

between those three models used. The predictions for minorities such as Native and Asian/Pacific Islander in logistic regression had a slightly larger value that the accuracy for Asian Pacific/Islander suggesting that the minorities live in a smaller life circle compared to the majority of the population.

References

1. https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-2016-crime-statistics#
2. https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a010.pdf
3. https://ebookcentral.proquest.com/lib/cit-ebooks/reader.action?ppg=16&docID=3019734&tm=1511259264201