Big Data Continual Assessment

Analysis of Homicide Perpetrators

Aoife Sayers

Institute of Technology, Tralee, North Campus, Dromtacker, Tralee, Co. Kerry , Ireland

Abstract

A Kaggle homicide dataset dating from 1980 to 2014 was used to

*Keywords:*

*Abbreviations: SVM = Scalable Vector Machine, MLP = Multi-Layer Perceptron*

1. Introduction

Homicides instances have decreased greatly since the peak period of homicides in 1980. Analysis of homicide perpetrators may help the law in targeting resources and area for improvement. The analysis of perpetrators may also aid police intelligence model better crime occurrences and analyse perpetrator’s activity patterns. The Murder Accountability Project is …

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6061179

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6061179

Each homicide data includes

the following important features:

1. State: The state where the homicides happened

2. Year: The year when the homicides happened

3. Crime Type: It includes two type of crime , i.e. Murder

or Manslaughter and Manslaughter by Negligence

4. Crime Solved:It indicates whether the homicide has

been solved or not.

5. Victim Sex

6. Victim Age

7. Victim Ethnicity

8. Victim Race

9. Perpetrator Sex

10. Perpetrator Age

11. Perpetrator Ethnicity

12. Perpetrator Race

1. Methods

The KDD (Knowledge Discovery in Databases) Process was used to analyze the profiles of homicide perpetrators. The KDD Process consists of selection, preprocessing and transformation of the data. Data Mining techniques, Logistic Regression, Neural Networks and Scalable Vector Machines (SVMs) were used for the prediction of data.

* 1. Selection & Pre-processing

The dataset selected for the KDD process analysis was the Murder Accountability Project’s Homicide Reports 1980-2014. The Murder Accountability Project is an American organization which compiles information about homicides, particularly serial homicides in the USA (Project, 2018). The dataset was selected from Kaggle <https://www.kaggle.com/murderaccountability/homicide-reports> and is in the form in of CSV file. The dataset contains information about the ages, different races, sexes, ethnicities of victims and perpetrators and the relationship between the victims and perpetrators. The CSV file was read into a Panda’s dataframe. The dataset contained no missing values, therefore no objects had to be removed. Feature Reduction was used to remove Record Id, Agency Code, Agency Name and Agency Type. The listed columns were irrelevant in predicting the features of Perpetrators. The columns Perpetrator Age and Victim Count were cased to numeric fields to allow numeric processing e.g. the mean. A perpetrator with age 0 was removed from the dataset as the age skewed the dataset and appears to incorrect data.

* 1. Transformation

The CSV file has already been read in from a .CSV file to a pandas dataframe. The previous step of pre-processing is executed on the pandas dataframe which exists in persistent memory.

* 1. Data Mining – Descriptive & Visual Analysis

The dataset was read in using the pandas module function *pd.readcsv(filename).* To get the size of the column df.shape() was used to print out the number of rows by column. The dataset shape is 638454 rows x 24 columns. df.info was used to print out the 24 columns and the associated datatype. The function df.head(5) give the first 5 rows of the dataset.

* + 1. Solved vs Unsolved cases

In the dataset of 63,454 rows, 29.9% cases, 190282 cases remained unsolved and 70.2% were solved suggesting not all perpetrators had been apprehended and identified by the police force. As a result, this explains why contains Unknown values in some columns of the dataset.

* + 1. Homicide distribution in terms of weapon type

An analysis of the most common weapon types was plotted on a horizontal bar chart. The bar chart visualized thrs an outlier of weapons – the handgun. Also, from the histogram, it’s evident that gun violence makes up a large portion of homicides. There are 5 types of guns listed within the dataset: handguns, guns, firearms, shotguns and rifles.

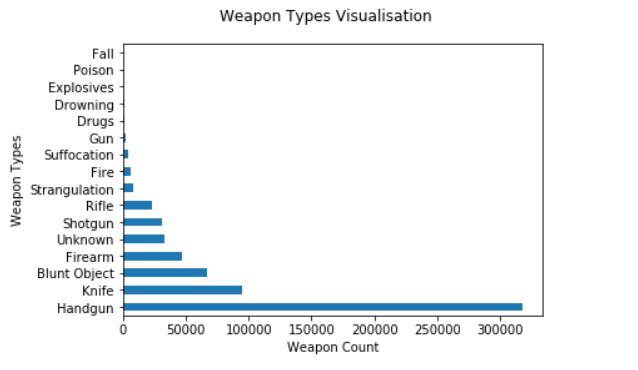


Fig. 2. Percentage of Crimes vs Unsolved Crimes.

* + 1. State with most cases of gun related homicide

The interpretation of gun homicides findings is where the following weapons have been used have been used by a perpetrator against a victim: Handgun, Rifle, Shotgun, Firearm. A total of 28,4390 cases of the full dataset of 63,454 cases of homicide account for gun related violence. The state in the USA with the most instances of homicide account for California. This suggests California is the state where people are most susceptible to die from homicide by gun violence.

* + 1. Weapons used by Perpetrator against an Unknown/Stranger Victim

In the dataset where the Perpetrator’s relationship to the Victim is Unknown/Stranger accounts to 40% of all homicides. The most common weapon used against an unknown/stranger victim is a handgun, reinforcing the point about gun violence. The three most common relationships between a perpetrator and their victim are strangers, acquaintances and unknown. This may suggest that the crime is random, and the Perpetrator was in pursuit of something e.g. money or it may have to do with a psychological disorder of the Perpetrator. The perpetrator may be a serial killer.

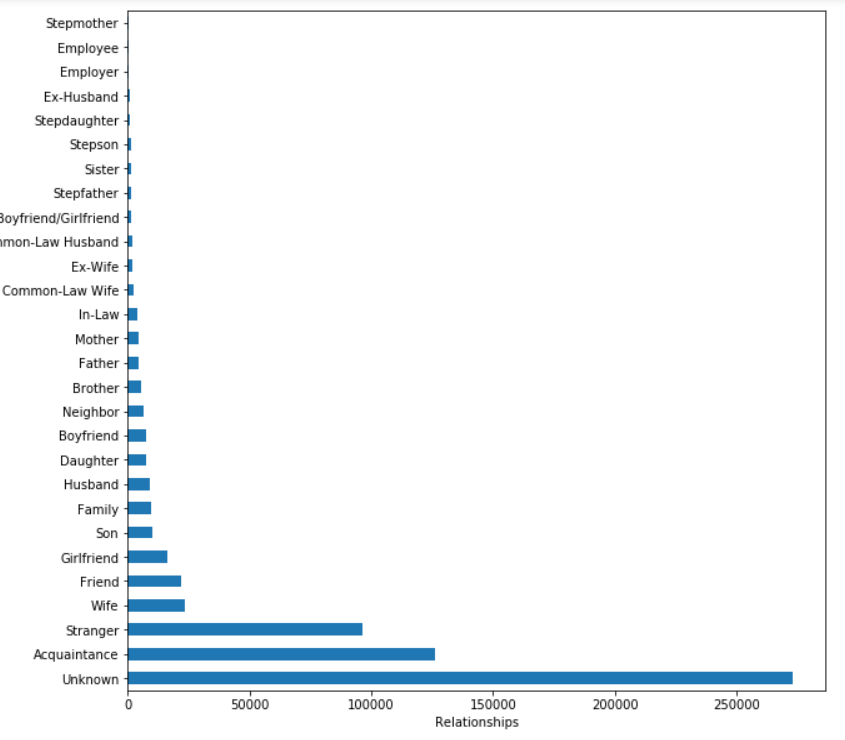
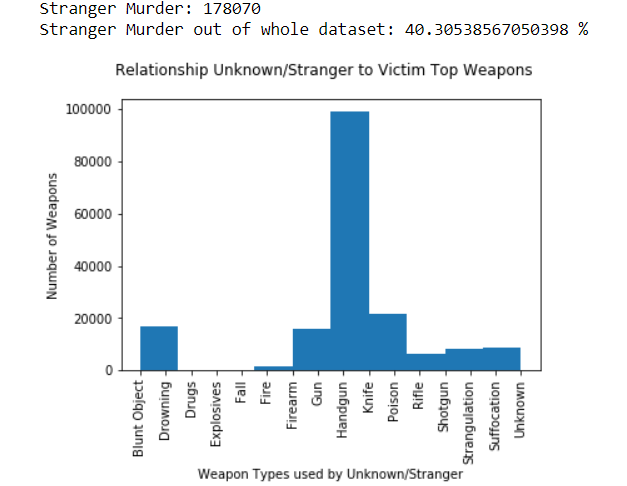


Fig. 4. (a) (b) Left – Bar Chart of Unknown/Stranger Relationship Weapons, Right – Horizontal bar chart of all the relationships

* + 1. Cases of Familicide

For analyzing cases of familicide only the immediate family was included. Immediate family consisted of the following relationships: Wife, Husband, Son, Daughter, Brother, Sister, Father, Mother and Family

From the histogram and mode() function the most common Familicide fatality is homicide against the wife by a significant amount. Familicide is least common against the father within this dataset. The results suggest evidence of domestic violence.

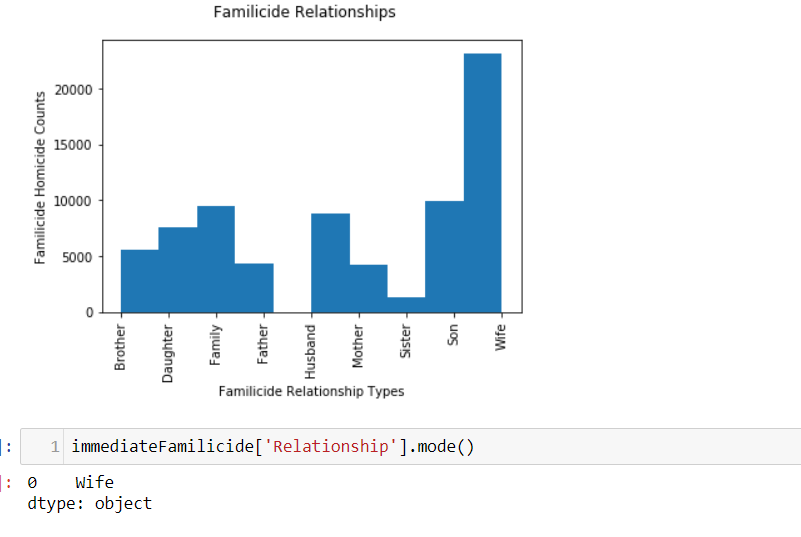


Fig 5. Familicide Relationships plotted on a histogram

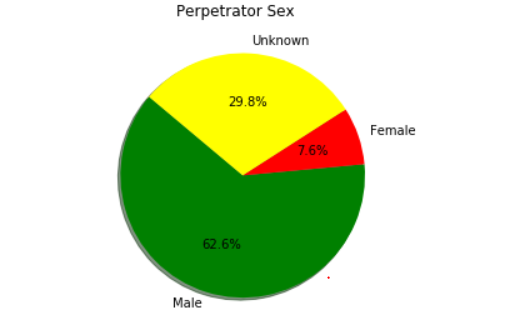
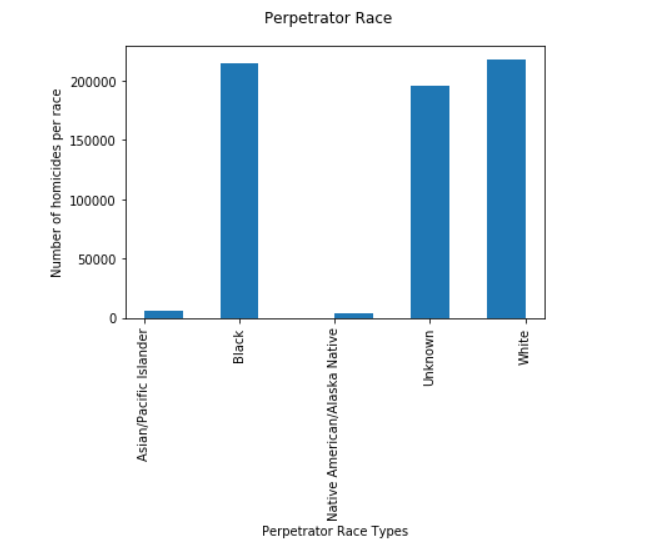
* + 1. Perpetrator profile – Sex, Race, Age

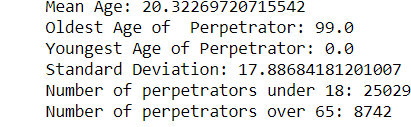
The pie chart in figure 6(a) indicates that that there are more Male perpetrators. The evidence indicates males are most likely to commit a homicide as opposed to their female counterparts. There is also 29.8% of unknown perpetrator sexes as they may not have been apprehended by the police yet.

The perpetrator race types histogram indicates that Perpetrators of a white race are most likely to commit homicide.

The perpetrator’s age profile is an average of 20 year old. The average of a perpetrator is young which highlights the need for additional resources for young offenders in the American Justice Sytem. However, the youngest age of a perpetrator is 0 and oldest age of a perpetrator is 99 which suggests some data may be inaccurate and as a result skews, the average age.

From the profiling of a perpetrator it is most likely that a White Male around 20 years old is most likely to be guilty of homicide.

 Fig 6. (a)Top Left - Perpetrator Sex on a pie chart (b) Top Right – Perpetrator Races on a histogram (c) Bottom Left – Age profiling of a perpetrator

* 1. Data Mining – Predictive Analysis
     1. Perpetrator Sex Prediction

Predictive Analysis was performed on the Perpetrator Sex using Logistic Regression and a SVM following the method in the journal Prediction on Homicide Reports, 1980-2014 (Liu, et al., 2018). The dataframe was preprocessed to only include rows where the crime was solved; the perpetrator was not unknown, and the victim sex was not unknown. The dimension of the dataframe was reduced further to include the following columns: X = df[['Year', 'Victim Ethnicity', 'Victim Sex', 'Victim Age', 'Victim Race']]. The get\_dummies() method in the pandas module was used to transform categorical data. The variable to be predicted is specified. A train- test split using sklearn is created: 70% train: 30% test.

A model for Logistic Regression is created. The X and Y train values are fit into the model and then predicted with the X test values.

A model for a SVM is created The X and Y train values are fit into the model and then predicted with the X test values.

* + 1. Perpetrator Race Prediction

Predictive Analysis was performed on the Perpetrator Race using Logistic Regression folowing the method in the journal Prediction on Homicide Reports, 1980-2014 (Liu, et al., 2018). A Multi-Layer Perceptron also used to predict a Perpetrator’s Race. The dataframe was preprocessed to only include rows where the crime was solved; the perpetrator was not unknown, and the victim sex was not unknown. The dataframe’s dimension was reduced further to include the following columns: X = df[[ 'Victim Ethnicity','Victim Race']]. The get\_dummies() method in the pandas module was used to transform categorical data. The variable to be predicted is specified. A train- test split using sklearn is created: 70% train: 30% test.

A model for Logistic Regression is created. The X and Y train values are fit into the model and then predicted with the X test values.

A model for a SVM is created. The X and Y train values are fit into the model and then predicted with the X test values.

A model for a MLP is instantiated. The X and Y train values are substituted into the fit method. Following, fitting the training data, the neural network predicts on the test set. A classification report is generated to print the accuracy.

1. Results/Data Findings

Logistic Regression aims to model P(label | data) by training a classifier of the form (Liu, et al., 2018)



Predicting a Perpetrator’s Sex using Logistic Regression Results

Accuracy score of 0.891844787651%

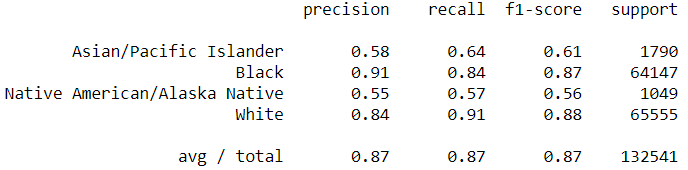
Predicting a Perpetrator’s Sex using SVM Results

This seems to run endlessly and as a result no results are available

Predicting a Perpetrator’s Race using Logistic Regression Results

Accuracy score of 0.865981092643%

Predicting a Perpetrator’s Race using a Multi-Layer Neural Network Results



1. Conclusion

# References

Liu, S., Wang, L. & Zhi, X., 2018. Prediction on Homicide Reports, 1980-2014. *UC San Diego,* I(1), p. 10.

Project, M. A., 2018. *Murder Accountability Project.* [Online]   
Available at: www.murderdata.org/  
[Accessed 1 May 2018].