Big Data Continual Assessment

Analysis of Homicide Perpetrators

Aoife Sayers

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

Abstract

A homicide dataset sourced from Kaggle containing homicides from 1980 to 2014 was used to perform descriptive and visualizations on the data. Predictive analysis was used to predict the features of the perpetrator. First the dataset was analyzed using descriptive and visualization techniques (Liu, et al., 2018)on the solved vs unsolved columns, the most popular weapons used, the relationships between the perpetrator and the victim, cases of stranger violence and familicide, cases of gun violence and analysis on the perpetrators. Predictive analysis was performed on the perpetrator sex using Logistic Regression and a SVM. Predictive analysis was also performed on the Perpetrator’s Race using Logistic Regression and a Multi-Layer Perceptron Neural Network.

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

1. Introduction

Homicide instances have decreased greatly since the peak period of homicides in 1980. Analysis of homicide perpetrators may help the law in targeting perpetrator and areas of improvement in the justice system. The analysis of perpetrators may also aid police to model crime occurrences more efficiently. Analysis of homicides also facilitate analysis of a perpetrator’s activity patterns. The homicide dataset was compiled by the Murder Accountability Project in the USA. This dataset includes the age, race, sex, ethnicity of victims and perpetrators, in addition to the relationship between the victim and perpetrator and weapon used (Kaggle, 2018). The dataset contains 638454 rows and 24 columns. The research should predict the most likely perpetrator race most likely to commit homicide and the perpetrator gender most likely to commit a homicide. Young Adults with access to firearms in heavily populated environments are more at susceptible to engaging in homicide.

1. Literature Review

While the number of homicides has reduced significantly since the 1980s, homicide is still a leading cause of death in the United States of America (Rothenberger, et al., 2011). The benefits of analyzing a perpetrator guilty of homicide help the justice system target resources for preventing homicides, monitoring possible offenders, and help gain intelligence in targeting homicide crime. A combination of visualizations, descriptive analysis, and predictive analysis on homicides would aid the justice system in discovering homicide related patterns. Predicting a perpetrator’s features could vastly aid the justice system in shortlisting a list of potentially perpetrators.

A significant number of homicides in US are caused by guns due to the nature of the lenient laws towards gun control in America. In relation to gun violence, President Trump sent additional Federals to Chicago to deal with the "epidemic proportions" of crimes and killings (usnews.com, 2017). Analysis of the homicide dataset concluded that Chicago was the state with the most instances of gun related homicides. This paper aims to identify features of the Perpetrator using predictive data mining techniques to predict the sex, race and ethnicity.

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 data frame in a Python notebook. The dataset contained no missing values, therefore no objects had to be removed. The irrelevant columns listed here were removed: Record Id, Agency Code, Agency Name and Agency Type. The removed columns were irrelevant in predicting the features of Perpetrators. The columns Perpetrator Age and Victim Count were casted to numeric fields to allow numeric processing e.g. to calculate 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 data frame. The previous step of pre-processing is executed on the Panda’s data frame 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 in figure 1 demonstrates the outlier of the 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.

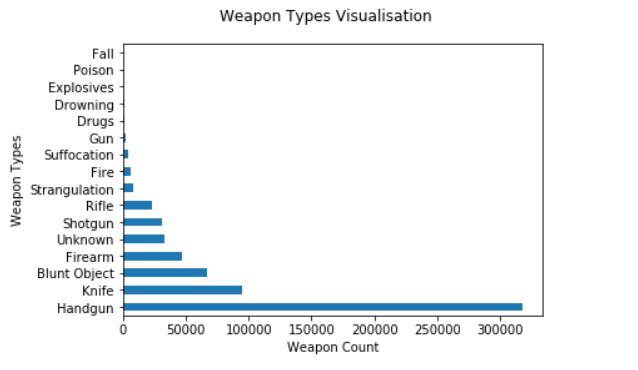


Fig. 1. Percentage of Crimes vs Unsolved Crimes.

* + 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 are in California.

* + 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 as seen in figure 4(a), reinforcing the point about gun violence. The three most common relationships between a perpetrator and their victim are strangers, acquaintances and unknown as seen in the bar chart in figure 4(b).

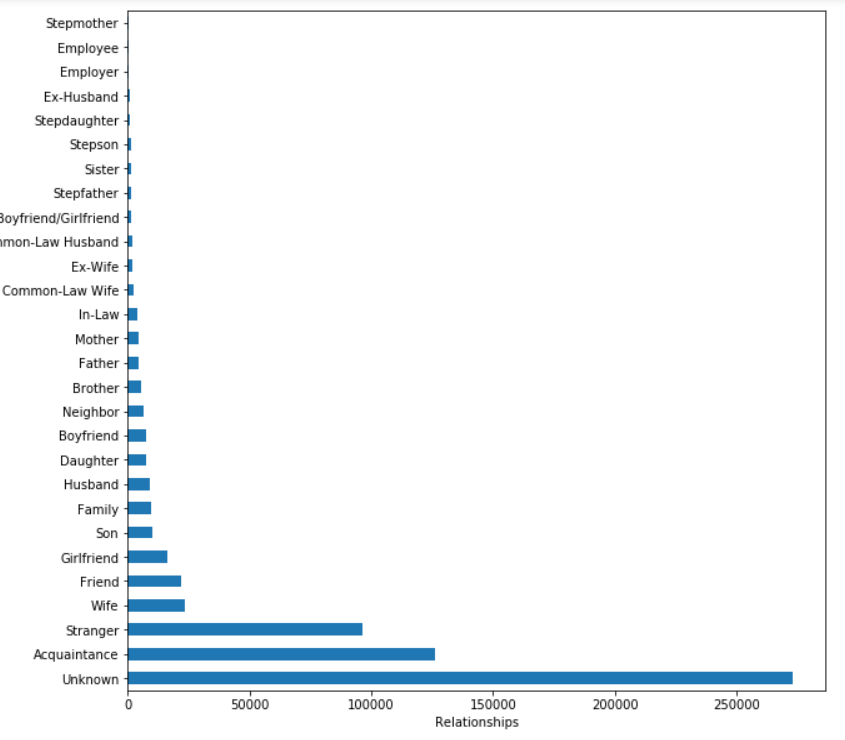
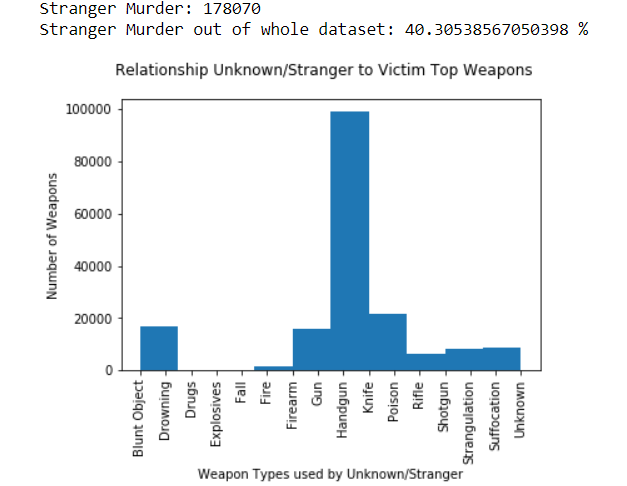


Fig. 2. (a) (b) Left – Bar Chart of Unknown/Stranger Relationship Weapons, Right – Horizontal bar chart of all the relationships

* + 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 in figure 3, and the 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.

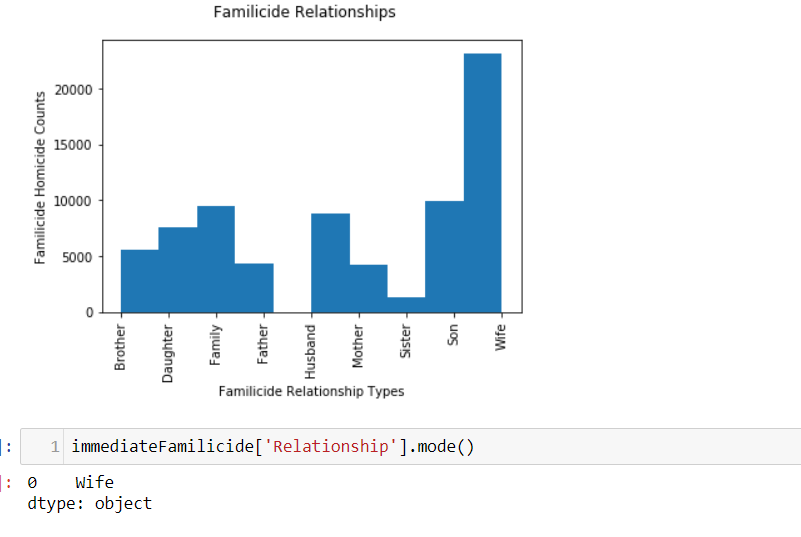


Fig 3. Familicide Relationships plotted on a histogram

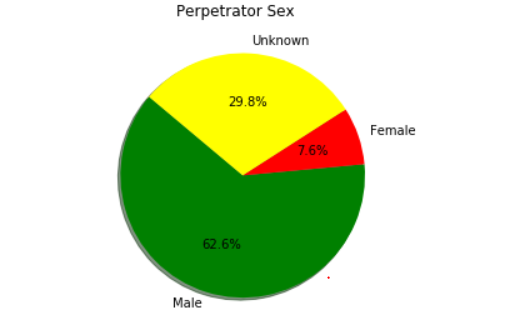
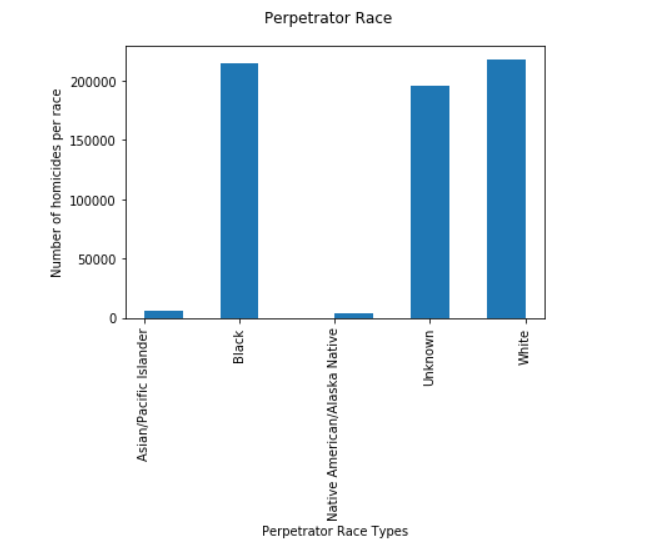
* + 1. Perpetrator profile – Sex, Race, Age

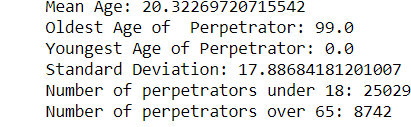
The pie chart in figure 4(a) indicates that that there are more Male perpetrators. Females only account to 7.6% of females. Unknown Perpetrators account for the 29.8% of all the Perpetrators sex. The evidence indicates males are most likely to commit a homicide as opposed to their female counterparts. There is also 29.8% of unknown perpetrators sexes as they may not have been apprehended by the police yet.

The perpetrator race type’s histogram indicates that Perpetrators of a white race are most likely to commit homicide as seen in figure 4(b).

The perpetrator’s age profile is an average of 20 year old as shown in figure 4(c). The average of a perpetrator is young which highlights the need for additional resources for young offenders in the American Justice System. 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 4. (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).

*Pre-processing for Logistic Regression and SVM*

The data frame was preprocessed to only include rows where the crime was solved; where the perpetrator was not unknown, and the victim sex was not unknown. The dimension of the data frame was reduced further to include the following columns: Year, Victim Ethnicity, Victim Sex, Victim Age, and Victim Race. The X variable should only contain the following:

X = df[['Year', 'Victim Ethnicity', 'Victim Sex', 'Victim Age', 'Victim Race']]

The get\_dummies() method in the pandas module was used to transform categorical into binary numbers. The variable to be predicted is specified. A train- test split using the Sklearn (Python module) is created: 70% train: 30% test. The train set contains approximately 44697 rows. The test set contains approximately 13409 rows of data.

*Data Mining – Predictive Analysis – Logistic Regression*

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 following 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 data frame was preprocessed to only include rows where the crime was solved; the perpetrator was not unknown, and the victim sex was not unknown. The data frame’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
   1. Predicting a Perpetrator’s Sex using Logistic Regression Results

The array used for predicting a Perpetrator’s Sex preprocesses the data frame so only solved crimes, known Perpetrator Races and Victim Sexes were included. The data frame dimensionality was reduced so only the following columns were used Year, Victim Ethnicity, Victim Sex, Victim Age, and the Victim Race.

After preprocessing the data frame and including the columns as specified in the journal Prediction on Homicide Reports, 1980-2014 by UC San Diego (Liu, et al., 2018), the accuracy result of predicting a perpetrator’s sex was the same. Logistic Regression on the perpetrator resulted in an accuracy score of 0.891844787651%.

* 1. Predicting a Perpetrator’s Sex using a SVM

The preprocessing for predicting a Perpetrator’s Sex using a SVM was the same as for the Logistic Regression as documented above. The data frame dimensionality also included the same columns: Year, Victim Ethnicity, Victim Sex, Victim Age, and the Victim Race.

Following the preprocessing stage, a SVM model yielded no results as it seemed to run endlessly and as a result no results are available.

* 1. Predicting a Perpetrator’s Race using Logistic Regression

The preprocessing stage for predicting a perpetrator’s stage involves discarding the rows where the crime has not been solved, the perpetrator race is unknown, and the victim sex is unknown as documented in Prediction on Homicide Reports, 1980-2014 by UC San Diego (Liu, et al., 2018). Also the data frame dimensionality was reduced to 2 columns as specified above in 2.4.2. Following this journal, predicting a Perpetrator’s Race using Logistic Regression yielded an accuracy score of 0.865981092643%.

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

Following the preprocessing stage and dimensionality reduction of the data frame as documented above for Logistic Regression the Multi-Layer Neural Network produced the results documented in the image below in figure 5. The precision, recall, f1-score and support values are listed for each Perpetrator Race.

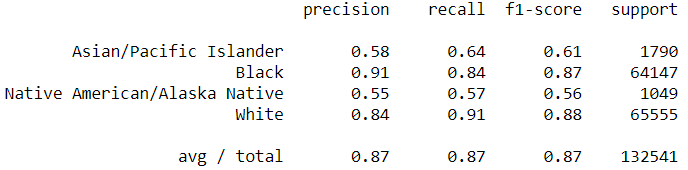


Fig 5. Perpetrator Race prediction using a Neural Network

1. Conclusion

Homicide remains as a large issue in the USA. Data Mining and machine learning techniques offer an insight into developing the most likely profiles of perpetrators for the justice system and police intelligence. The result of predicting a Perpetrators Sex using a Logistic Regression Model proved accurate 89% of the time. Predicting a perpetrators Race using a Logistic Regression Model proved accurate 86% of the time. Using a multi layer perceptron neural network to predict a perpetrator’s race proved slightly more accurate than a logistic regression model at 87% accuracy.

1. Source of Evidence

A GitHub repository containing a Python notebook of the conclusions found can be seen at the following link: <https://github.com/AoifeNicAntSaoir/BigData_KDDHomicide/blob/master/BigDataCa.ipynb>

# References

Fox, J. A., The Lipman Family Professor of Criminal Justice, N. U. & Wawitz, M., 2007. Homicide Trends in the United States. *BJS: Bureau of Justice Statistics ,* p. 183.

Kaggle, 2018. *Murder Accountability Project.* [Online]   
Available at: https://www.kaggle.com/  
[Accessed 2 May 2018].

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

Loeber, R. et al., 2005. The prediction of violence and homicide in young men. *Journal of Consulting and Clinical Psychology,* 73(6), p. 107.

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

Rothenberger, S. D., Wenskovitch, J. E. & Marai, G. E., 2011. Pexel and Heatmap Visual Analysis. *: Visual Analytics Science and Technology (VAST), 2011,* p. 2.

usnews.com, 2017. *U.S. News & World Report Unveils the 2017 Best Places to Live.* [Online]   
Available at: https://www.usnews.com  
[Accessed 2 May 2018].