Crime Analytics- Report

By Jayati Thakor

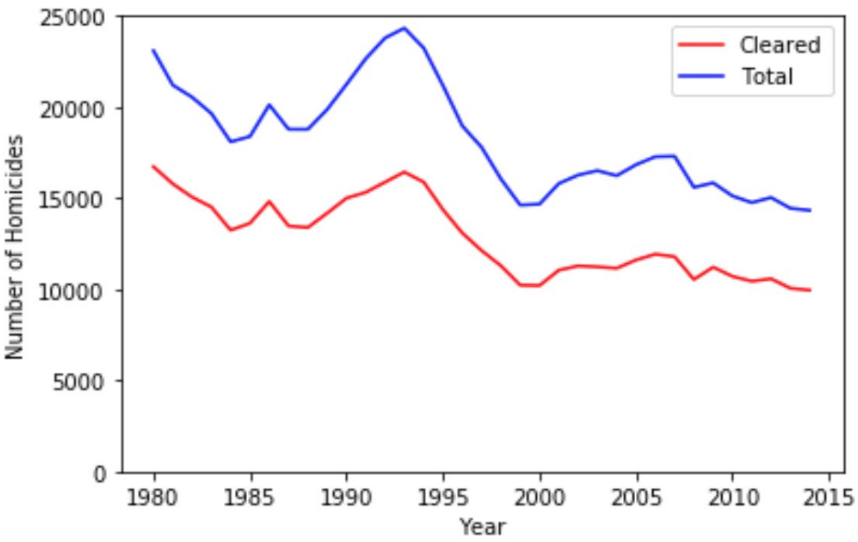
Problem Statement:

The main goal for this project is to explore the factors contributing to the declining clearance rates of Homicides in the USA. My hope is to use machine learning models to predict if the case will be solved or not. My main priority here is to see what factors contribute to the outcome.

Introduction and Background:

The project was inspired by my interest in true crime, and the efforts of a non-profit organization called Murder Accountability Project(MAP). USA does a poor job tracking and accounting for its unsolved homicides. There are more than 18000 different law enforcement agencies that don't collaborate or exchange information, and they perform little to no analysis. Many law enforcement agencies do not report to any higher authority. All the data reported to FBI is on a voluntary basis. Thus there might be many cases missing from the dataset because they were never recorded and reported. All of these factors make it very hard to see patterns and identify serial offenders. The rate of clearance for homicides had been steadily declining over the years.

More than 200,000 homicides remain unsolved, and about a third of the murderers go free. Moreover, with every passing year, 5000 new homicides go unsolved. No law enforcement agency in America is assigned to monitor failed homicide investigations or to investigate the reasons behind it. This project hopes to gain some insights into these unsolved cases.



Dataset:

There were various datasets used for this project. The main dataset was from kaggle, which was a clean version of the case-level data from Supplementary Homicide Report from FBI with additional 30,500 homicides obtained through the Freedom of Information Act for homicides not reported to the Justice Department. Another dataset was used to get clearance rates for agencies, this dataset was from Uniform Crime Report data from FBI, summarizing all homicides and homicide clearances reported from 1965 to the present. The final dataset after cleaning had 650,000 rows and 26 columns. Some of the important columns of the dataset can be classified as follows.

Agency info	**Case detail**	**Victim info**	**Perpetrator Info**	**Weapon**	**Relationship**	**Crime**
-Type	-City	-Sex	-Sex	E.g. :	E.g. :	***Solved or not (0/1)***
-Code	-State	-Age	-Age	-Knife	-Husband	
-Name	-Month	-Race	-Race (unknown when case is unsolved	-Rifle	-Son	
	-Year			-Firearm	-Stranger	

EDA and Cleaning:

Cleaning the data involved converting as many columns to numeric as possible. I used dictionaries and mapping to convert some categorical values to numerical. Cleaning also included removing some null and nan values and finding infinity values. Outliers were removed as they only belonged to a small number of observations. Data transformation included incorporating external data to get clearance rate for each agency to the original dataset.

EDA was mainly performed to see what states have a high percentage of unsolved cases. This was done by grouping the data by state and plotting it on a map using plotly.

Modelling:

For this project, Interpretation was a bigger priority than Accuracy. Accurately predicting the outcome does not help the case get solved. So the main goal was to see what features are strong predictors. I ran several classification models on the dataset like Logistic Regression, Decision trees, Random forests and a few more. I decided to stick with Logistic regression as other models did not show a drastic improvement in accuracy and do not offer a high level of interpretability.

Results:

The coefficients of the model shows what features are strong predictors for the outcome of the case. The biggest predictor is the clearance rate of the agency which shows its effectiveness. Consistently underperforming states and agencies contribute to an unsolved homicide. Additionally, as suspected, there is some gender and racial bias involved. Certain kinds of weapons are also harder to track and they contribute heavily to the outcome.

	STATE	WEAPON	AGENCY	VICTIM
Positive predictors	South Carolina, Illinois, Tennessee, Alabama	Rifle, Shotgun, Knife, Blunt object	County Police, Municipal Police	RACE: White, SEX: Female
Negative predictors	New York, DC, Maryland, California	Unknown, Firearm, Strangulation, Fire	State Police Sheriff	RACE: Hispanic, SEX: Male

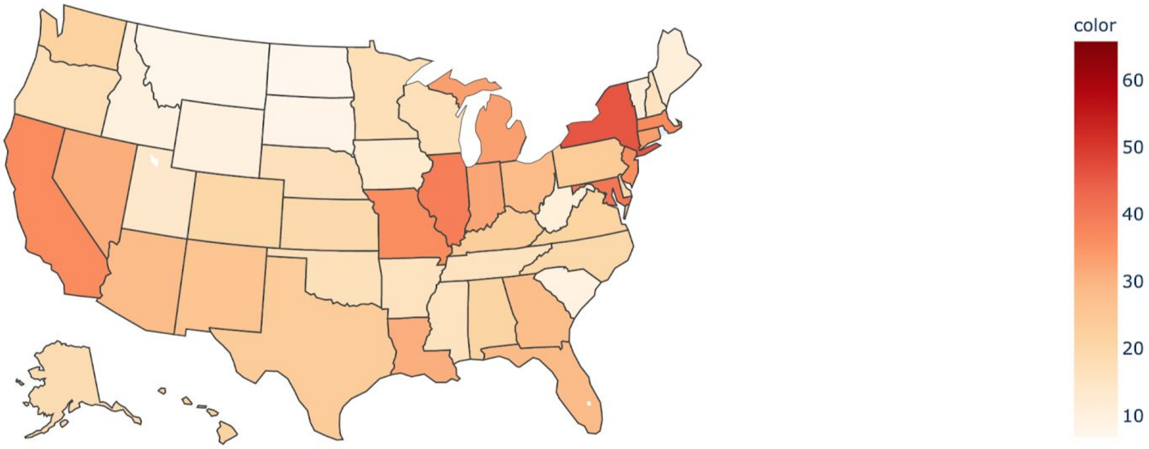
Potential for these findings:

Such findings can be very valuable to bring attention to hundreds of thousands of unsolved homicides that are forgotten. These results can be used to hold law enforcement agencies accountable for their shortcomings, and to emphasize the need for an overarching authority that deals with these unsolved cases, and makes reporting mandatory for all agencies. It can also be used to allocate resources and budgets more effectively, to make sure struggling agencies get the support they need. Better legislation for tracking gun sales is also needed to make sure it can be tracked to the registered buyer. We also need to bring to attention the gender and racial bias that plays a role, some protocols need to be established for cases that might be typically ignored.

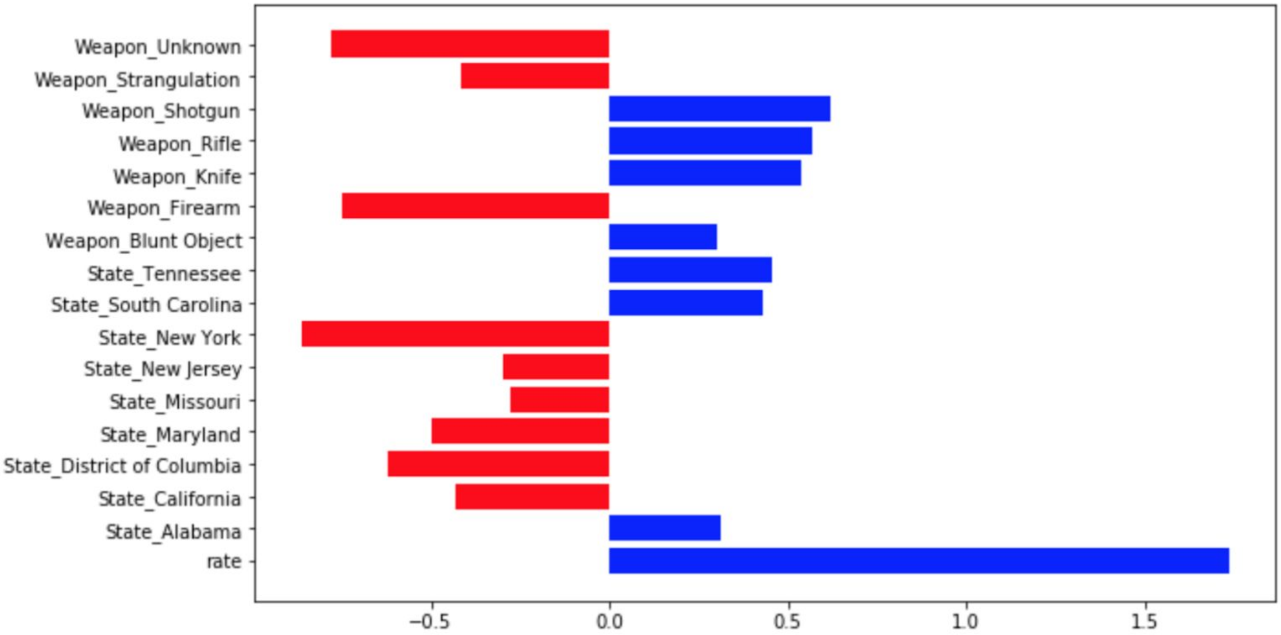
Future direction:

There is plenty of potential to perform advanced data analysis for historical crime data. Similar analysis can also be performed for missing person cases or other violent crimes to predict their outcome and contributing factors. Augmenting the dataset to include more geographical features can offer more insight to the patterns of homicides and to find local clusters of serial killers.

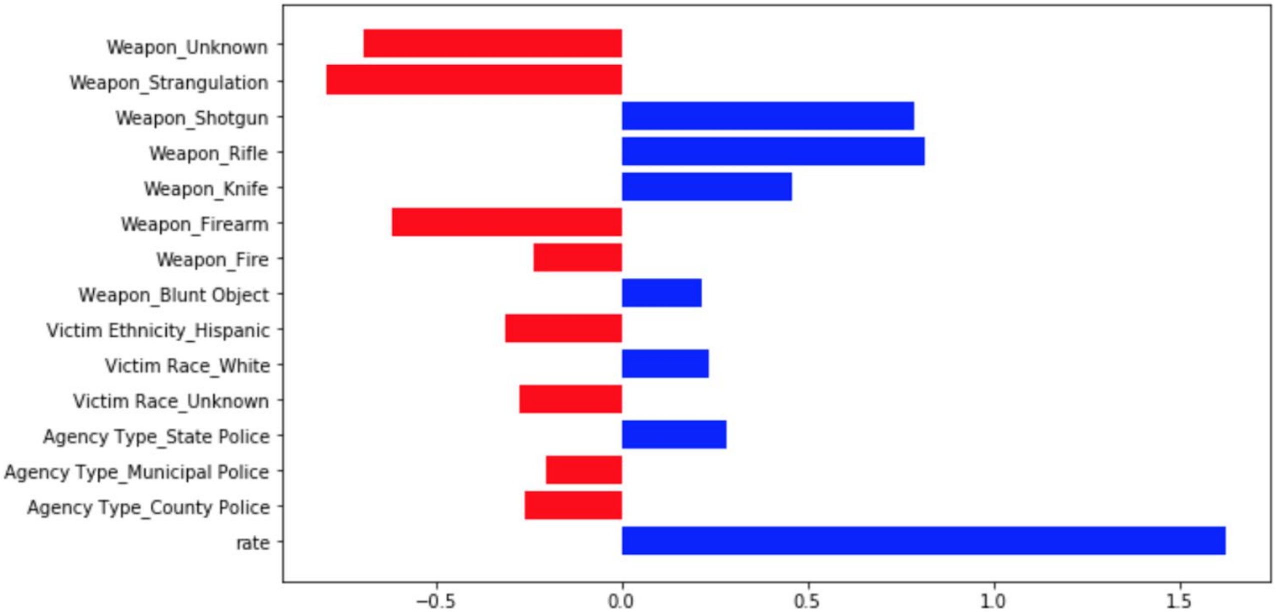
Appendix for figures



Percentage of Unsolved Homicides by state



Feature strength by coefficient values of Logistic regression model



Feature strength (without state) by coefficient values of Logistic regression model