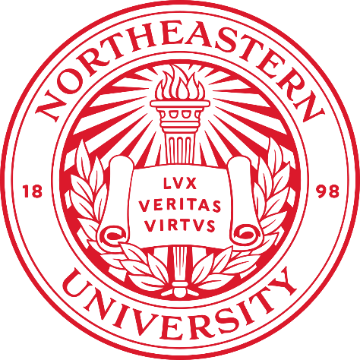
**Northeastern University**



**Module 6 Assignment: Capstone Project Report**

**Course :** ALY6140 Analytics Systems Technology

**Term:** Winter 2023

**Submitted By: Team 2**

Sreedevi Ravi

Rushikesh Prakash Sawant

**Submitted On:**

31st March 2023

**Submitted To:**

Professor Daya Rudhramoorthi

College of Professional Studies, Northeastern University, Boston, MA

This report is regarding Week 6 Assignment.

# Introduction

The purpose of this project is to find patterns and insights that can help Boston's law enforcement and crime prevention efforts. To accomplish this goal, the Crime Incident Reports dataset, which includes details on reported crimes from August 2015 to the present will be examined. We have selected the Crime Incident Reports – 2023 to Present dataset, which is the latest version available on the Boston Government website. This dataset contains crime incident reports from the year 2023 to the present day, sourced from a new system.

To uncover hidden trends in the dataset, we will employ visualization techniques like bar plots, geographic maps, and line graphs. In addition to these methods, we may also use statistical modeling and machine learning algorithms to create predictive models that can aid in identifying regions or communities where crimes are more likely to occur. In order to develop targeted crime prevention strategies, these models may also help identify specific crime types that are more likely to occur in particular locations.

# Business Questions

The business questions that has been formulated to find insights from the dataset are:

1. What are the most common types of crimes reported in Boston, and how have they changed over time?
2. Are certain areas of the city more prone to crime than others, and what factors contribute to this?
3. How does the frequency of crime incidents vary by time of day, day of the week, and month of the year?
4. Can the likelihood of a location or district having a crime due to a shooting be predicted?

# Exploratory Data Analysis

The data is provided in Comma Separated Values (CSV) format and comprises 11207 rows and 17 columns. Each row in the dataset corresponds to a specific crime incident and the columns provide various related attributes.

## Data Extraction

The dataset was loaded from the website using an API call. The data is received in JSON format which is loaded into csv format inside the function.

Figure : Function which gets the dataset using an API call.

Text

Description automatically generated

Figure 2: Glimpse of the columns in the dataset with the percentage of missing values

Table

Description automatically generated

## Data Cleanup

The columns OFFENSE CODE GROUP and UCR PART have no values, so they were removed from further analysis. The missing values in the columns ‘DISTRICT’, ‘Lat’, ‘Long’ and ‘Location’ were imputed. The missing values for DISTRICT were imputed using the measure mode (most frequent value). After filling in the missing values for 'DISTRICT', the mean latitude and longitude for each district were calculated. The missing values for latitude and longitude within each district were filled in with their respective mean values. This step ensures that each crime incident has a geographic location associated with it, which is important for subsequent visualizations and analyses. Finally, latitude and longitude values were converted into a string format to fill in missing location values.

## Data Visualization

By using a clustering algorithm to group crime incidents into geographic clusters, a map was generated which provides a way to identify areas with high crime rates and potential crime hotspots. This information can be used to inform law enforcement and crime prevention strategies, such as increasing police presence in high crime areas or implementing community policing programs in neighborhoods with a high concentration of crime incidents.

Figure 3: Map of Boston with crime incident clusters

Map

Description automatically generated

A bar plot was generated to display the number of crimes that occur on each day of the week in Boston.

Figure 4: Number of Crimes by Day

Chart, bar chart

Description automatically generated

The plot shows that crime incidents in Boston are most frequent on Monday and Wednesday, and relatively lower on Saturday and Sunday. This information can be used to inform law enforcement and crime prevention strategies, such as increasing police presence on weekends or implementing targeted crime prevention initiatives in neighborhoods that experience higher crime rates on certain days of the week.

It is also necessary to determine where crime due to shootings is likely to occur. The shooting locations were plotted on a map, which shows the street when we hover over the point.

Figure 5: Areas of shooting

Map

Description automatically generated

A treemap was created to visualize the top 5 offense descriptions in the crime dataset. Each rectangle in the treemap represents a different offense description, with the size of the rectangle indicating the frequency of the offense.

Figure 6: Treemap for offenseChart, treemap chart

Description automatically generated

A line plot showing the trend of crimes by date, which can imply how crime rates have changed over time was created. By visualizing the data, we can see whether crime has increased or decreased during certain time periods and identify any patterns or trends.

Figure 7: Trends of crimes

Chart

Description automatically generated

The crime rate saw a dip in mid of February but peaked again in March.

Figure 7: scatter plot for robberies in Boston

Map

Description automatically generated

The scatter map plot shows a concentration of red circles, which represent the locations of robberies, in Boston city. This concentration of robberies is clearly visible when looking at the scatter map plot, as the red circles are clustered together in a small geographic area. By analyzing the scatter map plot, it can be inferred that downtown Boston has a higher rate of robberies than other areas in the city, based on the number of red circles plotted in that area compared to other parts of the map.

Figure 8: bar plot for top 5 crimes in Boston

Chart, bar chart

Description automatically generated

From the bar plot of crimes in Boston, it can be observed that crimes involving investigations of persons and sick assistance are the most frequently occurring. These two categories stand out prominently, indicating a potential need for increased attention and resources to address these types of crimes in the city.

Figure 8: Crimes in each month

Chart, line chart

Description automatically generated

From 2023 dataset we can infer that January had the highest number of crimes in the entire year followed by February with significant decrease in March month.

Figure 8: Trends of crimes

Chart, bar chart

Description automatically generated

A bar plot comparing crime rates during daytime and nighttime in Boston reveals that overall, crime rates are significantly higher during the day. Crimes related to investigations of persons and sick assistance also exhibit higher rates during daytime, highlighting potential hotspots and areas of concern for law enforcement and public safety officials. The findings suggest the need for targeted strategies to address the unique challenges of daytime crime prevention and management.

Figure 8: HEATMAP for shooting incidents.

Map

Description automatically generated

The heatmap of shootings in Boston provides a clear visualization of the distribution of shootings across the city, with hotspots appearing in certain neighborhoods. The heatmap can help identify high-risk areas and inform targeted interventions to prevent and reduce gun violence.

# Predictive Models

1. Logistic Regression model to predict if shooting will happen or not:

This model aims to predict whether a shooting will occur or not, based on relevant attributes such as location, time, and other factors using logistic regression. Logistic regression is a statistical model that can be used to predict binary outcomes, such as whether a shooting will happen or not.

1. Decision Tree to predict the probability of shooting:

The categorical variables in encoded using the LabelEncoder() function from the sklearn.preprocessing module. The columns that are encoded are DISTRICT, DAY\_OF\_WEEK, and STREET which were chosen as the independent variables. The target variable is SHOOTING.

The train\_test\_split() function from the sklearn.model\_selection module were used to split the data into training and testing sets. The testing data is 20% of the original data, and the random state is set to 42 for reproducibility. The aim of this model is to find the likelihood of a location or district having a crime due to a shooting.

The Logistic Regression model had an accuracy of 99.15%.

Chart, line chart

Description automatically generated

An ROC curve area of 0.48 indicates that the model has limited predictive power. This suggests that the model is capturing some relationship between the predictor variables and the target variable, but there may be room for improvement through fine-tuning or incorporating additional data.

Chart

Description automatically generated

The confusion matrix showed that the model correctly predicted 3158 instances of the negative class (no shooting) and 0 instances of the positive class (shooting) (true positives). Moreover, the model predicted zero instances of the negative class as positive (false positive) and 27 instances of the positive class as negative (false negative) (false negatives).

The decision tree generated had an accuracy 99.12%. Since the data for shooting is imbalanced, the class weight for each class was set inversely proportional to its frequency in the dataset. The class weight for no shooting was set as 0.1 and for shooting occurred was set as 0.9.A picture containing text, sale, marketplace, different

Description automatically generated

Figure 8: Decision Tree for shooting

The confusion matrix showed that the model correctly predicted 3157 instances of the negative class (no shooting) and 0 instances of the positive class (shooting) (true positives). However, the model predicted one instance of the negative class as positive (false positive) and 27 instances of the positive class as negative (false negative) (false negatives). Overall, the model appears to perform well in predicting the negative class but poorly in predicting the positive class.

On running a feature selection, the significant feature returned was District.

Figure 9: Feature selection for decision tree

# Chart, bar chart Description automatically generatedApplication of Predictive Models

Based on historical data on shooting incidents and other relevant factors such as demographics and socioeconomic indicators, Boston police can use these predictive models to identify high-risk areas for potential shootings. This information can be used to inform targeted patrols and interventions in these areas in order to prevent shootings from occurring. Predictive models can also be used to forecast the likelihood of future shootings and provide early warning systems to allow for proactive response and prevention efforts. People can use predictive models to identify high-risk areas in Boston for shootings and avoid or take alternate routes to reduce their exposure to potential violence.

# Conclusion

Based on the comparison of logistic regression and decision tree models for predicting shootings in Boston, it can be concluded that the logistic regression model with an accuracy of 99.15% is marginally better than the decision tree model with an accuracy of 99.12%. While both models demonstrate high accuracy in predicting shootings, the logistic regression model appears to capture the relationships between predictor variables and the target variable more effectively.

We can predict the areas where shooting is bound to occur or not with the district variable. The most common offense type is investigating a person who committed a crime. The maximum number of crimes occur on Monday and Wednesday. The findings from logistic regression and decision tree suggest that the logistic regression model may be a more reliable and accurate tool for predicting shootings in Boston.

# References

Kim, S., Joshi, P., Kalsi, P. S., & Taheri, P. (2018, November). Crime analysis through machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 415-420). IEEE.