Prediction and Analysis of Crime Rate in Boston

DEVANSHI SHAH,

College of Information Studies, University of Maryland, College Park, USA, dshah2@umd.edu

ZACKARY RICHARDS,

College of Information Studies, University of Maryland, College Park, USA, zackaryr@umd.edu

1 INTRODUCTION

In this project, we will focus on the crimes occurring in the city of Boston and try to understand the current trends in the crime pattern by doing predictive modeling on the data related to the crimes.

1.1 Motivation

The Federal Bureau of Investigation (FBI) defines a violent crime as an offense which involves the use of force or a threat to any other individual. The FBI's Uniform Crime Reporting (UCR) program divides the crimes into different categories. Boston is the capital of Massachusetts. Crime within the city of Boston has been on a steady increase in recent years. There are many colleges and universities in Boston and people from many parts of the world are coming to study in Boston. It is also a major hub for startups in recent years. Crime is a big threat to any major city. Our motivation for this project is to analyze the crimes occurring in Boston and taking steps to ensure that the hub for education, business and tourist attraction is safe. This project will help the authorities take preventive measures to safeguard the people of Boston.

1.2 Background

We have taken the Boston crime data from 2015-2019. The exploratory data analysis of the crime data will aid the Boston Police department to understand the current trends in the crime throughout the years. The analysis will help the authorities in identifying the most common type of crime. Crime can be any action or omission that constitutes an offense that is punishable by law. Since the city of Boston is very diverse, there are different types of crime. The analysis will help the BPD to judge which areas are most susceptible to certain types of crime. The authorities will also get a brief idea of which areas might require an increase in policing according to the frequency of the crimes over the day, week, and the year. The study of the crime will also help in identifying whether there are any correlations between the reporting area of the crime and the type of crime.

1.3 Summary of the project

By working as prospective Data analysts for the Boston Police Department, we implemented the exploratory data analysis on the crime data from the year 2015 to 2019. This helped us to predict the severity level of crime in certain areas of Boston City. By studying the count/occurrence of crime in the past according to the hours of the day, days of the week, and months of the year, we derived the insights on the trends of crime for the whole dataset in general. We also used prediction algorithms like Decision Tree Classifier and Random Tree Algorithm to determine the prediction accuracy between the pattern of crime from 2015 to 2018 which was taken as the train set for the algorithm and for 2019 which was taken as the test set for the algorithm.

1.4 Specific Contribution

The search for the crime data from the official Boston site was a joint effort of the team in this project. Zackary focused on cleaning the dataset in Excel and combining it using the `pandas` library in Python. We both were responsible for further data cleaning using Python like dropping the null values, and transforming the whole dataset into correct data type as well as subsetting the dataset according to different years. The data visualization part of the project was done by both members of the team equally. Devanshi used `matplotlib` and `seaborn`. Zack used `altair`, `seaborn`, and `folium`.For the prediction of the data, the dataset was split into the train and test set by Zackary. Devanshi worked on implementation of the decision tree and random forest classifier algorithm for further crime forecasting.

2 LITERATURE REVIEW

According to Bogahawatte and Adikari[1], the usage of data mining techniques like exploratory data analysis, classification and prediction techniques are necessary for effective investigation of crimes. This can be done by developing a system named Intelligent Crime Investigation System (ICSIS). This could help in the identification of criminals based on the evidence collected from the crime location. Significantly such a system could suggest investigative leads to the investigators that might be overlooked by the investigators sometime.

Yu et al.[2] had an active discussion on the preliminary results of a particular crime forecasting model. This model was developed in collaboration with the police department of a city in the Northeast US. The process of the model was to architect the datasets from the original crime records by employing exploratory data analysis and prediction modeling techniques for better crime forecasting. We make use of these techniques to decide on the best possible method to find the crime trends for the city of Boston.

Sharma [3] proposed the idea of using prediction algorithms like decision tree classifier and random forest classifier algorithm to portray crime in society. To detect the suspected criminal activity, he suggested that the investigators prepare an active application aimed at analysis of crime and detection of patterns in crime. The use of this kind of technique can be implemented after dividing the data into training and testing sets and performing the test on the testing set for better classification and prediction.

With the use of the above mentioned techniques, we have focused more on the exploratory data analysis and data visualization aspect of the project instead of just focusing on the predictive modeling techniques. Only the use of statistical and predictive modeling will limit the target audience of our project. The target audience might need a prior background knowledge in the technology to just understand the modeling part of the project. So to make sure our project is being understood by an audience of diverse backgrounds, we are heavily relying on the data visualization aspect of the project by using python libraries like Altair, Folium, etc to create user friendly visualization. This would help the audience to get a better grasp on the crime patterns in general.

3 **METHODOLOGY**

3.1.1 *Collection*

The data that we used in our project came from the Boston Police Department. The first steps in this project involved exploring our data to understand how it was gathered and cleaned. We wanted to know where this data came from, how it was procured, and how it was prepared(if at all) before it was uploaded to their website for public use.

3.1.2 *Cleaning*

In this stage, we were able to identify some abnormalities in the data by visualizing it using the `missingno` Python library. This library visualizes the missing values within a dataset. This visualization exposed that in the crime dataset for 2019 was missing values for the column `OFFENSE_CODE_GROUP` and that the values for whether or not a shooting had occurred were reported differently in that dataset. When we saw this, we knew that our datasets were going to need some additional cleaning. Luckily for us, we only needed to clean the 2019 dataset. We changed the values for the `SHOOTING` column from 0 and 1 to null and Y, which was how that column was reported in previous years. In order to populate the `OFFENSE_CODE_GROUP` column, we combined the other datasets together in Excel and gathered the unique `OFFENSE_CODE` and `OFFENSE_CODE_GROUP` pairings and used that as a key for our VLOOKUP function that we used to assign `OFFENSE_CODE_GROUP` values to the 2019 dataset.

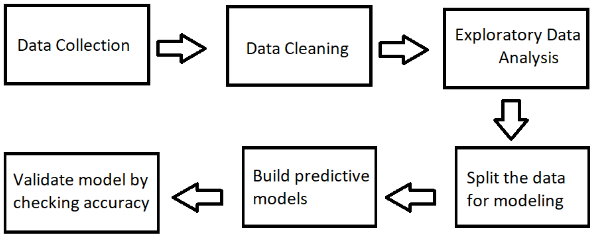
3.1.3 *Curation*

When finished, we combined all of the years' data together into one large dataset. Once we had our dataset cleaned and ready to use, we started brainstorming Python libraries that we wanted to use in order to complete our project. The list of libraries and their reasons for use are as follows:

- `pandas`: data manipulation and analysis
- `numpy`: data manipulation and analysis
- `math`: data manipulation and analysis
- `matplotlib`: data visualization
- `seaborn`: data visualization
- `altair`: data visualization
- `datetime`: data manipulation
- `sklearn`: data modeling
- `folium`: data visualization
- `missingno`: data visualization

When we first set off on this project, we knew that a big part of what we were creating was going to be the included visualizations. We wanted to use libraries that would visualize multiple dimensions of the data or that would offer functionality that would aid the understanding of the data(like a tooltip when you hover over the graph). Thus, this is what led us to choosing `matplotlib`, `seaborn`, `altair`, and `folium` for our visualizations. After we built our visuals, we then turned our attention to building our predictive model. Part of this project was to roleplay as a data analyst for the Boston Police Department. We tasked ourselves with building a model that would assist the department in preparing for crime in the future. We settled on making a model that would be able to predict the reporting area of crimes. This was in hopes that this model could be used to predict the reporting area of crimes and what crimes will be most heavily associated with those areas. We did some research on models that work well with large datasets like ours with some seasonality and we decided on using the decision tree and the random forest models.

3.1.3 *Solution Framework*


Data Collection
Data Cleaning
Exploratory Data Analysis
Split the data for modeling
Build predictive models
Validate model by checking accuracy

Solution Framework

4 **RESULT**

4.1.1 *Experiment Setup*

The result of our semester long work on our progress is a collection of documents that include this report, a Google Colab, and a “readme.txt” file. The only experimental aspect of our project would be the predictive model that we created. Thus, that is what we will be focusing on in this section. The setup to attempt to create an acceptable predictive model was very simple. We tried various methods of splitting the training and test data. At first, we did about a 80%-20% split of all the data from all years. We scrapped this idea in favor of splitting the data by year, as we thought that would be a much more realistic way of training the model. This is how we landed on our final method of splitting the data up by year. 2015-2018 would be used to train the model and 2019 would be used to test it.

4.1.2 *Results*

The result for our decision tree and random forest models were an accuracy of 72.09% and 56.41% respectively. There were 879 reporting areas as a whole.

4.1.3 *Findings*

We found that the most effective model for prediction was the decision tree model. We thought this was interesting as usually in situations like ours where you have a large dataset, random forest generally performs better. We think that this could be due to a couple of reasons, mainly because our laptops do not have enough processing power to leverage the power of the random forest model and that it is possible the random forest model could be benefiting a bit from overfitting of the model.

5 **LIMITATION**

We believe that a couple aspects of our project were limited. One limitation, as we mentioned before in this report and our project presentations, is processing power. For example, the random forest model creates 100 trees by default. We had to lower the number of trees to 4 or less, otherwise our machines would crash. We believe this limitation affected the accuracy of the random forest model. This limit on processing power also limits the amount of data that we can work with. This in turn, limited the scope of our project too. For example, it would be amazing if we could have included more states, or even the entirety of the United States in our project, however this is impossible. We were also limited by time. We initially wanted to host our project on a webpage, however we did not have enough time to do so.

6 CONCLUSION AND FUTURE WORK

In today's advanced era, many techniques can be found which consist of crime prevention and investigation strategies, but this project presents a way of finding criminal records using existing evidence when there are no witnesses present. Our present work in this project accounts for some major steps. After the data cleaning and collection, we visualized the dataset from different perspectives. This helped us to better understand the data and analyze the current trends in data. Furthermore, the prediction modeling helped us to do crime forecasting on the test dataset. We applied two models – decision tree and random forest classifier algorithm. From the results, we can see that the decision tree model has a slightly better performance in terms of accuracy as compared to the random forest algorithm due to time and memory constraints. Random forest algorithm might slightly improve its performance as compared to decision tree model by increasing the number of trees in the model. Both the models did not perform as well as they should have due to limited processing power of personal computers. To improve the accuracy of predictive modeling in our project, we will focus on applying another method to make an improved predictive model, possibly a neural network. We would also like to incorporate demographic information into our visualizations.

REFERENCES

[1] Kaumalee Bogahawatte and Shalinda Adikari. Intelligent criminal identification system | IEEE Conference Publication. Retrieved December 10, 2022 from https://ieeexplore.ieee.org/document/6553986

[2] Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng. Mining Location-based Social Networks for Criminal Activity Prediction. 2015. In Proceedings of 24th IEEE International Conference on Wireless and Optical Communication,185-190. https://ieeexplore.ieee.org/document/7346202

[3] Mugdha Sharma. 2014 . Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree. International Conference on Data Mining and Intelligent Computing. 1-6.. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6954268

[4] Crime incident reports (August 2015 - to date) (source: New system). Analyze Boston. (n.d.). Retrieved December 10, 2022, from https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

[5] Police. Boston.gov. (2016, February 2). Retrieved December 10, 2022, from https://www.boston.gov/departments/police