New York City Crimes Detection using Machine Learning

Adam Ben Rhaeim, Amir Jribi, Mohamed Saket

1. Abstract

Crime is a critical social issue that impacts communities globally, influencing the quality of life, economic stability, and national reputation. In recent years, the escalation in crime rates has highlighted the urgent need for innovative methods to enhance crime analysis and community protection. Effective crime prediction can significantly aid in reducing criminal activities, yet it remains a challenging task due to the complex interplay of factors influencing crime occurrences.

This study focuses on utilizing advanced visualization techniques and machine learning models to forecast crime patterns in New York City. The process begins with preprocessing raw data and employing visualization tools to gain insights into variable relationships and trends. Subsequently, various machine learning algorithms are implemented to predict crime categories based on user-provided inputs. Finally, an interactive user interface is developed to facilitate seamless user engagement, making the system accessible and efficient. This approach aims to contribute to a safer community through data-driven insights and predictive capabilities.

2. Introduction

2.1 General overview

Crime is a pervasive social issue that adversely affects the quality of life, economic development, and the reputation of nations. It significantly influences individual decisions, such as choosing where to live, when to travel, and which areas to avoid. Crimes tarnish the image of communities and impose a financial burden on governments, necessitating additional resources for law enforcement and judicial systems.

The alarming rise in crime rates necessitates proactive measures to mitigate these issues. In New York City, overall index crimes increased by 1.3% in 2021 compared to 2020, with a 15.8% rise in robberies and a 13.8% increase in felonious assaults, despite a 13.7% decrease in burglaries [1]. Such statistics highlight the importance of analyzing and predicting crime patterns to enable preventive actions.

Real-time crime prediction and mass surveillance can substantially reduce crime rates by protecting lives and property. By analyzing historical crime data, patterns and trends can be identified, enabling predictions about future criminal activities. This predictive capability aids in preemptive planning, allowing communities and law enforcement agencies to act proactively.

Criminal behaviors often exhibit repetition, with offenders targeting familiar areas and operating under similar conditions. This tendency provides an opportunity to forecast crimes, though not universally accurate, as studies suggest a high probability of repeated patterns.

This paper introduces a crime prediction framework implemented as a user-friendly web application. The system employs data preprocessing, visualization, and machine learning techniques to analyze historical data and predict crimes based on patterns.

Key steps in the framework include data cleaning and transformation, generating visual insights through reports and maps, and building machine learning models to classify crimes based on location and other factors. These stages, detailed in subsequent sections, aim to provide a comprehensive solution for crime prediction and prevention.

2.2 Review of existing approaches

Various machine learning techniques have been explored for crime prediction, with their effectiveness largely dependent on the nature of the datasets and the chosen predictive features.

In [2], crime detection and classification were undertaken using data compiled from multiple online platforms and newsletters. The study compared the performance of Naive Bayes and decision tree algorithms, finding the former to be more effective. Similarly, [3] conducted an extensive analysis of crime prediction strategies, examining methods such as Support Vector Machines (SVM) and

Artificial Neural Networks (ANN). This analysis highlighted that no single approach could comprehensively address the diverse challenges posed by crime datasets.

The work in [4] applied both supervised and unsupervised learning techniques to crime records, aiming to identify relationships and patterns within the data. This analysis was instrumental in improving the precision of crime prediction. Additionally, [5] explored clustering methods to detect crime trends, while [6] utilized classification algorithms for predictive tasks, contributing further to the field of crime analytics.

3. Dataset

This work is based on the NYPD Complaint Data Historic dataset [7], which includes data on felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to 2019. The dataset comprises 6,901,167 complaints and contains 35 columns, offering a variety of spatial and temporal information related to crime incidents, as well as detailed descriptions and penal classifications.

3.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the process of analyzing and visualizing datasets to summarize their main characteristics and uncover patterns, trends, and relationships. The objective of EDA is to gain insights, identify potential issues, and inform the choice of models and further analyses.

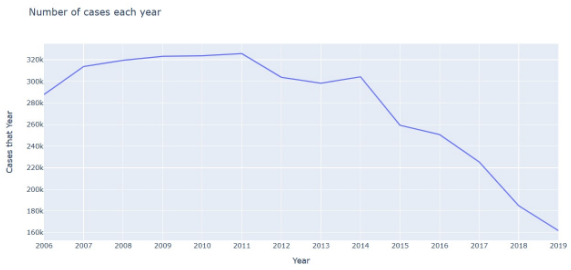

Figure 1 Number of Crimes Reported by Year in New York City

Crime cases peaked in 2011, but it started to decline slowly after that. We see a small rise in 2014, and a hard decline after that, which could be because NYC had elected a new Mayor, Bill De Blasio, and the new Mayor also chose Bill Bratton as the new Police Commissioner. Bill has been named "the city's most significant police leader of the past quarter-century" by the New York Times.

Figure 2 Change in top 5 crime's rate last 5 years

In the Last Five Years, the Top 5 crimes were:

- dangerous drugs
- assault 3 & related offenses
- petit larceny
- vehicle and traffic laws
- other offenses related to theft

We knew that the crime count has been declining in the past 5-6 years, so all the trend lines of the crime slowly declining make sense.

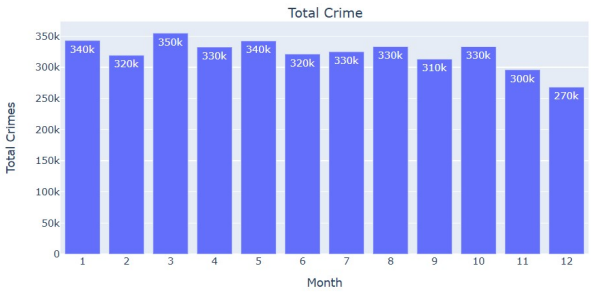

Figure 3 Crime's numbers by months

March seems to be most popular for crime, but we don' t know any particular reason for it. December seems to be least popular month for crime, it could be because it' s a holiday season so people are with - their families.

Staten Island has the least amount of crime throughout the year, it' s pretty steady with very little decline in holiday season. We conclude that Brooklyn contributes with the highest amount of crime. So, we can say that the chance of occurring of a crime is same throughout the year in Staten Island but other boroughs have higher

Figure 4 Crime's numbers by borough

chance of crime occurring in March and less chance of crime occurring in December

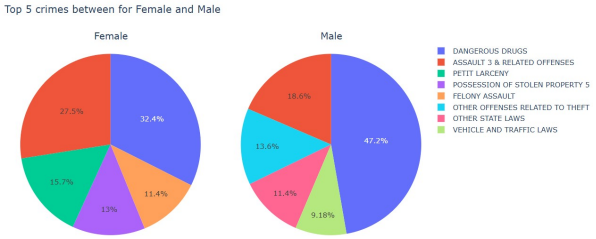

Figure 5 Top 5 crimes for each gender

Both Female and Male commit heavy amount of crimes related to drugs and Assault cases Female population commit more crime related to Felony Assault and Assault 3 Related Offenses Male population commit more crime related to theft and, traffic laws.

Figure 6 Crimes by top 3 age groups in the past 5 years

Age group 25-44 have committed the most crime, and age group 65+ have committed the least amount of crime. But as we saw earlier of how the crime rate is falling, we can see how the crime committed by this 3 age group is also declining. The biggest change we see is for age group of 18-24, where the rate fell by 51.(We see changes in other two groups but its not as drastic change as the 18-24 age group.) Big Drop in crime count from years 2016-2017, could be because new Police Commissioner was assigned Similarly, tables should be formatted as shown in Table **??**.

3.2 Data Cleaning

In the data cleaning process, we focused on removing duplicate entries, handling missing values, and eliminating non-logical data. Duplicate records were identified and removed to ensure data accuracy, while missing values were addressed through appropriate imputation or removal. Additionally, non-logical entries that did not conform to expected patterns or values were corrected or discarded to improve the integrity of the dataset.

3.3 Data Modeling

In this study, we applied the XGBoost algorithm to classify crimes based on their severity. XGBoost, a scalable and efficient implementation of gradient boosting, was chosen for its high performance in handling imbalanced datasets and large volumes of data. It is widely recognized for its effectiveness in various machine learning competitions and real-world applications.

The decision to focus solely on XGBoost was driven by its superior performance compared to other classifiers, as it demonstrated the best balance between accuracy and computational efficiency during preliminary evaluations.

3.4 Experimental Result

The XGBoost algorithm yielded strong performance metrics, including high precision, accuracy, and F1 scores, in classifying crimes based on severity. These results underscore the model's capability to effectively handle the dataset and provide reliable predictions across different crime categories.

Confusion Matrix Analysis

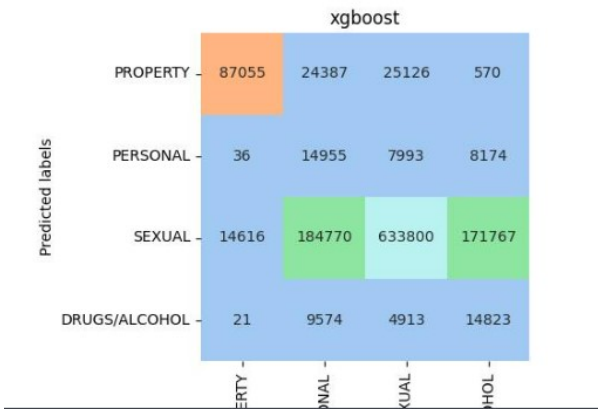

Figure 7 Confusion matrix

The confusion matrix highlights the distribution of pre-

dicted labels versus the actual labels across four crime categories: **PROPERTY**, **PERSONAL**, **SEXUAL**, and **DRUGS/ALCOHOL**.

- **PROPERTY Crimes:**
  - A total of **87,055** instances were correctly classified as PROPERTY crimes.
  - Misclassifications included predictions into the SEXUAL category (**14,416**) and DRUGS/ALCOHOL category (**21**).
- **PERSONAL Crimes:**
  - **14,955** instances were correctly classified as PERSONAL crimes.
  - Notable misclassifications were into the SEXUAL category (**184,770**) and DRUGS/ALCOHOL category (**9,574**).
- **SEXUAL Crimes:**
  - This category achieved the highest correct classification count (**633,800**), showcasing the model's strength in identifying these instances.
  - Misclassifications occurred primarily into the PROPERTY category (**25,126**) and DRUGS/ALCOHOL category (**4,913**).
- **DRUGS/ALCOHOL Crimes:**
  - **14,823** instances were correctly classified as DRUGS/ALCOHOL crimes.
  - Misclassifications occurred into the PROPERTY category (**570**) and PERSONAL category (**8,174**).

Observations

The model performed exceptionally well for the **SEXUAL** category, achieving high true positive rates and relatively lower misclassifications. For the other categories, there were some overlaps, particularly with PROPERTY and DRUGS/ALCOHOL crimes being misclassified into SEXUAL crimes, indicating potential similarities in their features.

This highlights areas where further refinement of features or additional data could improve the classifier's distinction between categories. Overall, the model demonstrated robust classification performance, and the results align with the goals of the study.

4. User Interface and features

For the user interface, we aimed to provide a modern and user-friendly web design to ensure an intuitive experience for users. The interface integrates several key

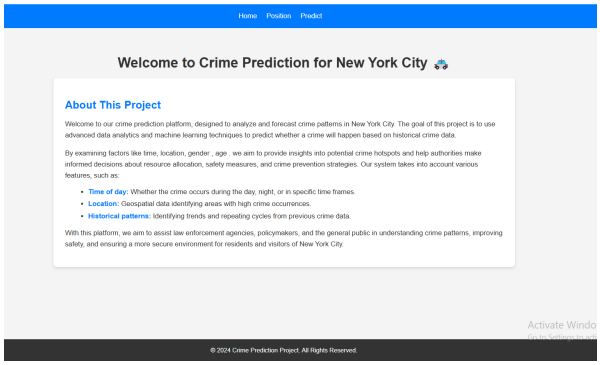

Figure 8 Home page

features to enhance functionality and usability:

- **Interactive Map:** Allows users to input a specific location and assess its safety. This feature provides a visual representation of crime patterns in the selected area.
- **Detailed Results:** Highlights the crime category with the highest probability of occurrence, enabling users to quickly understand the predominant risks in a given location.

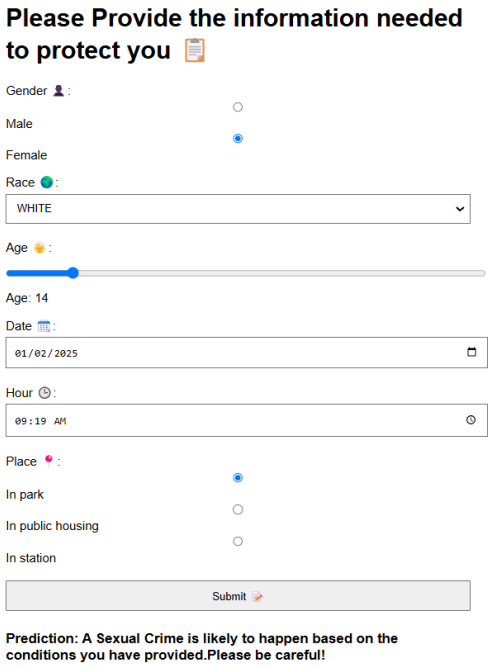

Figure 9 Predictions page

- **Geolocation Services:** Leverages the user's current location to provide tailored insights about the area's safety and crime probabilities.

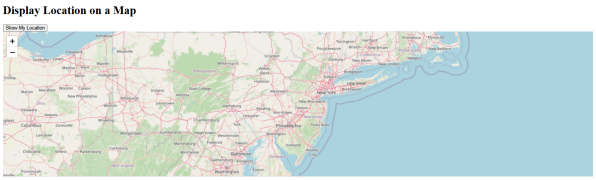

Figure 10 Localisation service

By incorporating these features into a sleek and modern web interface, we ensured that users can access the information they need seamlessly and efficiently, enhancing the overall usability of the platform.

5. Conclusion

The ability to predict and prevent crime has become a critical priority in modern society, with the aim of reducing the occurrence of criminal activities by forecasting potential crime types. This study demonstrated the efficacy of the Random Forest model in identifying and predicting crime patterns. Through a comprehensive analysis, we observed that the model, when optimized, produced highly accurate results, validating its potential as a reliable tool for crime prediction.

The findings highlight the importance of selecting the most appropriate predictive model based on the specific characteristics of the dataset at hand. While the Random Forest model proved effective for this study, it is crucial to recognize that the success of any predictive model is inherently dependent on the quality, structure, and features of the underlying data. Therefore, it is essential to tailor the modeling approach to the unique attributes of the data to achieve optimal outcomes.

Ultimately, the study underscores the transformative potential of machine learning techniques, like Random Forest, in contributing to more informed, data-driven decision-making in crime prevention strategies. Future work should focus on exploring additional models and integrating other data sources to further enhance prediction accuracy and applicability in real-world crime prevention efforts.

6. References

1 New York Police Department (NYPD). NYPD announces citywide crime statistics for October 2021. https://www1.nyc.gov/site/nypd/news/pr1103/nypd-citywide-crime-statistics-october-2021, 2021.

2 Shiju Sathyadevan, Devan M. S., Surya S Gangadharan, "Crime Analysis and Prediction Using Data Mining," International Conference on Networks Soft Computing (ICNSC), 2014.

3 Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma, and Nikhilesh Yadav, "Crime Pattern Detection, Analysis and Prediction," International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2017.

4 Amanpreet Singh, Narina Thakur, Aakanksha Sharma, "A Review of Supervised Machine Learning Algorithms," 3rd International Conference on Computing for Sustainable Global Development, 2016.

5 Bin Li, Yajuan Guo, Yi Wu, Jinming Chen, Yubo Yuan, Xiaoyi Zhang, "An Unsupervised Learning Algorithm for the Classification of the Protection Device in the Fault Diagnosis System," in China International Conference on Electricity Distribution (CICED), 2014.

6 R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. Shariat Panahy, and N. Khanahmadliravi, "An Experimental Study of Classification Algorithms for Crime Prediction," Indian J. of Sci. and Technol., vol. 6, no. 3, pp. 4219-4225, Mar. 2013.

7 New York Police Department (NYPD). NYPD Complaint Data Historic. https://data.cityofnewyork.us/ Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i, 2016.