Illuminating NYC's Crime Landscape with Predictive Precision

1st Akram Dhouib
3rd Aymen Houidi
Higher School of Communication (Sup'Com)
{akram.dhouib, aymen.houidi}@supcom.tn

2nd Oussema Louhichi
4th Soulaimene Turki
Higher School of Communication (Sup'Com)
{oussema.louhichi, soulaimene.turki}@supcom.tn

***Abstract*—In today's society, the escalating concern of crime poses a persistent threat, causing unrest among citizens. The use of artificial intelligence in crime prediction has gained prominence, necessitating the establishment of comprehensive crime databases for future analysis. Forecasting crimes based on factors like time and location is crucial for preemptive measures. Despite the challenging task of accurate prediction due to the rising crime rates, effective crime prediction and analysis methods are essential to identify and mitigate future criminal activities.**

We have experimented with various machine learning methods, highlighting the versatility of machine learning in predicting crimes. This study emphasizes the importance of machine learning to predict and prevent violent crimes in specific regions, contributing to the overall reduction of crime rates.

Key words: Machine Learning, Crime Prediction

I. Introduction

The escalating prevalence and complexity of crime pose a persistent challenge for law enforcement agencies. As criminal behaviors evolve, explaining patterns becomes increasingly difficult. Crimes, ranging from kidnapping and theft to murder and rape, necessitate advanced methods for data collection, especially through Information Technologies (IT). With the surge in crime numbers, the analysis of crime patterns becomes imperative for effective risk reduction.

This paper addresses the need for a crime prediction and analysis tool, providing a practical approach to identify and analyze crime patterns. The proposed methodology focuses on predicting the occurrence of specific crimes at particular places and times, utilizing the Random Forest Classifier. The classification process extracts features and predicts future trends based on similarities in crime data.

The organization of this paper is structured as follows: Section I introduces the study, highlighting the growing challenge of crime. Section II reviews related works, offering insights into existing approaches. Section III details the methodology employed for crime prediction. Section IV discusses the practical implementation of the proposed method. Finally, Section V presents the conclusion, summarizing the significance of the study in advancing crime analysis and prediction methodologies.

II. Related Work

Numerous algorithms for crime prediction have been proposed, each emphasizing the critical role of data type and attribute selection in determining prediction accuracy. In a study referenced as [3], data gathered from diverse sources such as websites and newsletters were employed for crime prediction and classification. The Naive Bayes algorithm and decision trees were utilized, revealing the superior performance of the former. Another comprehensive investigation, as presented in [4], scrutinized various crime prediction methods, including Support Vector Machine (SVM) and Artificial Neural Networks (ANN). The study concluded that no single method universally addresses diverse crime dataset challenges, underscoring the complexity of the issue. In [5], a focus on supervised and unsupervised learning techniques was observed in the context of crime records. This research aimed at establishing connections between crime and crime patterns, contributing to knowledge discovery and subsequently enhancing predictive accuracy. The exploration involved clustering approaches for crime detection and classification methods for crime prediction, as outlined in [7].

III. Methodology

A. Exploratory Data Analysis

Exploratory Data Analysis (EDA) plays a pivotal role in understanding the underlying structures and patterns within the dataset pertinent to this study. By employing various graphical representations and visualizations, EDA aids in uncovering trends, anomalies, and relationships in the data, with the end goal of explicating the phenomena behind the occurrence and distribution of criminal activity.

The exploration begins with a broad analysis of the dataset to ascertain the prevalent trends and notable discrepancies. This initial phase sets the stage for a more granular investigation of specific data attributes.

1) Victim Profile Analysis: An in-depth investigation into victim demographics is imperative for the predictive aspect of this research, which aims to bolster citizen safety. Dissecting the data by victim profiles, such as race and gender, sheds light on which groups are disproportionately affected by criminal acts. The results of this analysis are encapsulated in the figures that follow, each elucidating a different facet of the data.

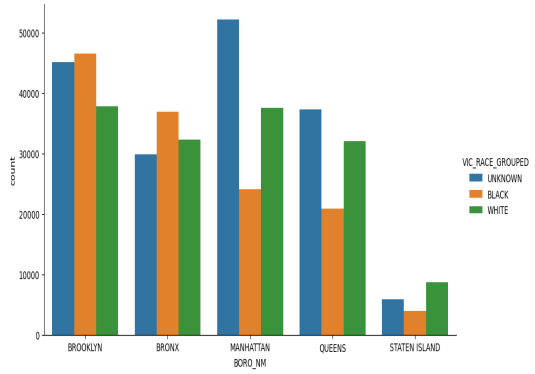

Fig. 1. Number of victims by Race in the different boroughs

Figure 1 demonstrates a significant racial imbalance in victimization rates, with individuals of black ethnicity experiencing a higher incidence of crime compared to other racial groups. A further breakdown by gender uncovers that black females, in particular, stand out as being especially vulnerable to assault.

2) Geo-Spatial Distribution Analysis: The EDA proceeds to integrate geo-spatial data with the previously mentioned demographic attributes, enriching the analysis with a spatial dimension.

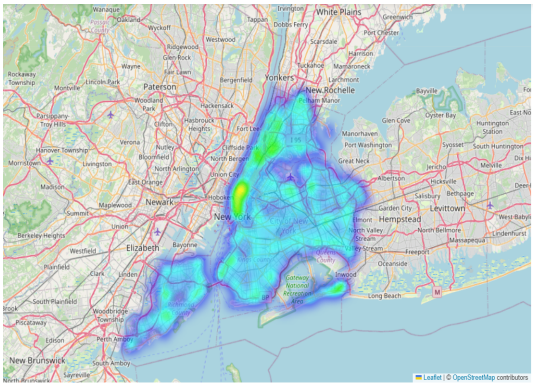

Fig. 2. Crimes Heatmap

As depicted in Figure 2, mapping crime occurrences onto a geographic backdrop reveals clusters and hotspots of heightened criminal activity. This geo-spatial perspective is instrumental in understanding the localized nature of crime, paving the way for targeted interventions and preventative strategies.

The synthesis of demographic and spatial analyses not only enhances the accuracy of crime prediction but also furthers our overarching objective: the protection and well-being of the populace.

B. Data Preprocessing

To address the challenge of imbalanced target data with 63 different crime categories, a two-fold approach was implemented.

1) Handling Imbalanced Target Data: Due to the significant variation in crime frequencies across categories, the following strategy was employed:

a) Preservation of High-Frequency Crimes: To maintain granularity and meaningful distinctions, a threshold was identified, preserving the most frequent crimes as separate categories. The selection of this threshold was based on a pragmatic consideration, such as the top 10 or a predetermined minimum number of occurrences.

b) Clustering Low-Frequency Crimes: Low-frequency crimes, representing a large portion of the dataset, were addressed through clustering.

2) Clustering Low-Frequency Crimes:

a) Text Preprocessing: To prepare the crime descriptions for clustering, text preprocessing was performed. This included converting the text to lowercase, removing punctuation, and filtering out common words (stopwords).

b) Feature Extraction (TF-IDF): The preprocessed text was transformed into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency).

c) Clustering Algorithm (K-Means): The K-Means clustering algorithm was applied to group similar crime descriptions into clusters. The number of clusters was predetermined based on an assessment of the dataset, and in this case, we chose 5 clusters.

d) Assigning Cluster Labels: Each low-frequency crime was assigned to a specific cluster based on the similarity of its description to the cluster's centroid.

e) Naming Clusters: Each cluster was given a name based on the analysis of key terms within the cluster. The naming process aimed to capture the thematic essence of the crimes grouped in each cluster.

This dual strategy allowed for a balance between preserving distinctions in high-frequency crimes and aggregating low-frequency crimes based on their shared characteristics. The subsequent analysis and model training were performed on the refined target data, enabling a more effective and interpretable crime prediction model.

This process not only mitigated the imbalance in our target data but also helped in making the machine learning model's predictions more interpretable. By reducing the number of target classes from 63 to 15, we improved the model's performance by simplifying the problem space and focusing on the most relevant information.

C. SMOTE for Class Imbalance

Class imbalance can impact the performance of machine learning models, especially in crime prediction where certain crime types may be less frequent. To address this issue, we employ SMOTE, a technique that synthetically generates minority class instances.

IV. Model Building and Evaluation

In the pursuit of predictive accuracy within the complex domain of crime data, the Random Forest algorithm was harnessed for its robust performance and adaptability. As an ensemble learning method, Random Forest constructs multiple decision trees during training and outputs the class that is the mode of the classes predicted by the individual trees. This

algorithm is highly regarded for its classification prowess, effectively managing unbalanced datasets and missing data, and it offers a clear advantage through its feature importance evaluation capability.

A. Random Forest Classifier

The Random Forest algorithm serves as the cornerstone of our predictive model, renowned for its efficacy in addressing complex datasets. It is an ensemble technique that amalgamates the predictions from multiple decision trees to improve accuracy and control over-fitting. The algorithm is particularly advantageous in classification tasks for its capability to manage unbalanced and incomplete data, and for its provision of a measure of feature importance.

B. Results

Upon training the Random Forest model, we attained an accuracy of 77%. This metric signifies the proportion of total predictions that were correctly classified, reflecting the model's substantial capability in predicting new instances accurately.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100$$

C. Confusion Matrix

A pivotal aspect of model evaluation, the confusion matrix, provides an in-depth look at the prediction results of the classifier. It offers a breakdown of each actual versus predicted class, allowing for a detailed analysis of the model's performance across different crime categories.

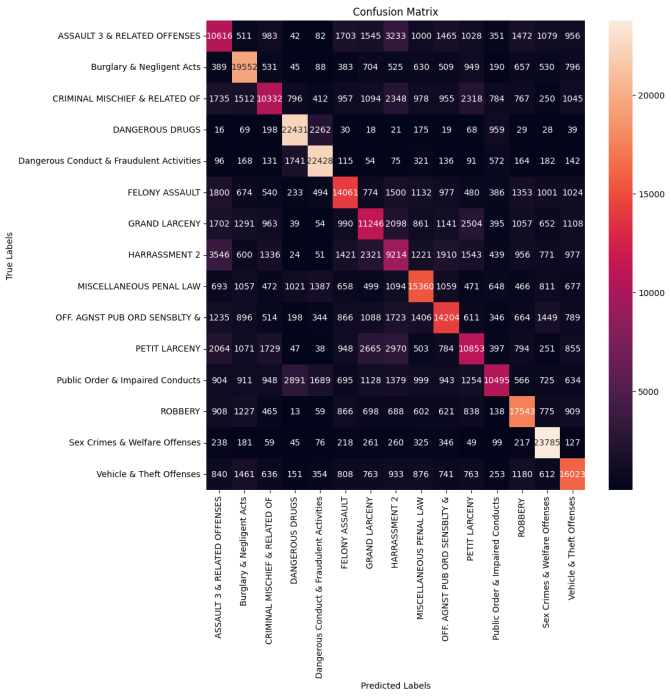

Fig. 3. Confusion Matrix of the Random Forest model

In the provided confusion matrix 3, each row represents the actual classes (True Labels) of crimes, while each column represents the predicted classes (Predicted Labels) as classified by the Random Forest model. The diagonal cells, which represent the number of correct predictions for each class, show high values for crimes like "ASSAULT 3 & RELATED OFFENSES", "PETIT LARCENY", and "GRAND LARCENY". This suggests that the model is particularly adept at predicting these categories of crimes.

However, the off-diagonal cells reveal instances where the model has made incorrect predictions. For example, there are significant numbers of misclassifications between "GRAND LARCENY" and "PETIT LARCENY", which could indicate similarities between these categories that the model is confusing. Additionally, "DANGEROUS DRUGS" seems to be frequently mistaken with other categories, potentially due to less distinctive features within the data or a lower number of training samples for this class.

V. Conclusion

Our study leveraged machine learning, specifically the Random Forest algorithm, to predict and analyze crime patterns in New York City. With a notable accuracy of 77

Exploring victim demographics, geo-spatial distributions, and employing data preprocessing techniques like SMOTE enhanced the model's effectiveness. Acknowledging limitations in human-labeled target data, future research should focus on data quality improvements and augmentation techniques.

The Random Forest model, while robust, suggests ongoing enhancements for real-time data integration and exploration of advanced algorithms. This research contributes to urban safety and sets the stage for further advancements in crime prediction systems.

References

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data",IEEE,Proceedings of the 16th international conference on multimodal interaction, 2014, pp. 427-434Springer Berlin Heidelberg, pp. 319–345, 1999. doi: 10.1007/3-540-46805-6_19.

[2] Ubon Thansatapornwatana,"A Survey of Data Mining Techniques for Analyzing Crime Patterns", Second Asian Conference on Defense Technology ACDT, IEEE,Jan 2016, pp. 123–128.

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

[4] 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

[5] Amanpreet Singh,Narina Thakur,Aakanksha Sharma,"A review of supervised machine learning algorithms",3rd International Conference on Computing for Sustainable Global Development,2016

[6] 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

[7] 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.