New York City Crimes Detection using Machine Learning

Amen Allah Ben Othmane, Youssef Abdelhedi and Mohamed Khalil Drira

***Abstract*— This study focuses on the application of machine learning techniques for crime detection in New York. The initiative involves a user-friendly web app that allows users to input personal information and choose a specific location within the city. Through advanced machine learning algorithms, the system predicts potential criminal activities in the specified New York area. The paper outlines the methodology, model selection process, and development of the Web application, while also addressing ethical considerations and potential social effects. The findings highlight the effectiveness of this approach in raising awareness of crime and helping decision making for both users and law enforcement within the context of New York City.**

I. INTRODUCTION

The rising threats to urban security call for the adoption of changes in technology that enhance crime detection as well as public safety. This research tries to solve this problem using machine learning techniques in the setting of New York City. The work aims at creating a viable model for predicting crimes and placing it into a user-friendly web application. This confluence of sophisticated machine learning with human-centered design seeks to enrich people's lives by providing them with practical information and equipping law enforcement agencies with instruments to prevent and reduce crime opportunities. The purpose of the paper is to explain the construction of the system, the selection of models, and the design and deployment of the web application in ethical terms. Importantly, the research attempts to contribute to the literature on the use of new technologies in preventing crime and promoting the well-being of citizens with a particular focus on the distinct security issues facing New York City.

II. LITERATURE REVIEW

In the case of crime forecasting, several crime forecasting models have been developed, and their prediction accuracy largely depends on the dataset and features being used. In [1], for prediction and classification of crimes, data scraped from various websites and newsletters were used. An Algorithm using the Naive Bayes measure along with decision trees was used, where the former was reported to yield better results. In [2] a review was conducted on numerous approaches toward crime prediction, such as Support Vector Machine (SVM) and Artificial Neural Networks (ANN). They informed that there is no one single solution that can work across different crime dataset problems. In [3] the papers presented the application of supervised and unsupervised methods of learning techniques [4] that were used on crime data with the aim of identifying associations between crimes and crime trends for purposes of knowledge discovery. This explorations aids to improve predictive precision of crime. In [5], clustering techniques were used for crime detection while classification techniques were used for prediction of crime.

III. METHODOLOGY

The goal of this research is to develop a robust machine learning model capable of accurately predicting offense descriptions, categorizing them into Personal, Property, Sexual and Drugs/Alcohol categories,

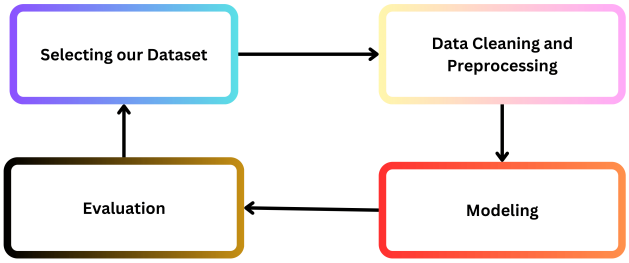


Fig. 1: Workflow

A. Data Collection

Data collection is the process of gathering, measuring, documenting and recording information on specific variables relevant to research, analysis, or decision-making. It serves as a critical step in the research and analysis workflow, with the quality of the data collected playing a key role in determining the accuracy and trustworthiness of the findings or conclusions that follow.

B. Data Cleaning and Preprocessing

Data cleaning focuses on detecting and correcting errors within a dataset, such as managing missing data, removing duplicates, addressing outliers, and standardizing formats. Its goal is to improve the dataset's reliability and consistency by resolving inconsistencies. Conversely, data preprocessing involves converting raw data into a usable form for analysis or machine learning applications. This process may include normalizing numerical values, encoding categorical variables, balancing datasets, engineering features, and addressing challenges specific to particular data types, such as time series or textual data. Together, data cleaning and preprocessing are essential steps for preparing data, enabling meaningful analysis and the creation of dependable machine learning models, thereby enhancing the credibility and relevance of the results.

C. Modeling

Gradient boosting algorithms have gained prominence in machine learning for their effectiveness in predictive modeling. Three notable implementations of gradient boosting—XGBoost, LightGBM, and CatBoost—stand out for their unique features and capabilities.

- **XGBoost (eXtreme Gradient Boosting)** is renowned for its efficiency, scalability, and regularization techniques. It has become a staple in machine learning competitions and real-world applications. XGBoost's key strengths lie in its ability to handle complex datasets, mitigate overfitting, and deliver high performance. It employs a gradient boosting framework that sequentially builds decision trees, continuously improving predictive accuracy.
- **LightGBM (Light Gradient Boosting Machine)** Developed by Microsoft, LightGBM is designed for distributed and efficient training. What sets LightGBM apart is its novel approach to handling large datasets using a histogram-based learning method. This enables faster training times and reduced memory usage, making LightGBM particularly suitable for scenarios where efficiency is crucial. The algorithm excels in capturing intricate patterns in data and is well-suited for applications in both research and industry.
- **CatBoost** CatBoost, short for Category Boosting, is a gradient boosting algorithm developed by Yandex. CatBoost is recognized for its ability to handle categorical features seamlessly without the need for extensive preprocessing. It incorporates a robust handling of categorical variables, making it user-friendly and efficient. CatBoost's optimizations, such as the implementation of ordered boosting and advanced strategies for dealing with overfitting, contribute to its competitive performance in various machine learning tasks.

D. Evaluation

Model evaluation is the process of assessing the performance and effectiveness of a machine learning model based on its predictions or classifications. It is a crucial step in the model development pipeline, providing insights into how well the model is likely to perform on new, unseen data. The goal of model evaluation is to measure the model's accuracy, generalization ability, and suitability for the intended task.

IV. Implementation

A. Data Collection

NYPD Complaint Data Historic This dataset includes all valid felony, misdemeanor, and violation crimes re- ported to the New York City Police Department (NYPD) from 2006 to the end of 2019. The data contains 6901167 complaint and 35 columns including spatial and temporal information about crime occurrences along with their description and penal description.

B. Data Cleaning and Exploratory data analysis

1) Data Cleaning: The dataset underwent thorough preprocessing to enhance its quality and suitability for analysis.

```
Complaint_ID have  0.0  % missing values
Start_Date have  0.009379644624597266  % missing values
Start_Time have  0.0006873632701994943  % missing values
End_Date have  23.89125798504899  % missing values
End_Time have  23.826144635265717  % missing values
Precinct have  0.03101726756775218  % missing values
Report_Date have  0.0  % missing values
Offense_Code have  0.0  % missing values
Offense_Description have  0.26957528253136415  % missing values
Precinct_Code have  0.08398719957750071  % missing values
Precinct_Description have  0.08398719957750071  % missing values
Crime_Status have  0.00010024047690409292  % missing values
Offense_Level have  0.0  % missing values
Borough have  0.1560028221990269  % missing values
Location_Type have  21.19358627060604  % missing values
Premises_Type have  0.5689649469076314  % missing values
Jurisdiction have  0.0  % missing values
JURISDICTION_CODE have  0.08398719957750071  % missing values
Park_Name have  99.6413109335009  % missing values
Housing_Development have  95.04485546540437  % missing values
PSA_Code have  92.31142654084292  % missing values
X_Coordinate have  0.3445981194600131  % missing values
Y_Coordinate have  0.3445981194600131  % missing values
Suspect_Age have  67.41276608297592  % missing values
Suspect_Race have  47.81516572543246  % missing values
Suspect_Gender have  49.72423128800278  % missing values
Transit_District have  97.75054641799964  % missing values
Lat have  0.3445981194600131  % missing values
Lon have  0.3445981194600131  % missing values
Coordinates have  0.3445981194600131  % missing values
Patrol_Borough have  0.09053147071252506  % missing values
Station_Name have  97.75054641799964  % missing values
Victim_Age have  23.46262970580709  % missing values
Victim_Race have  0.0044105809837800885  % missing values
Victim_Gender have  0.004396260915650933  % missing values
```

Fig. 2: Dataset before cleaning

The initial steps focused on managing missing values by either removing columns with a high proportion of null entries or imputing binary indicators for certain categorical variables. Date and time columns were standardized by converting them into datetime objects and excluding rows that lacked crucial temporal information. Additional temporal variables, such as year, month, day, hour, and weekday, were derived from incident dates to provide richer temporal context.

Subsequent cleaning addressed missing or inconsistent values in categorical demographic variables. Missing or unknown entries were imputed to ensure a more complete dataset. Columns deemed redundant or irrelevant were removed to streamline the data. Moreover, a new categorical variable was created to classify crimes into specific categories, enabling a more structured and interpretable analysis.

These preprocessing measures ensured the dataset's consistency and reliability, enhancing its usability and establishing a robust foundation for further exploratory data analysis and modeling efforts.

2) Exploratory data analysis: Exploratory Data Analysis (EDA) is a vital step in the data analysis process, focusing on examining and visualizing data to identify patterns, rela-

tionships, and key insights. Utilizing statistical and visual techniques, EDA helps uncover the dataset's underlying structure, detect potential outliers, and inform subsequent analytical approaches. It provides an essential foundation for understanding the data, supporting better decision-making and the formulation of meaningful hypotheses.

The following visualizations showcase critical patterns and trends identified during the Exploratory Data Analysis, offering a detailed perspective on the dataset's characteristics.

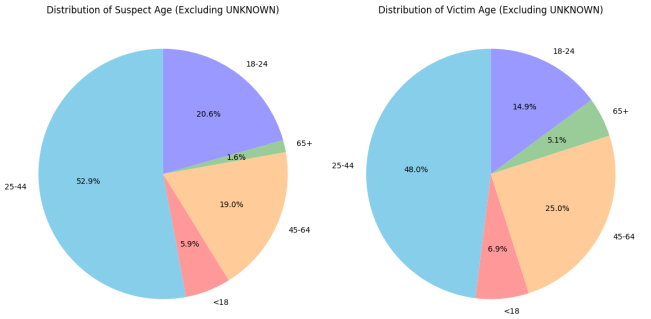


Fig. 3: Age of Suspect/Victim

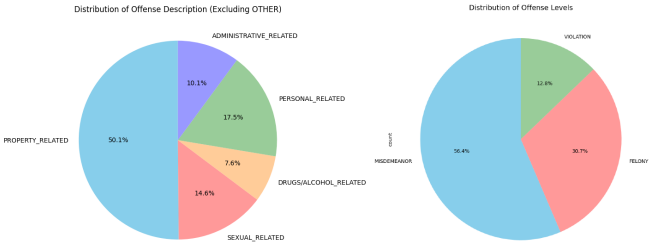


Fig. 4: Crimes Success rate / Level of Offense

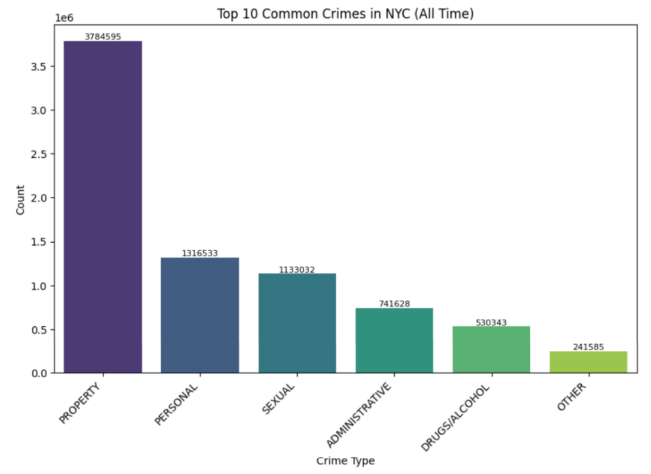


Fig. 5: Top Commun Crimes in New York

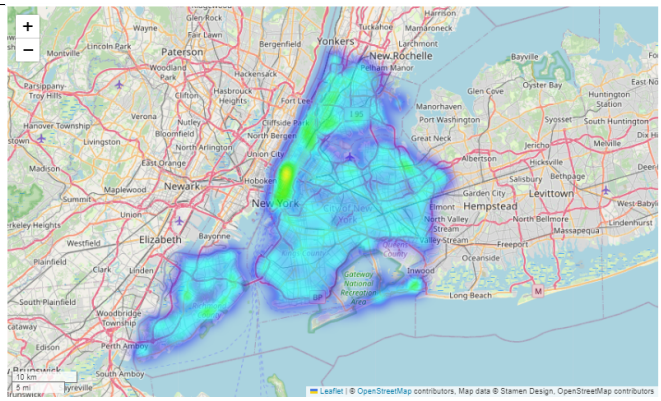


Fig. 6: Crimes Heatmap from NYC

C. Data Preprocessing

Before transitioning to the modeling phase, several preprocessing steps were applied to refine the dataset. Instances associated with specific categories were filtered out, and the target variable underwent encoding to facilitate analysis. Balancing techniques were employed to address class distribution issues, ensuring a more representative dataset.

The feature selection process focused on key aspects, including temporal information, geographic coordinates, and relevant crime-related attributes. Binary classification within certain columns was established to simplify subsequent analyses. Exploratory correlation analysis was performed to understand relationships among selected features, visualized through a heatmap. Categorical variables were encoded, and boolean columns were transformed to streamline compatibility with machine learning algorithms.

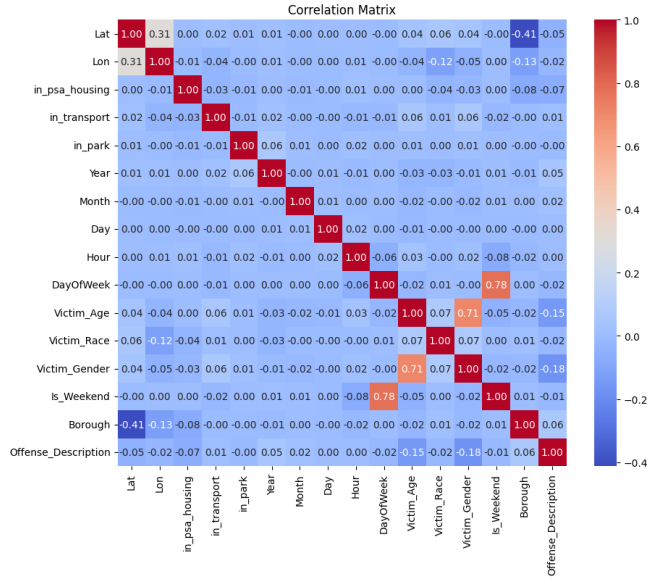


Fig. 7: Numeric Variables Correlation Matrix

D. Modeling

In the training phase, the dataset is partitioned into training and testing sets, with approximately 15% reserved for testing to ensure robust generalization evaluation. Shuffling introduces randomness, and a specific random state is set for reproducibility. This division enables effective model training on one subset and testing on another, facilitating a comprehensive performance assessment. Hyperparameter tuning with Optuna was executed for each algorithm, enhancing model configurations and optimizing predictive capabilities.

The primary objective of the model is to classify and predict the likelihood of specific crimes occurring within categories such as 'DRUGS/ALCOHOL,' 'PROPERTY,' 'PERSONAL,' and 'SEXUAL' Evaluation metrics will gauge the model's ability to discriminate between these crime types, offering valuable insights into its effectiveness in predicting and distinguishing among specific criminal activities.

E. Evaluation and Metrics

- **ROC Curve:**

The ROC curve visually represents the trade-off between sensitivity (true positive rate) and specificity (true negative rate) across various threshold values. In crime prediction, it illustrates how well the model distinguishes between positive (occurrence of crime) and negative (non-occurrence of crime) instances. The area under the ROC curve (AUC-ROC) quantifies the model's overall performance, with a higher AUC indicating superior class discrimination.

- **Confusion Matrix:**

The Confusion Matrix breaks down predictions into true positives, true negatives, false positives, and false negatives. This matrix enables the assessment of precision, recall, and F1 score, providing nuanced insights into the model's performance.

- **Accuracy:** Measures overall correctness using the formula:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

- **Precision:** Quantifies accuracy of positive predictions:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

- **F1 Score:** Balances precision and recall:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$$

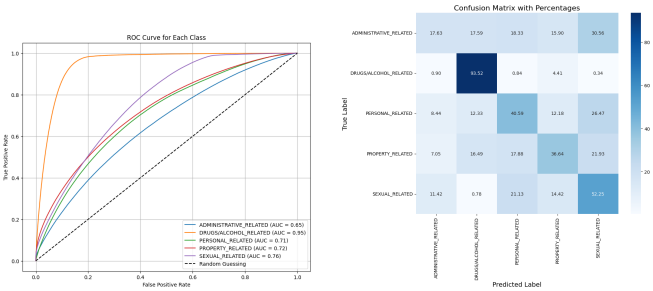
- **Recall:** Ratio of true positive predictions to total actual positives:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

These metrics comprehensively evaluate the effectiveness of our crime prediction model in the specified New York area.

F. Obtained Results

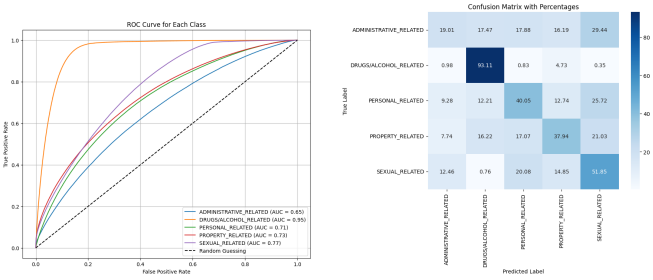
1) LightGBM: This section encompasses the assessment of the LightGBM (LGBM) model.



(a) ROC Curve for the LGBM Model (b) Confusion Matrix for the LGBM Model

Fig. 8: Evaluation Metrics for the LGBM Model

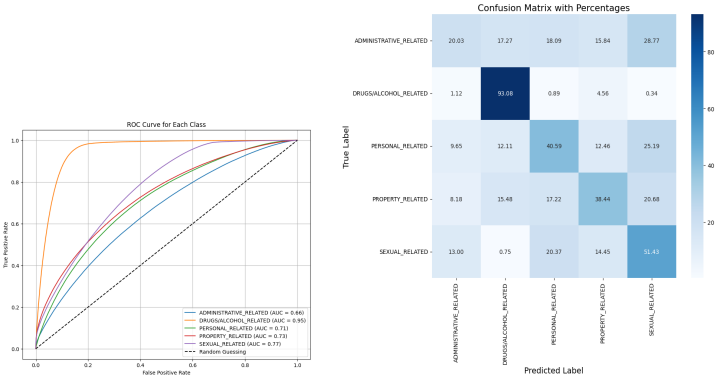
2) XGBoost: This section encompasses the assessment of the XGBoost model.



(a) ROC Curve for the XGBoost Model (b) Confusion Matrix for the XGBoost Model

Fig. 9: Evaluation Metrics for the XGBoost Model

3) CatBoost: This section focuses on assessing the performance of the CatBoost model.



(a) ROC Curve for the CatBoost Model (b) Confusion Matrix for the CatBoost Model

Fig. 10: Evaluation Metrics for the CatBoost Model

G. Models Comparison

As observed, the performance of the three models is quite comparable. Nevertheless, it is noteworthy that the XGBoost model exhibited a slightly superior performance, particularly evident when comparing their confusion matrices.

TABLE I: Comparison of different models

Model	Accuracy (%)	F1 Score
CatBoost	43.2	40.6
LightGBM	42.72	39.63
XGBoost	**43.4**	**40.9**

V. USER INTERFACE

After training and saving the model weights, we created a Streamlit and Folium-based web app for interactive crime prediction. Users provide input on gender, race, age, date, and time, and can select a location on the map, specifying a category like a park, public housing, or station. Destination selection is flexible, allowing users to click on the map or type its name.

This information is then transformed to match the model's input. We utilized various shapefiles to determine the police precinct and borough from coordinates. Subsequently, using the loaded model weights file, we make predictions regarding the type of crime. The predicted crime type, along with potential subtypes, is then sent back to the user. The web application is deployed using Streamlit, and you can access it here.

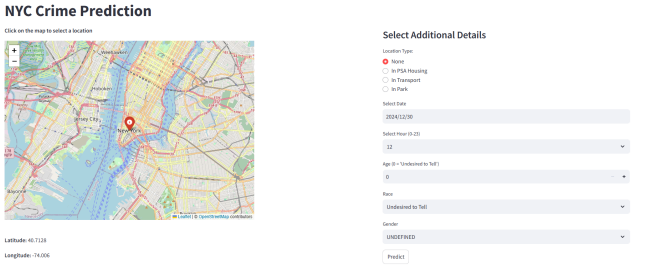


Fig. 11: Prediction when user selects Map Destination

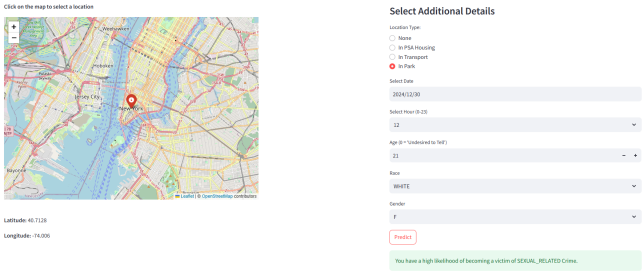


Fig. 12: Prediction when user inputs text destination

VI. CONCLUSIONS

Predicting and preventing crime has become a key priority in modern society. The primary aim of crime prediction is to reduce the occurrence of criminal activities by anticipating the types of crimes that are likely to happen in the future. This study utilizes the Random Forest model to analyze and predict crime patterns. The results highlight the model's effectiveness, especially when trained properly, leading to high accuracy. It is important to note that selecting the most appropriate model depends on the specific characteristics of the dataset, stressing the need to customize approaches based on the unique features of the data for the best predictive results.

REFERENCES

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

[2] 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.

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

[4] 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

[5] 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.