New York City Crimes Detection using Machine Learning

Nour Mabrouk Bacem Ahmed

Abstract— **This project uses machine learning for crime detection in New York, featuring a web app where users can input personal data and select a location to predict potential criminal activities. The paper discusses the methodology, model selection, and app implementation, addressing ethical and societal impacts. Results highlight the approach's effectiveness in improving crime awareness and aiding decision-making for users and law enforcement.**

I. INTRODUCTION

This research aims to enhance urban security by applying machine learning for crime prediction in New York. The project develops a crime prediction model integrated into a user-friendly web app, helping individuals make informed decisions and assisting law enforcement in managing potential crimes. The paper covers the methodology, model selection, app implementation, and ethical considerations, contributing to the use of technology for crime prevention and community safety in urban settings, with a focus on New York's unique challenges.

II. LITERATURE REVIEW

Several crime prediction algorithms have been proposed, with accuracy depending on the data and features used. In [1], Naive Bayes and decision trees were tested, with Naive Bayes performing better. In [2], a study on various methods, including Support Vector Machine (SVM) and Artificial Neural Networks (ANN), concluded that no single method can universally address crime dataset challenges. In [3], supervised and unsupervised learning techniques were applied to crime records to uncover patterns and improve predictive accuracy. Clustering was used for crime detection, while classification methods were applied for prediction in [5].

III. METHODOLOGY

The goal is to develop a robust machine learning model that accurately predicts offense descriptions and categorizes them into Personal, Property, Sexual, and Drugs/Alcohol categories.

A. Data Collection

Data collection involves gathering, measuring, and recording information on variables of interest for research, analysis, or decision-making. It is a crucial step in the research process, and the quality of the data collected directly impacts the validity and reliability of the findings.

B. Data Cleaning and Preprocessing

Data cleaning involves identifying and correcting errors in a dataset, such as handling missing values, removing duplicates, addressing outliers, and standardizing formats. Its goal is to improve data accuracy and integrity. Data preprocessing, on the other hand, focuses on transforming raw data into a format suitable for analysis or machine learning. This includes normalizing numerical features, encoding categorical variables, addressing imbalances, and performing feature engineering. Both processes are essential for preparing data for analysis and building reliable machine learning models, ensuring the validity and effectiveness of the results.

C. Modeling

Gradient boosting algorithms are widely used in machine learning for their effectiveness in predictive modeling. Three notable implementations—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 a machine learning model's performance based on its predictions or classifications. It is a critical step in the development pipeline, offering insights into how well the model will perform on unseen data. The goal of evaluation is to measure accuracy, generalization, and the model's suitability for the intended task.

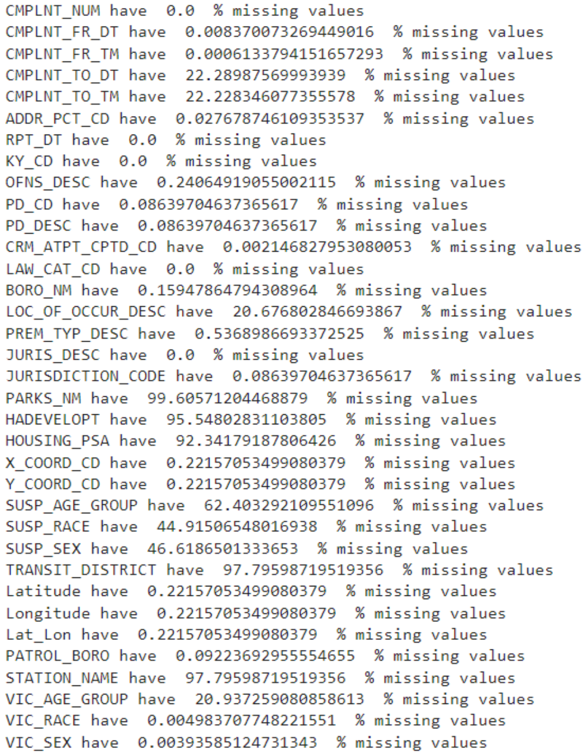
IV. Implementation

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 pre-processing to enhance its quality and suitability for analysis.



```
CMPLNT_NUM have  0.0  % missing values
CMPLNT_FR_DT have  0.008370073269449016  % missing values
CMPLNT_FR_TM have  0.0006133794151657293  % missing values
CMPLNT_TO_DT have  22.28987569993939  % missing values
CMPLNT_TO_TM have  22.228346077355578  % missing values
ADDR_PCT_CD have  0.027678746109353537  % missing values
RPT_DT have  0.0  % missing values
KY_CD have  0.0  % missing values
OFNS_DESC have  0.24064919055002115  % missing values
PD_CD have  0.08639704637365617  % missing values
PD_DESC have  0.08639704637365617  % missing values
CRM_ATPT_CPTD_CD have  0.002146827953080053  % missing values
LAW_CAT_CD have  0.0  % missing values
BORO_NM have  0.15947864794308964  % missing values
LOC_OF_OCCUR_DESC have  20.676802846693867  % missing values
PREM_TYP_DESC have  0.5368986693372525  % missing values
JURIS_DESC have  0.0  % missing values
JURISDICTION_CODE have  0.08639704637365617  % missing values
PARKS_NM have  99.60571204468879  % missing values
HADEVELOPT have  95.54802831103805  % missing values
HOUSING_PSA have  92.34179187806426  % missing values
X_COORD_CD have  0.22157053499080379  % missing values
Y_COORD_CD have  0.22157053499080379  % missing values
SUSP_AGE_GROUP have  62.403292109551096  % missing values
SUSP_RACE have  44.91506548016938  % missing values
SUSP_SEX have  46.6186501333653  % missing values
TRANSIT_DISTRICT have  97.79598719519356  % missing values
Latitude have  0.22157053499080379  % missing values
Longitude have  0.22157053499080379  % missing values
Lat_Lon have  0.22157053499080379  % missing values
PATROL_BORO have  0.09223692955554655  % missing values
STATION_NAME have  97.79598719519356  % missing values
VIC_AGE_GROUP have  20.937259080858613  % missing values
VIC_RACE have  0.004983707748221551  % missing values
VIC_SEX have  0.00393585124731343  % missing values
```

Fig. 1: Dataset before cleaning

Initial steps involved handling missing values by dropping or imputing binary indicators for categorical variables. Date and time columns were standardized, and new variables like year, month, day, hour, and weekday were derived. Further cleaning addressed missing values and inconsistencies in demographic data, and redundant columns were removed. A new categorical column for crime types was added.

These preprocessing steps ensured dataset integrity, improved usability, and set the stage for further analysis and modeling.

2) Exploratory data analysis: Exploratory Data Analysis (EDA) is a crucial step in data analysis, involving the examination and visualization of data to uncover patterns, relationships, and insights. Using statistical and graphical methods, EDA helps identify structure, outliers, and informs further analysis. The following plots highlight key patterns and trends from the dataset.

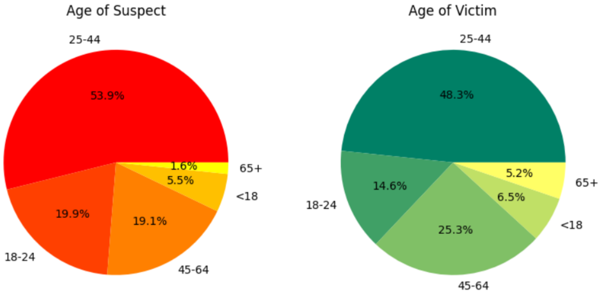


Fig. 2: Age of Suspect/Victim

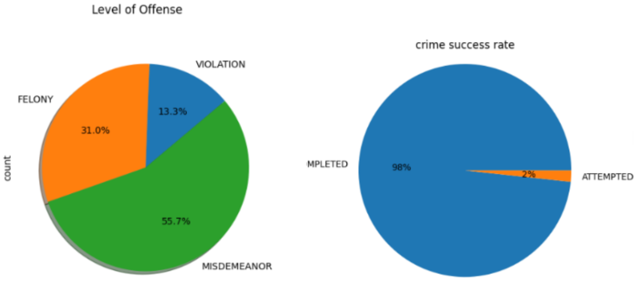


Fig. 3: level of offense / Crimes Success rate

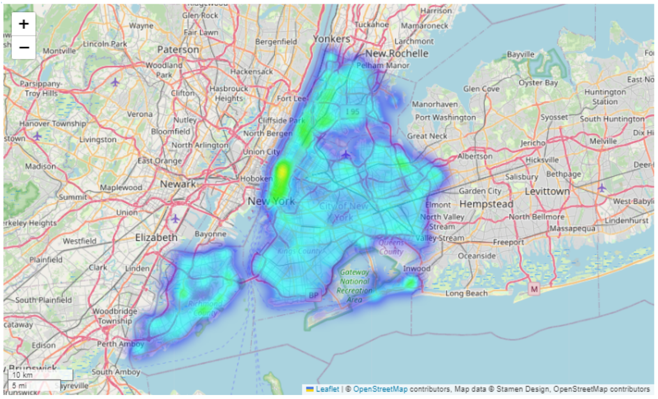

+
−
Yonkers
New Rochelle
Paterson
Newark
New York
Elizabeth
Hempstead
Levittown
Long Beach
10 km
5 mi
Leaflet | © OpenStreetMap contributors, Map data © Stamen Design, OpenStreetMap contributors

Fig. 4: Crimes Heatmap from NYC

C. Data Preprocessing

Before modeling, several preprocessing steps were applied to refine the dataset. Instances were filtered by category, and the target variable was encoded for analysis. Balancing techniques addressed class distribution issues, ensuring a representative dataset.

Feature selection focused on temporal data, geographic coordinates, and relevant crime attributes. Binary classification was applied to simplify analysis, and correlation analysis was conducted to explore relationships among features, visualized in a heatmap. Categorical variables were encoded, and boolean columns were transformed for compatibility with machine learning algorithms.

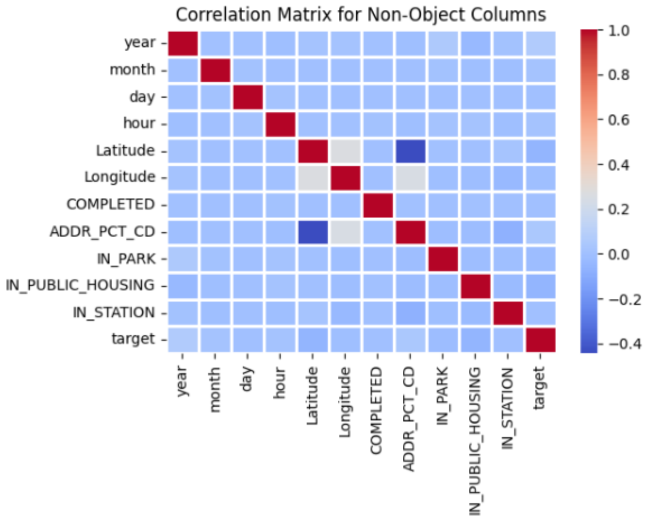


Fig. 5: numeric variables correlation matrix

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) at various threshold values. In crime prediction, it shows how well the model distinguishes between crime (positive) and non-crime (negative) instances. The area under the ROC curve (AUC-ROC) quantifies the model’s overall performance, with a higher AUC indicating better class discrimination.

- **Confusion Matrix:**

The Confusion Matrix breaks down predictions into true positives, true negatives, false positives, and false negatives. It helps assess precision, recall, and F1 score, offering detailed insights into the model’s performance.

- **Accuracy:** Measures overall correctness
- **Precision:** Quantifies accuracy of positive predictions
- **F1 Score:** Balances precision and recall:
- **Recall:** Ratio of true positive predictions to total actual positives

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

F. Models Comparison

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

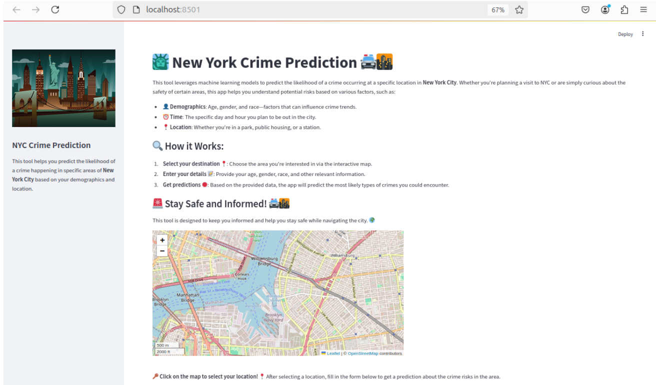
TABLE I: Comparison of different models

Model	Accuracy (%)	F1 Score
XGBoost	61.2	59.64
CatBoost	63.38	61.29
LightGBM	**64.6**	**65.31**

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 .

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 .



NYC Crime Prediction

This tool helps you predict the likelihood of a crime happening in specific areas of **New York City** based on your demographics and location.

🗽 New York Crime Prediction 🚔🌆

This tool leverages machine learning models to predict the likelihood of a crime occurring at a specific location in **New York City**. Whether you're planning a visit to NYC or are simply curious about the safety of certain areas, this app helps you understand potential risks based on various factors, such as:

- 👤 **Demographics**: Age, gender, and race—factors that can influence crime trends.
- ⏰ **Time**: The specific day and hour you plan to be out in the city.
- 📍 **Location**: Whether you're in a park, public housing, or a station.

🔍 How it Works:

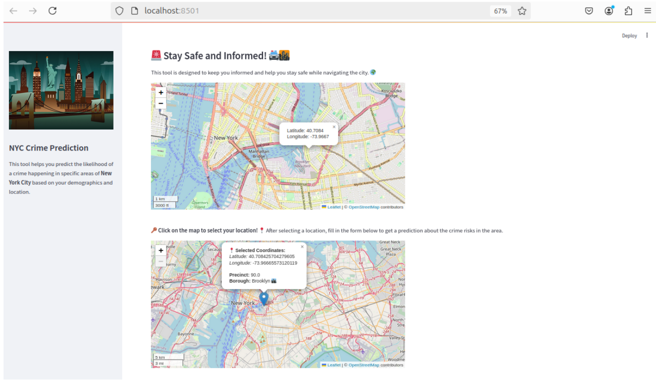
1. **Select your destination** 📍: Choose the area you're interested in via the interactive map.
2. **Enter your details** 📝: Provide your age, gender, race, and other relevant information.
3. **Get predictions** 🔴: Based on the provided data, the app will predict the most likely types of crimes you could encounter.

🚨 Stay Safe and Informed! 🚔🌆

This tool is designed to keep you informed and help you stay safe while navigating the city. 🌍

🔎 **Click on the map to select your location!** 📍 After selecting a location, fill in the form below to get a prediction about the crime risks in the area.

Fig. 6: User Interface



NYC Crime Prediction

This tool helps you predict the likelihood of a crime happening in specific areas of **New York City** based on your demographics and location.

🚨 Stay Safe and Informed! 🚔🌆

This tool is designed to keep you informed and help you stay safe while navigating the city. 🌍

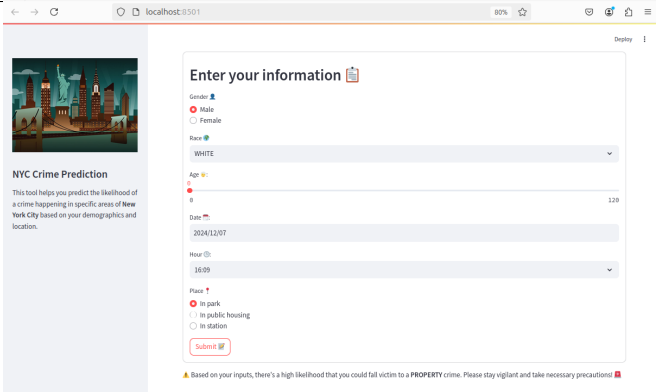
Latitude: 40.7084
Longitude: -73.9667

🔎 **Click on the map to select your location!** 📍 After selecting a location, fill in the form below to get a prediction about the crime risks in the area.

📍 **Selected Coordinates:**
Latitude: 40.708425704279605
Longitude: -73.96665573120119

Precinct: 90.0
Borough: Brooklyn 🌆

Fig. 7: Selected Position on map



NYC Crime Prediction

This tool helps you predict the likelihood of a crime happening in specific areas of **New York City** based on your demographics and location.

Enter your information 📋

Gender 👤
● Male
○ Female

Race 🌍
WHITE

Age 👶:
0
0 120

Date 📅:
2024/12/07

Hour 🕒:
16:09

Place 📍
● In park
○ In public housing
○ In station

⚠️ Based on your inputs, there's a high likelihood that you could fall victim to a **PROPERTY** crime. Please stay vigilant and take necessary precautions! 🚨

Fig. 8: Form with Prediction Result

VI. Conclusions

Predicting and preventing crime is a key focus in society. This study uses the Random Forest model to analyze and predict crime patterns, showing its effectiveness when trained optimally. The choice of model depends on the dataset's characteristics, highlighting the importance of tailoring approaches for optimal 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.