Crime Prediction in New York City Using Machine Learning

Angela Ben Frej, Tasnim Benhamed, and Mohamed Kaouech

***Abstract*— This study explores the use of machine learning for crime detection in New York City. By employing advanced algorithms and integrating them into a web application, the system predicts crime categories based on user inputs. The results demonstrate its potential for enhancing decision-making for individuals and law enforcement while addressing ethical and societal concerns.**

I. Introduction

Urban crime remains a significant challenge, and technological advancements in machine learning present opportunities for better detection and prevention. This project aims to develop a machine learning-based crime prediction system integrated into a web application. The study explores data preprocessing, model development, and evaluation, alongside ethical considerations.

II. Related Work

Several studies have demonstrated the effectiveness of machine learning in crime prediction:

- Naive Bayes and decision trees for classification, with Naive Bayes showing superior performance[1].
- SVM and ANN methods evaluated for dataset-specific performance[2].
- Clustering techniques to identify crime patterns, and classification for predictive tasks[3,4,5].

III. Data Overview and Preparation

A. Dataset Description

The dataset used is the NYPD Complaint Data Historic (2006–2019), consisting of over 6.9 million records across 35 features, including temporal, spatial, and descriptive crime data.

B. Data Cleaning

Key cleaning steps included:

- Handling missing values and standardizing formats for dates and times.
- Removing redundant columns and creating derived features, such as day of the week.
- Encoding categorical variables and normalizing numerical features.

C. Exploratory Data Analysis (EDA)

EDA identified patterns in the dataset, such as crime hotspots and temporal trends.

```
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  99.6057120446879  % missing values
HADEVELOPT have  95.5480283110380 5  % 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

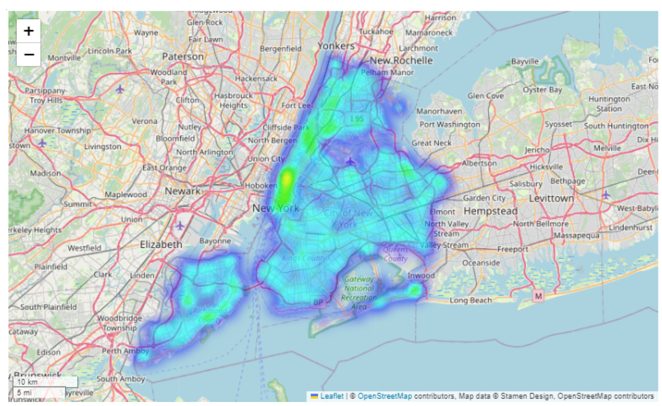

Fig. 2: Heatmap of NYC crime locations

IV. Methodology

A. Workflow Overview

The project follows a structured pipeline:

- Data collection, cleaning, and feature engineering.
- Model selection and hyperparameter tuning.
- Evaluation using performance metrics.

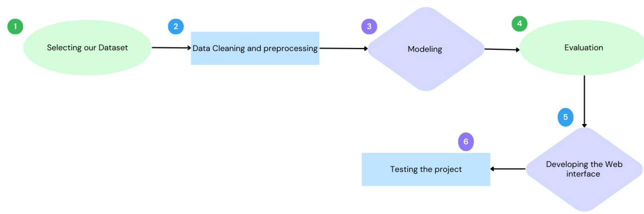

Fig. 3: Workflow of the crime prediction system

B. Modeling

The study evaluated three gradient boosting algorithms:

- **XGBoost:** Known for scalability and regularization capabilities.
- **LightGBM:** Efficient for large datasets with histogram-based learning.
- **CatBoost:** Designed for handling categorical data effectively.

C. Evaluation Metrics

Key metrics used include:

- ROC Curve and AUC for class discrimination.
- Confusion Matrix for precision, recall, and F1 Score.
- Overall accuracy of the models.

V. Results

A. Model Performance

TABLE I: Comparison of model performance

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

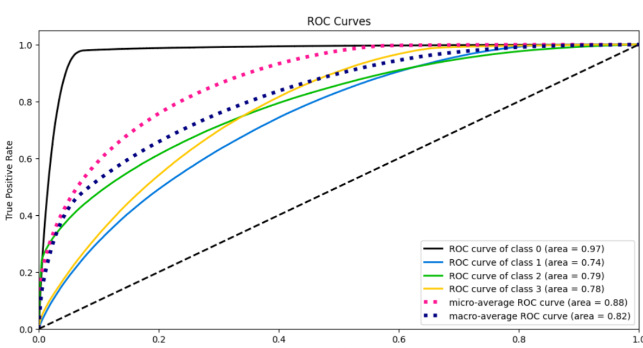

Fig. 4: ROC Curve for LightGBM

B. Discussion

LightGBM showed superior performance, particularly in distinguishing between crime categories. However, other models also performed reasonably well.

VI. Web Application

A. User Interface

The web app, built using Streamlit, allows users to input demographic data, select a location on a map, and receive crime predictions.

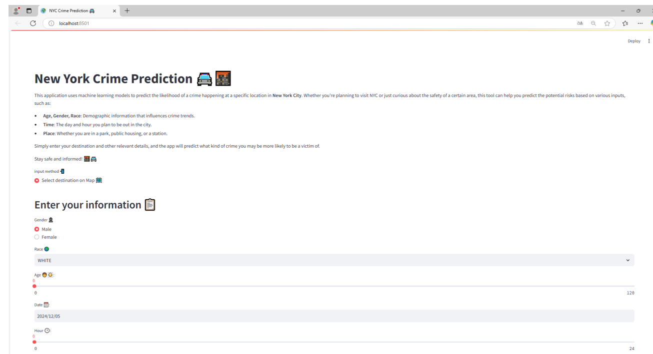

Fig. 5: Web application interface

B. Prediction and Mapping

Users receive crime type predictions and visualizations of crime hotspots based on input data the web interface

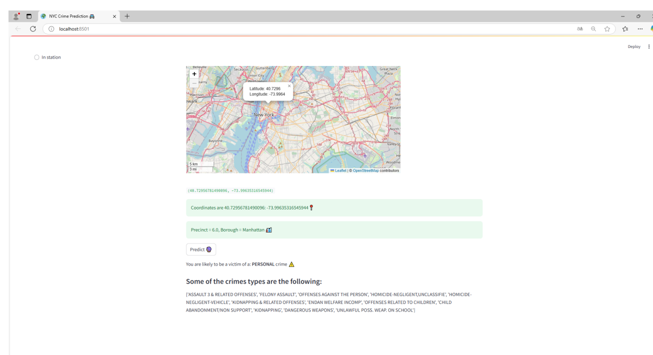

Fig. 6: Prediction when user chooses the destination

VII. Ethical Considerations

The project considers:

- Addressing biases in data and algorithms.
- Ensuring privacy and security of user data.
- Assessing societal impacts of crime prediction technologies.

VIII. Conclusion and Future Work

The study highlights the potential of machine learning in crime prediction. Future work includes real-time data integration, expansion to other cities, and improvements in user experience.

References

[1] Shiju Sathyadevan et al., "Crime Analysis and Prediction Using Data Mining," ICNSC, 2014.
[2] Sunil Yadav et al., "Crime Pattern Detection," ICECA, 2017.
[3] Amanpreet Singh et al., "Review of Supervised ML Algorithms," 2016.