New York City Crime Prediction Using Advanced Machine Learning Models

Nour KRICHEN   Zied KALLEL   Ines MEKKI   Dhia Elhak EZZEDDINI

***Abstract*—This paper presents an advanced machine learning framework for predicting crime categories in New York City using historical NYPD complaint data spanning from 2008 to 2024. Four supervised learning models—Decision Tree, Random Forest, LightGBM, and XGBoost—were trained and evaluated on more than 9.4 million crime records. Extensive data preprocessing, feature engineering, and hyperparameter optimization were applied to enhance predictive performance. Experimental results show that XGBoost achieved the best overall performance, reaching an accuracy of 77.21% and a weighted F1-score of 0.76. The selected model was deployed within an interactive Streamlit-based web application, allowing users to assess crime risk based on spatial, temporal, and demographic inputs. The proposed system provides actionable insights to support public safety awareness and data-driven decision-making.**

***Index Terms*—Crime Prediction, Machine Learning, XGBoost, LightGBM, NYPD Data, Streamlit, Public Safety**

I. Introduction

A. Motivation

Urban crime represents a persistent challenge for large metropolitan areas. New York City alone records millions of criminal complaints each year. Conventional crime prevention strategies are largely reactive, relying on historical analysis rather than predictive intelligence. With the increasing availability of large-scale open crime datasets and advances in machine learning, it has become possible to move toward proactive, data-driven crime prediction systems.

This research aims to:

- Develop accurate machine learning models for multi-class crime prediction
- Identify key temporal, spatial, and demographic factors influencing crime patterns
- Provide an accessible tool for public crime risk assessment
- Support data-driven resource allocation for law enforcement

B. Problem Statement

Given a specific location, time context, and demographic information in New York City, the objective is to predict the most likely category of crime. This is formulated as a multi-class classification problem with the following crime categories:

- **PERSONAL**: Assaults, homicide, weapons offenses
- **PROPERTY**: Burglary, larceny, robbery, theft
- **SEXUAL**: Sexual offenses and related crimes
- **DRUGS/ALCOHOL**: Drug-related and alcohol-related crimes

C. Contributions

The main contributions of this work are:

1) Advanced feature engineering incorporating temporal patterns and spatial density
2) A systematic comparison of four modern machine learning algorithms
3) Deployment of a real-time crime prediction web application
4) A fully reproducible end-to-end crime prediction pipeline

II. Literature Review

A. Crime Prediction Research

Early crime prediction studies focused on binary classification using traditional algorithms such as Naive Bayes and Decision Trees [1]. Later research explored more complex models, including Support Vector Machines [2] and ensemble techniques.

B. Ensemble Learning Approaches

Random Forests [5] and gradient boosting methods such as XGBoost [3] and LightGBM [4] have demonstrated strong performance in structured tabular data, making them well-suited for crime prediction tasks.

C. Research Gaps

Most existing works:

- Focus on binary crime prediction
- Use limited or outdated datasets
- Lack deployment-oriented solutions

This work addresses these limitations through multi-class modeling, large-scale data usage, and real-world deployment.

III. Methodology

A. Dataset Description

Source: NYPD Complaint Data Historic (2008–2024)
Records: 9,490,262 incidents
Features: 23 original attributes

B. Data Preprocessing

Preprocessing included:

- Missing value handling
- Date-time standardization
- Removal of redundant attributes
- Crime category consolidation

C. Feature Engineering

Temporal features such as weekend, night-time, rush hours, and seasonality were created. Spatial density features were computed using rounded latitude-longitude grids.

D. Feature Encoding and Scaling

Categorical variables were label-encoded, and numerical features were standardized using:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

E. Train-Test Split

- Training set: 80%
- Test set: 20%
- Stratified sampling with random state 42

F. Models Implemented

1) Decision Tree: Decision Trees are non-parametric supervised learning algorithms that recursively partition the feature space based on splitting criteria such as Gini impurity or entropy. Each internal node represents a decision based on a feature threshold, while leaf nodes represent class predictions. Decision Trees are interpretable and require minimal data preprocessing, making them suitable as baseline models. However, they are prone to overfitting and may exhibit high variance on unseen data. In this study, the Decision Tree serves as a baseline to evaluate the performance gains achieved by more sophisticated ensemble methods.

2) Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions for classification tasks. By introducing randomness through bootstrap sampling and feature subsampling, Random Forest reduces variance and improves generalization compared to individual decision trees. The algorithm is robust to overfitting, handles high-dimensional data effectively, and provides feature importance metrics. Random Forest was selected for its proven effectiveness in structured tabular datasets and its ability to capture complex non-linear relationships in crime data.

3) LightGBM: LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that uses tree-based learning algorithms optimized for efficiency and scalability. Unlike traditional gradient boosting methods that grow trees level-wise, LightGBM employs a leaf-wise growth strategy, which reduces training time and memory consumption while maintaining high accuracy. It incorporates advanced techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to handle large-scale datasets efficiently. LightGBM was chosen for its computational efficiency and its capability to process millions of crime records with minimal resource requirements.

4) XGBoost: XGBoost (Extreme Gradient Boosting) is an optimized distributed gradient boosting library designed for speed and performance. It implements a regularized boosting framework that prevents overfitting through L1 and L2 regularization terms. XGBoost builds an ensemble of weak learners sequentially, where each new tree corrects the errors of the previous ensemble. The algorithm supports parallel processing, handles missing values natively, and incorporates advanced features such as tree pruning and column subsampling. XGBoost has consistently demonstrated state-of-the-art performance in machine learning competitions and real-world applications, making it a strong candidate for multi-class crime prediction.

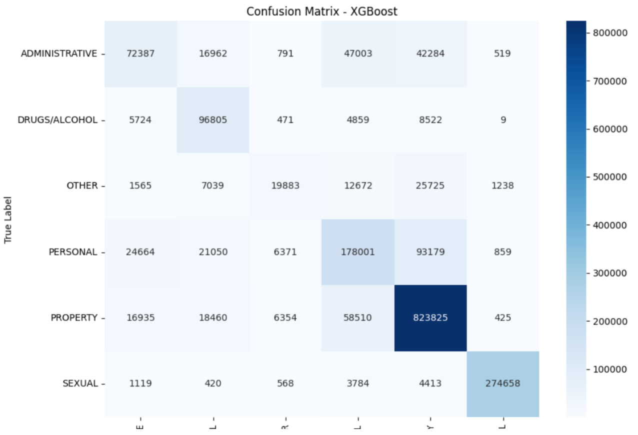


Fig. 1: Confusion Matrix for XGBoost Model

IV. Results

A. Model Performance Comparison

TABLE I: Model Performance Comparison

Model	Accuracy	F1-Score
Decision Tree	0.6824	0.7044
Random Forest	0.7177	0.7326
LightGBM	0.7058	0.7236
XGBoost	**0.7721**	**0.7628**

Key Findings:

- XGBoost achieved the highest overall accuracy and F1-score
- Random Forest showed competitive performance with higher memory cost
- LightGBM provided fast training with slightly lower accuracy
- Decision Tree served as a baseline model

B. Per-Class Performance

XGBoost demonstrated strong predictive capability for PROPERTY and SEXUAL crimes, while lower recall was observed for less frequent classes due to data imbalance. [Fig.1]

C. Feature Importance

Geographic location, hour of occurrence, precinct, and location density were identified as the most influential features.

Macro-average AUC of 0.85 indicates strong discriminative ability across all classes.

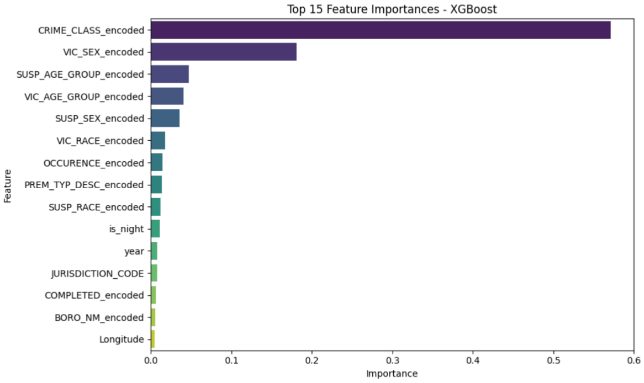


Fig. 2: Top Feature Importances (XGBoost)

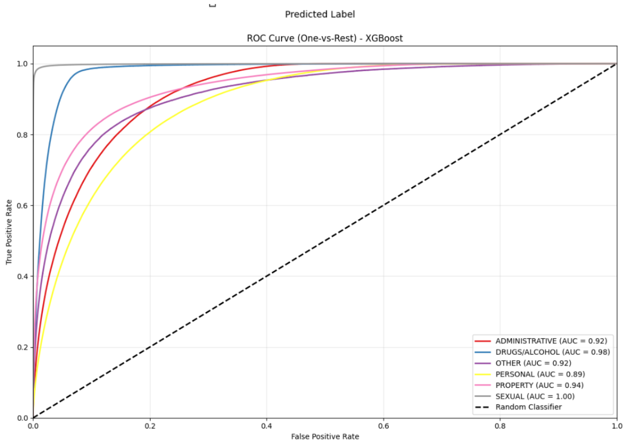
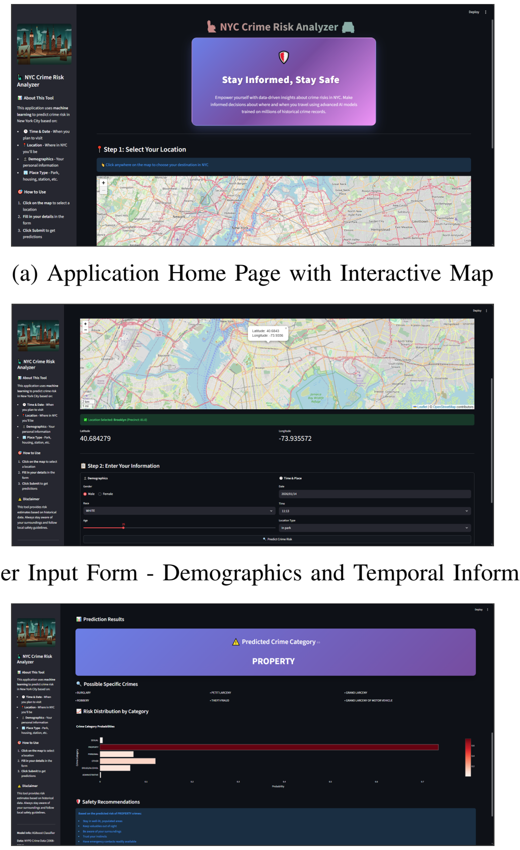


Fig. 3: ROC Curves for the XGBoost Model

V. Web Application Implementation

A Streamlit-based web application was developed to operationalize the trained model, allowing users to interactively assess crime risk using real-world inputs.



NYC Crime Risk Analyzer

Stay Informed, Stay Safe

Empower yourself with data-driven insights about crime risks in NYC. Make informed decisions about where and when you travel using advanced AI models trained on millions of historical crime records.

Step 1: Select Your Location

Click anywhere on the map to choose your destination in NYC

NYC Crime Risk Analyzer

About This Tool

This application uses **machine learning** to predict crime risk in New York City based on:

- **Time & Date** - When you plan to visit
- **Location** - Where in NYC you'll be
- **Demographics** - Your personal information
- **Place Type** - Park, housing, station, etc.

How to Use

1. **Click on the map** to select a location
2. **Fill in your details** in the form
3. **Click Submit** to get predictions

(a) Application Home Page with Interactive Map

Latitude: 40.6843
Longitude: -73.9356

Location Selected: Brooklyn (Precinct: 81.0)

Latitude
40.684279

Longitude
-73.935572

Step 2: Enter Your Information

Demographics

Gender
● Male ○ Female

Race
WHITE

Age
25

Time & Place

Date
2026/01/14

Time
11:13

Location Type
In park

Predict Crime Risk

Disclaimer

This tool provides risk estimates based on historical data. Always stay aware of your surroundings and follow local safety guidelines.

(b) User Input Form - Demographics and Temporal Information

Prediction Results

Predicted Crime Category

PROPERTY

Possible Specific Crimes

- BURGLARY
- ROBBERY
- PETIT LARCENY
- THEFT-FRAUD
- GRAND LARCENY
- GRAND LARCENY OF MOTOR VEHICLE

Risk Distribution by Category

Safety Recommendations

Based on the predicted risk of PROPERTY crimes:

- Stay in well-lit, populated areas
- Keep valuables out of sight
- Be aware of your surroundings
- Trust your instincts
- Have emergency contacts readily available

Model Info: XGBoost Classifier

Data: NYPD Crime Data (2008-

(c) Prediction Results Dashboard with Probability Chart

Fig. 4: Streamlit Web Application Screenshots

VI. Conclusion

This study demonstrates the effectiveness of advanced machine learning techniques for large-scale crime prediction. By combining extensive feature engineering, ensemble learning, and interactive deployment, the proposed system provides valuable insights for public safety applications. While promising, such systems must be used responsibly to support—not replace—human decision-making.

References

[1] S. Sathyadevan et al., "Crime Analysis and Prediction Using Data Mining," ICNSC, 2014.

[2] S. Yadav et al., "Crime Pattern Detection, Analysis and Prediction," ICECA, 2017.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," KDD, 2016.

[4] G. Ke et al., "LightGBM," NeurIPS, 2017.

[5] L. Breiman, "Random Forests," Machine Learning, 2001.

[6] NYC Open Data, "NYPD Complaint Data Historic."