Integrated Machine Learning for Crime Prediction: A New York City Case Study with Interactive Visualization

Ahmed Guermazi - Emna Belguith - Fatma Abid - Rimel Hammami

***Abstract*—Crime prediction is a pivotal domain in urban planning and public safety. This paper utilizes large-scale historical crime data from New York City to build predictive models for crime categorization. By leveraging extensive data cleaning, geospatial feature engineering, and class balancing techniques, we evaluate three gradient boosting classifiers: LightGBM, XGBoost, and CatBoost. Furthermore, we demonstrate the practical utility of these models by deploying them via an interactive, geospatial web application. Our results indicate that while behavioral unpredictability limits accuracy in certain categories, the models perform robustly in distinguishing property and drug-related offenses.**

***Index Terms*—Crime Prediction, Machine Learning, Gradient Boosting, NYC Open Data.**

I. Introduction

Crime prediction is an increasingly important topic in the intersection of data science, urban planning, and public policy. With the proliferation of open data initiatives in major metropolitan areas, researchers now have unprecedented access to detailed incident-level crime data. New York City, in particular, provides rich historical data regarding reported offenses, demographic context, and geographic coordinates — making it an ideal case study for machine learning applications.

The objective of this paper is to explore how large-scale crime data can be cleaned, engineered, and modeled to produce reliable predictions of crime categories. Moreover, we seek to demonstrate how these predictive models can be deployed in an accessible, interactive interface for public use. To achieve this, we evaluate multiple gradient boosting classifiers, perform extensive data preprocessing, and contextualize our findings within recent advances in crime prediction research.

II. Literature Review

A. Crime Prediction with Machine Learning

Machine learning has been widely used to model crime occurrences. Early work focused on statistical methods such as regression and time-series analysis [9]. More recent studies emphasize supervised learning approaches for classification and forecasting [5], [7]. Gradient boosting algorithms, including XGBoost and LightGBM, have shown strong performance in high-dimensional, imbalanced datasets typical of crime records [1].

B. Spatial Analytics in Crime Modeling

Geospatial patterns are central to understanding crime. Studies emphasize the use of GIS and heatmap visualizations to identify crime hotspots [2]. Predictive models incorporating spatial coordinates often include clustering and spatial regression techniques to account for adjacent risk areas [6]. Our work incorporates latitude, longitude, precinct boundaries, and borough partitions to spatially inform model training.

C. Handling Class Imbalance

Crime data typically exhibit strong class imbalance — rare but severe crimes coexist with more frequent but lower-risk offenses. Techniques such as oversampling (SMOTE), class weighting, and ensemble methods are commonly used to mitigate imbalanced class learning [3], [4]. In this study, we balanced each offense category to ensure equitable learning.

D. Deployment and Public-Facing Interfaces

Translating predictive models into practical tools is a growing trend. Interactive dashboards have been used to support decision-making in health and transportation [8]. Tools like Folium and Streamlit facilitate rich, geospatially enabled interfaces, lowering the barrier of entry for non-expert audiences.

III. Data Preparation

A. Data Quality Assessment

The raw dataset consists of approximately 7.8 million records from NYPD Complaint Data Current. An initial evaluation revealed substantial missingness. Fields such as `HOUSING_PSA` ($\approx$ 92% missing) and `TRANSIT_DISTRICT` ($\approx$ 97% missing) were almost entirely empty. Demographic attributes also displayed incomplete reporting; categories such as "UNKNOWN" in victim race and suspect age reflect reporting gaps.

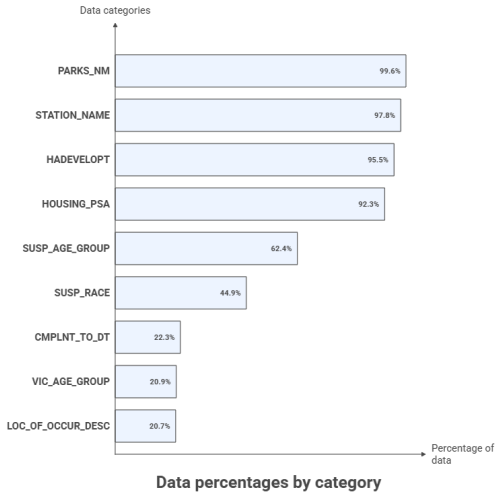

Fig. 1: Missing data percentages

B. Data Cleaning

Based on quality assessment, several steps were performed:

- **Removal of Sparse Columns:** Columns with extreme sparsity were removed due to the impossibility of meaningful imputation.
- **Redundancy Removal:** `X_COORD_CD` and `Y_COORD_CD` were dropped in favor of Latitude/Longitude.
- **Imputation:** Missing demographic fields (VIC_RACE, VIC_AGE_GROUP) were recoded as "UNKNOWN".
- **Geospatial Filtering:** Records with coordinates outside NYC boundaries were filtered out.

After cleaning, the dataset contained 7,807,331 valid observations.

C. Feature Engineering

Several transformations were applied to enhance modeling quality:

1) *Binary Contextual Indicators:* Sparse textual fields were converted to binary indicators (e.g., `IN_PARK`, `IN_STATION`).
2) *Temporal Decomposition:* Datetime fields were converted into year, month, day, hour, and weekday.

D. Exploratory Analysis and Target Engineering

Although distinct from preprocessing, exploratory insights guided our data strategy. Preliminary distributions revealed that fields like `HOUSING_PSA` were 92% empty, justifying their removal.

Crucially, the target variable `OFNS_DESC` exhibited extreme class imbalance, dominated by *Petit Larceny* ($\approx$ 1.3M records) and *Harassment* ($\approx$ 1M). Raw classification on these labels would bias the model toward frequent, low-severity crimes.

Fig. 2: Distribution of crime types over the dataset

To mitigate this, we consolidated the offenses into six thematic categories. This mapping, detailed in Table I, ensures a more stable and interpretable classification framework.

TABLE I: Consolidation of Crime Categories

Final Category	Representative Original Offenses
PROPERTY	Petit Larceny, Burglary, Robbery, Forgery, Arson, Fraud.
PERSONAL	Assault 3, Felony Assault, Dangerous Weapons, Homicide, Kidnapping.
SEXUAL	Rape, Sex Crimes, Harassment 2.
DRUGS/ALC.	Dangerous Drugs, Intoxicated & Impaired Driving.
ADMIN	Criminal Trespass, Traffic Laws, Public Administration.
OTHER	Miscellaneous Penal Law, Gambling, etc.

These categorizations allow the model to learn distinct behavioral patterns rather than overfitting to specific legal codes. So the dataset distribution became :

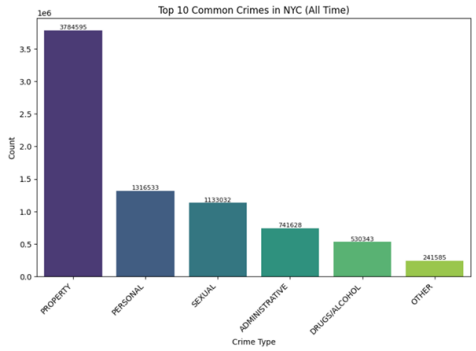

Fig. 3: onsolidation of the 6 Crime Categories

IV. Methodology

A. Target Encoding and Balancing

The target variable was transformed into numerical labels using Label Encoding. Because the original dataset exhibited

substantial class imbalance—dominated by non-personal offenses—an oversampling strategy was applied. The resulting balanced dataset contains 530,343 instances per class.

B. Feature Selection

A set of relevant predictors was selected based on domain knowledge:

- **Temporal:** year, month, day, hour.
- **Geospatial:** latitude, longitude, precinct code.
- **Contextual:** IN_PARK, IN_PUBLIC_HOUSING.
- **Demographic:** age group, race, sex.

C. Train–Test Split

The dataset was partitioned into training (85%) and testing (15%) subsets using a fixed random seed (`random_state=42`) to ensure reproducibility.

V. Model Development

Three gradient boosting algorithms were evaluated: LightGBM, XGBoost, and CatBoost. All models were optimized using Optuna.

A. LightGBM

LightGBM was selected for its efficiency with large datasets.

- **Config:** num_leaves=189, learning_rate=0.071.
- **Result:** Accuracy = 0.634. Strong recall for Drugs and Property crimes, but lower for Personal crimes.

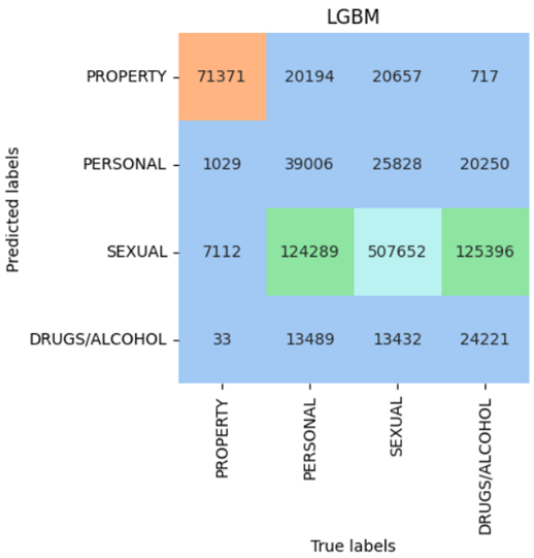

Fig. 4: LigtGBM confusion matrix

B. XGBoost

Optimized with a GPU-accelerated histogram tree algorithm.

- **Config:** max_depth=5, n_estimators=60.
- **Result:** Accuracy = 0.633. Results were nearly identical to LightGBM.

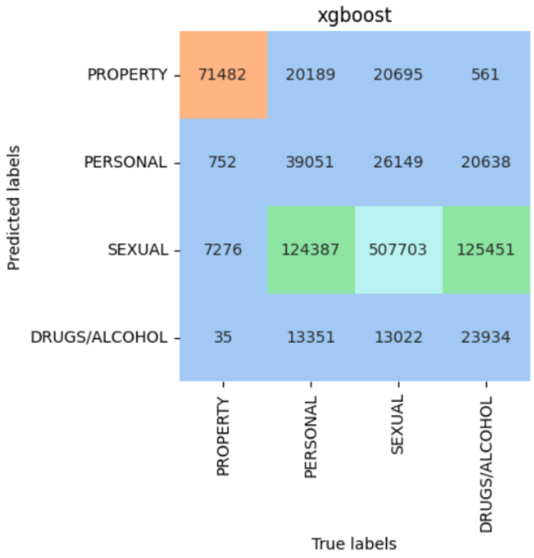

Fig. 5: XGBoost confusion matrix

C. CatBoost

CatBoost inherently handles categorical variables well.

- **Config:** iterations=1200, depth=7.
- **Result:** Accuracy = 0.634. Best generalization for Property crimes.

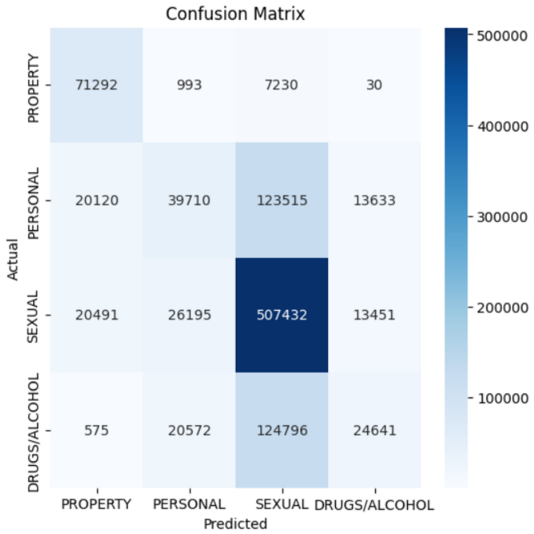

Fig. 6: CatBoost confusion matrix

TABLE II: Model Performance Comparison

Model	Accuracy	Strength	Weakness
LightGBM	0.634	Efficiency	Minority Recall
XGBoost	0.633	Stability	Tuning Sensitivity
CatBoost	0.634	Categorical Feats	Training Time

VI. Deployment

To bridge the gap between complex machine learning models and public utility, we developed *NYC Crime Prediction*, an interactive web application. The system is built using the Streamlit framework, chosen for its rapid prototyping capabilities and seamless integration with Python's data science stack.

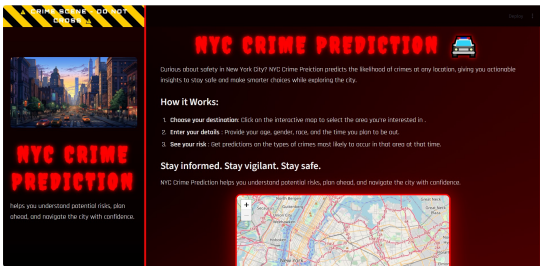

Fig. 7: Streamlit interface

A. System Architecture

The application architecture relies on three core components:

1) **Geospatial Engine:** A Leaflet-based interactive map (via `folium`) allows users to pinpoint precise locations within New York City. The application automatically resolves the latitude and longitude of the user's click to identify the associated NYPD Precinct and Borough using spatial polygon mapping.
2) **Contextual Input Module:** A dynamic form collects specific risk factors utilized by the model, including demographic data (Age, Race, Gender) and environmental context (e.g., proximity to stations, parks, or public housing), as shown in Fig. **??**.
3) **Inference Backend:** The inputs are preprocessed in real-time to match the training schema (Label Encoding and feature scaling) before being passed to the trained Gradient Boosting model for classification.

B. User Experience (UX) Design

The interface features a high-contrast, dark-mode design with red accents, chosen to simulate an "alert" or "warning" aesthetic appropriate for crime risk assessment. The user flow consists of three intuitive steps:

1) Location Selection: The user lands on a dashboard displaying a heatmap of NYC. By clicking on the map (Fig. 7), the system captures the spatial coordinates. This eliminates the need for users to know their precinct number manually.

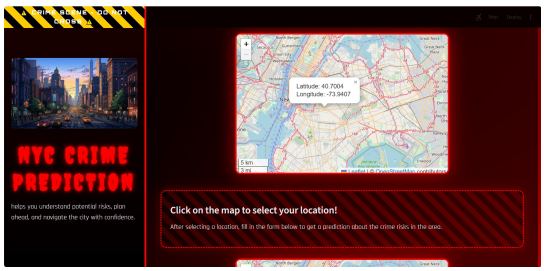

Fig. 8: Interactive Map Selection

2) Profile Configuration: Once a location is selected, a detailed input form appears (Fig.8). The user inputs temporal data (Date, Hour) and personal characteristics. Notably, the interface asks for specific situational contexts—such as "In station" or "In park"—which are high-weight features in our model.

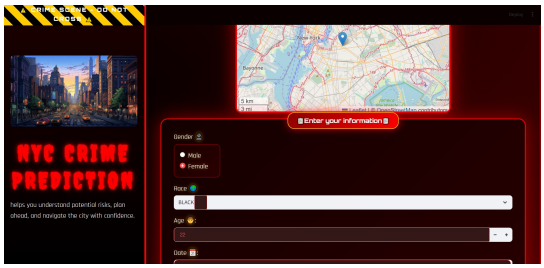

Fig. 9: Demographic & Context Input

3) Risk Prediction and Warning: Upon submission, the model predicts the most probable crime category. The result is displayed not as a raw numerical code, but as a natural language warning (Fig. 9), e.g., *"You may face a higher risk of a SEXUAL crime. Stay aware..."*. This actionable insight empowers users to take necessary precautions based on data-driven forecasts.

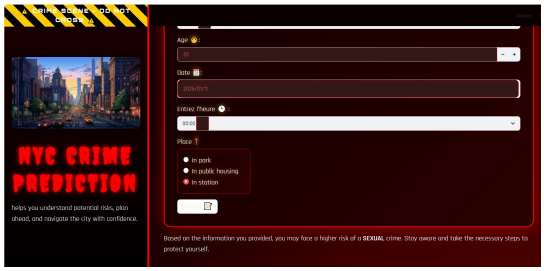

Fig. 10: Prediction & Safety Warning

C. Practical Significance

This deployment demonstrates that historical crime data can be effectively operationalized. By requiring inputs like "Race" and "Age," the tool highlights how victimization risks vary across different demographics, providing a personalized risk assessment rather than a generic regional average.

VII. Conclusion

This study demonstrated that while crime data is inherently noisy, advanced gradient boosting methods can achieve robust

baseline predictions ($\approx 63\%$ accuracy) when coupled with rigorous spatial engineering. The deployment of these models via a Streamlit interface illustrates the potential for accessible, data-driven public safety tools.

REFERENCES

[1] R. Berk, B. Kriegler, and J. Baek, *Machine learning methods for crime prediction and justice analytics*. Springer, 2019.

[2] S. Chainey and J. Ratcliffe, *GIS and crime mapping*. Wiley, 2005.

[3] N. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[4] A. Fernández et al., *Learning from Imbalanced Data Sets*. Springer, 2018.

[5] M. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems*, vol. 61, pp. 115–125, 2014.

[6] G. Mohler et al., "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 100–108, 2011.

[7] X. Wang and R. Green, "Crime prediction and classification using big data," *ACM Trans. Knowl. Discov. Data*, 2019.

[8] Q. Wang, R. Zhang, and Y. Fu, "A web-based crime forecasting tool using random forest," *Int. J. Geogr. Inf. Sci.*, vol. 32, no. 1, pp. 123–142, 2018.

[9] W. Wang, "A comparative analysis of statistical models for crime forecasting," *Journal of Urban Analytics*, vol. 4, no. 2, pp. 34–48, 2012.