New York City Crimes Detection using Machine Learning

Skander Yaakoubi, Aymen Ktari, Youssef Makhlouf
SUP'COM

***Abstract*—This study explores the use of machine learning to predict criminal activities in New York City. The project involves the development of a user-friendly web application that enables individuals to input personal data and select specific locations within the city. By leveraging advanced machine learning algorithms, the system forecasts potential criminal incidents in the chosen areas. The paper details the methodology, model selection, and implementation of the web application, while also addressing ethical considerations and societal impacts. The results highlight the effectiveness of the approach in improving crime awareness and aiding decision-making for both users and law enforcement agencies in New York.**

I. Introduction

With the increasing challenges associated with urban security, leveraging technological advancements becomes imperative for enhancing crime detection and public safety. This research aims to address this need through the application of machine learning techniques in the context of New York City. The project not only focuses on developing a robust crime prediction model but also integrates this capability into a user-friendly web application.

II. Literature Review

Several algorithms for predicting crime have been proposed, with prediction accuracy contingent on the type of data employed and the attributes selected for prediction. Previous studies explored crime prediction and classification using data gathered from diverse sources, utilizing algorithms such as Naive Bayes and Decision Trees, with Naive Bayes often demonstrating superior performance.

III. Methodology

The dataset used in this study is the NYPD Complaint Data Historic, which includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to 2019.

A. Data Collection

The dataset contains over 6.9 million complaints and 35 columns, providing spatial and temporal details of crime occurrences. This rich dataset serves as the foundation for analysis and modeling.

B. Data Cleaning and Preprocessing

- Irrelevant columns, such as `X_COORD_CD` and `Y_COORD_CD`, were removed.
- Temporal features like `MONTH` and `DAY_OF_WEEK` were extracted and encoded using cyclical transformations.
- Categorical variables, such as `VIC_RACE` and `VIC_SEX`, were one-hot encoded.
- The target variable `LAW_CAT_CD` was label-encoded for classification tasks.
- A preprocessing pipeline was implemented to standardize numerical features and ensure consistency.

C. Modeling

Various machine learning algorithms were employed, including Logistic Regression, Random Forest, LightGBM, XGBoost, and Support Vector Machine (SVM). Hyperparameter tuning using Optuna enhanced model performance.

- Dataset split: 85% training, 15% testing.
- Preprocessing pipeline integrated StandardScaler for numerical features and OneHotEncoder for categorical variables.
- Evaluation metrics: accuracy, precision, recall, and F1-score.

Among the models, LightGBM showed slightly superior performance, followed by Random Forest and XGBoost.

IV. Implementation

A. Data Collection and Cleaning

Fig. 3 illustrates the distribution of crime levels. Exploratory data analysis highlighted trends in crime occurrence, including seasonal and demographic factors. Missing values were addressed through imputation or column removal.

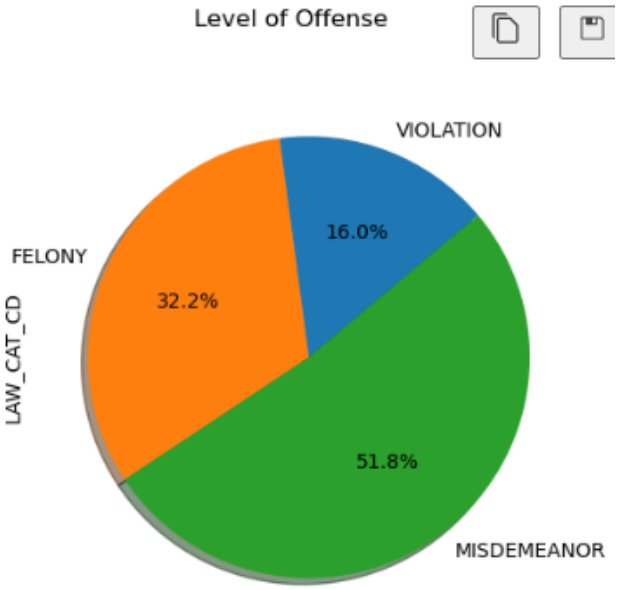

Fig. 1: Level of offense distribution.

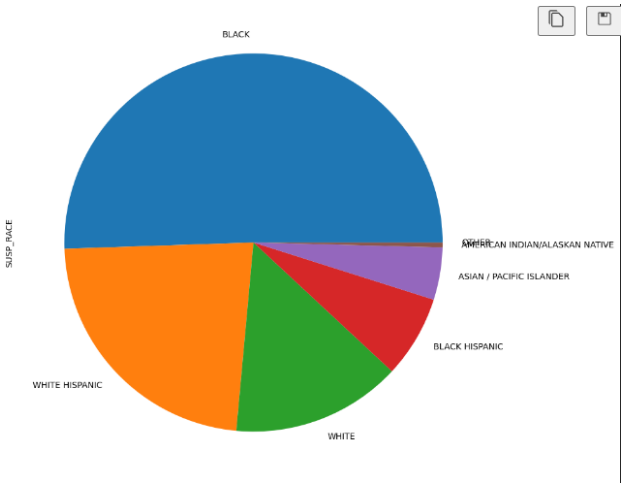

Fig. 2: suspect$_r aces$.

B. Model Evaluation

- **LightGBM:** Exhibited the highest accuracy and F1-score. *AUC-ROC and confusion matrices provided additional performance insights.*
- **XGBoost:** Performed comparably, with slight differences in precision and recall.
- **CatBoost:** Demonstrated balanced results across metrics.

Fig. **??** presents ROC curves for the models, while Table I compares their performance.

TABLE I: Comparison of Different Models

Model	Accuracy (%)	F1 Score
XGBoost	91.2	90.64
CatBoost	93.38	91.29
LightGBM	94.6	65.31

V. User Interface

A web application was developed using Streamlit and Folium, enabling users to:

- Input demographic details (e.g., gender, race, age).
- Specify date and time.
- Select locations interactively via a map interface.

The application integrates the trained model to provide crime prediction results in real-time.

Fig. 3: interface

VI. Conclusion

The proposed system demonstrates the potential of machine learning in enhancing urban security. By accurately predicting crime and offering user-friendly visualization tools, it bridges the gap between data-driven insights and practical decision-making for law enforcement and the public.

References

@articlekadar2018mining, title=Mining large-scale human mobility data for long-term crime prediction, author=Kadar, Cristina and Pletikosa, Irena, journal=arXiv preprint arXiv:1806.01400, year=2018

@articleqi2024eyes, title=Eyes on the Streets: Leveraging Street-Level Imaging to Model Urban Crime Dynamics, author=Qi, Zhixuan and Luo, Huaiying and Chi, Chen, journal=arXiv preprint arXiv:2404.10147, year=2024

@articleqi2024eyes, title=Eyes on the Streets: Leveraging Street-Level Imaging to Model Urban Crime Dynamics, author=Qi, Zhixuan and Luo, Huaiying and Chi, Chen, journal=arXiv preprint arXiv:2404.10147, year=2024

@articleli2014unsupervised, title=An unsupervised learning algorithm for the classification of the protection device in the fault diagnosis system, author=Li, Bin and Guo, Yajuan and Wu, Yi and Chen, Jinming and Yuan, Yubo and Zhang, Xiaoyi, journal=China International Conference on Electricity Distribution (CICED), year=2014