# End-to-End New York City Crimes Detection using Machine Learning

Ghazi Tounsi, Mohamed Karaa, and Riadh Tebourbi

Higher School of Communication of Tunis

{*ghazi.tounsi, mohamed.karaa, riadh.tebourbi*}*@supcom.tn*

## Abstract

*Crimes are a precarious and common social problem faced worldwide. Crimes affect the quality of life, economic growth, and reputation of a nation. In the last few years, the crime rate has been enormously increased. As a response to that increase, there is a need for advanced systems and new approaches for improving crime analytics for protecting communities. While accurate real-time crime prediction helps reduce crime rates, it remains a difficult problem for the scientific community because crime occurrences depend on many complex factors. In this work, various visualization techniques and machine learning algorithms are adopted to predict the crime distribution over New York City.*

*In the first step, a raw dataset was processed, and multiple visualization techniques were adopted to understand better the data and the relations between the different variables.*
*Afterward, multiple machine learning algorithms were used to predict crime types based on user input and consider the users' geographic location.*
*The last step is to develop a user interface using Streamlit to make user interaction more straightforward. The final code is available at: https://github.com/mohamedkaraa/New-York-Crime-Prediction*

*Keywords: Crime Analysis; Crime prediction; Data Visualization; Crime Maps; Machine Learning, Folium, Streamlit*

## 1. Introduction

### 1.1. General overview

Crimes are common social problems that affect the quality of life, economic growth, and reputation of countries. Crimes are significant factors that affect various vital decisions of an individual's life, like moving to a new place, roaming at the right time, avoiding risky areas, etc.
Crimes affect and defame the image of a community. Crimes also affect a nation's economy by placing the financial burden on the government due to the need for additional police forces, courts, etc.
As crimes are growing drastically, we are alarmed to reduce them at an even faster rate. Overall, New York City index crime increased by 1.3% in 2021 compared to 2020. Only Burglary saw a 13.7% decrease compared to 2020, but Robbery increased by 15.8%, and Felonious Assault 13.8% [1]. We can reduce these numbers if we can analyze and predict crime scenes and locations and take preventive measures in advance.
The crime rates can be significantly reduced by real-time crime forecasting and mass surveillance, which help save the most valuable lives. Proper analysis of previous crime data helps predict the crimes and thus supports reducing the crime rate. The analysis process involves investigating crime reports and identifying new patterns, series, and trends as quickly as possible.
This analysis helps in preparing statistics, queries, and maps on demand. Crimes type can be predicted as the criminals are active and operate in their comfort zones, and they are likely to reproduce the same crime if they complete the first crime successfully. Criminals generally find a similar location and time for attempting the next crime.
Although it may not be accurate for all the cases, the possibility of repetitions is high, as per studies, making the crimes predictable.

This paper proposes a web application & visual-based crime prediction tool interface built with python using various libraries such as Streamlit for the UI, Pandas for data processing, Folium for geographical data visualization, etc. The proposed framework uses different visualization techniques to show the trend of crimes and various ways to pre-

dict the crimes using machine learning algorithms.
Most importantly, the steps followed are data pre-processing, Data Visualization, and Model building, which are discussed more in detail in the following sections. In brief, the pre-processing phase consists of cleaning and transforming data. The visualization phase generates various reports and maps for the diagnosis and analysis process, and finally, in the model building phase, different machine learning algorithms are used for the classification of crime that can happen in a particular location.

### 1.2. Related work

Analyzing and predicting crime is an important activity that can be optimized using various techniques and processes. A lot of research work is done by multiple researchers in this domain. The current work is limited to using the datasets to identify crime locations. But none of them considered the type of crime and the date of crime as the factor. ToppiReddy et al. [2], provide the static maps with no interactive features. The same with Almuhanna et al. [3], who have proposed only machine learning and data processing methods without interactive maps.

The proposed framework provides visualization techniques that consider the location and many other information introduced by the user to predict the type of crime to overcome these limitations. Few papers focused on the usage of decision trees for crime prediction [4] [5]. Nasridinov et al. [5], used the attributes population of a country, Median Household income, percentage of people who are unemployed with age greater than 16, type of crime, etc. which only predicts whether in an area there will be a high, medium or low percentage of violent crimes that can happen in future. The methods proposed by them did not predict the type of crime that could occur.

## 2. Dataset

This work relies on NYPD Complaint Data Historic dataset [6]. This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to 2019. The dataset contains 6901167 complaint and 35 columns including spatial and temporal information about crime occurrences along with their description and penal classification.

### 2.1. Exploratory data analysis

To further understand the data in hand and analyze the different distributions and relations between features, we undergo an Exploratory data analysis (EDA) in order to answer questions about what, where and when crimes occur.

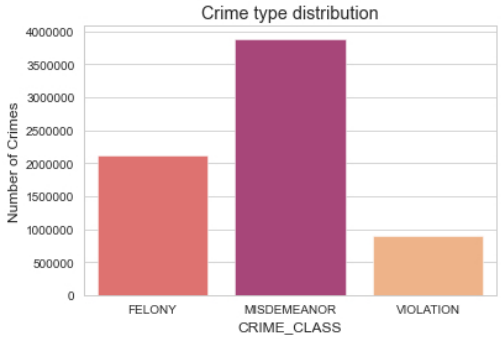



Figure 1: Crime count per class.

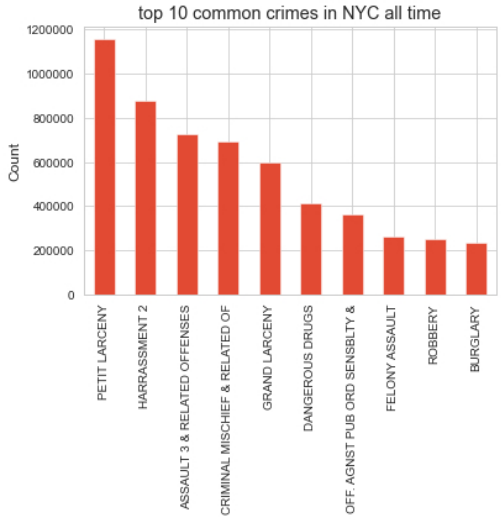



Figure 2: Top 10 crime types reported to NYPD.

#### 2.1.1 Crime classification

Crime types are classified according to penal code to 3 types: felonies, misdemeanors and violations, ordered from the most to the less serious. Figure 1**??** shows that the most common crimes are misdemeanors. It is also shown that the dataset is unbalanced, to heal this problem in model training, we take a subset of 800000 complaint for each class.

It is also important to know which specific crimes are more common (Figure 2**??**) In addition to that, we analyze the profile of the victims (Figure 33) to know which categories are more susceptible of being victims of crime. We can see that females are more affected by crimes. Also, people aging between 25 to 44 years old are most likely to be victims more than any other age group. These features can be helpful to predict the probability of a crime once combined with other spatio-temporal features.

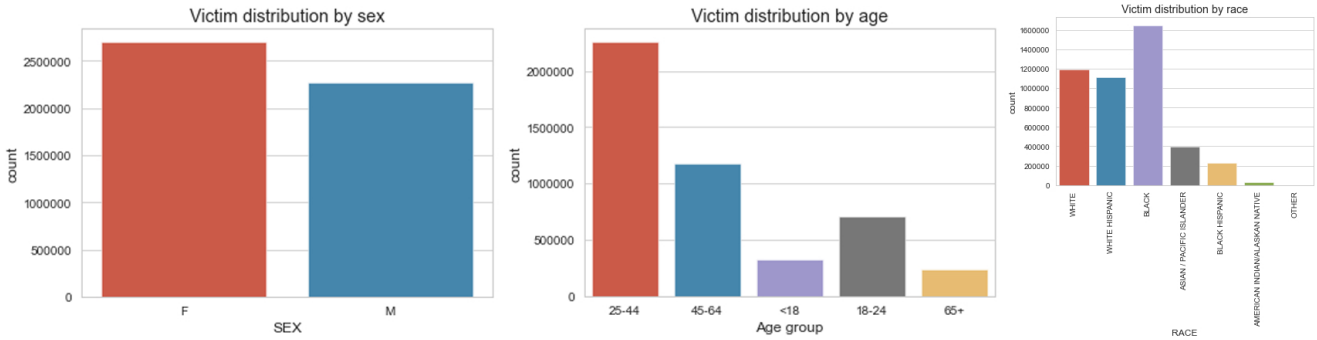



Figure 3: Crime victims distribution by sex, age and race.

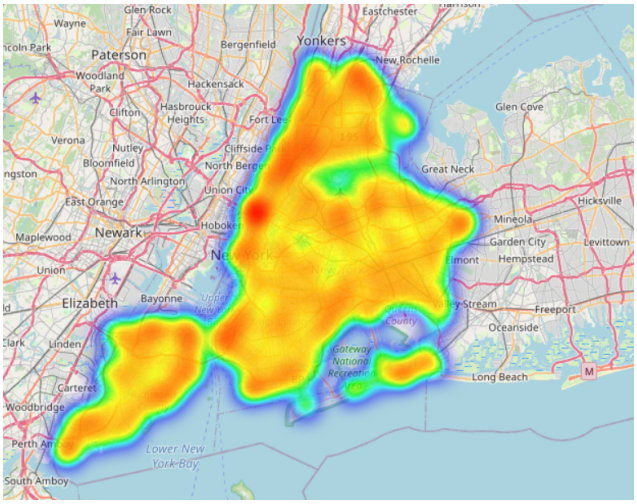


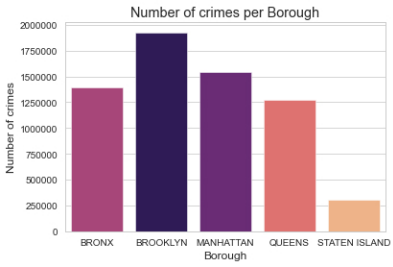

Figure 4: Crime heat map for the 2nd of February, 2018.

### 2.1.2 Geographic distribution

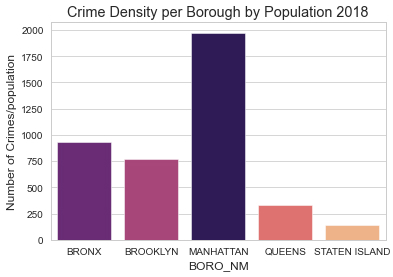
To study the geographic partition of crimes in New York City and their patterns, the dataset provides geographic coordinates (latitude and longitude) of the crime scene. We use these coordinates to plot a heat map to discover the hot spots. Figure 44 illustrates an example of such heat maps for the day 02/04/2018. The result shown goes well with the distribution of crimes along the five boroughs of New York (Figure 55). It shows high crime levels in Manhattan, Brooklyn and the Bronx.

### 2.1.3 Time distribution

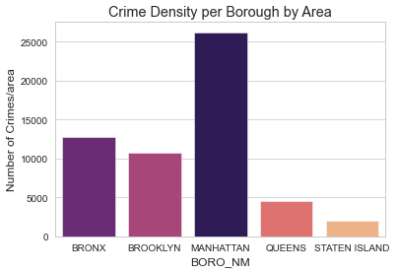
We study the temporal distribution of crime occurrence across month, days and day hours. Analysis proves that most crimes occur in the summer season month, on work days, and in the evening hours of the day as shown in Figure 66.




(a) Crimes by borough.




(b) Crimes by borough by population.

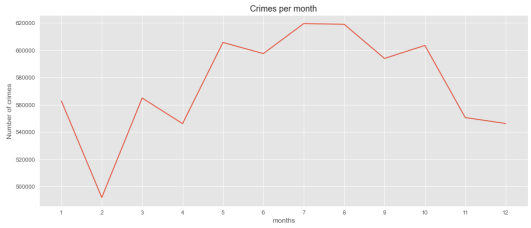



(c) Crimes by borough by area.

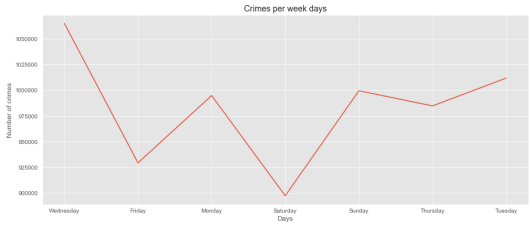
Figure 5: Crime distribution by borough.

## 2.2. Data cleaning

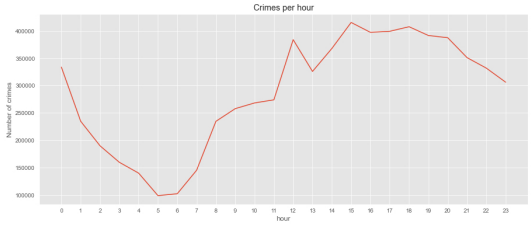
To prepare the dataset for modeling phase, we went through a data cleaning process. First, we manually deleted redundant features such as crime codes, report dates and




(a) Crimes distribution by month.




(b) Crimes distribution by week days.




(c) Crimes distribution by hours.

Figure 6: Crime temporal distribution.

geographic coordinates in other systems. Then, time features were transformed to extract years, months, days and hours each on its own. Next, we dealt with missing data either by deleting rows in case only little values are missing for the column, or by considering values as a booleans and the missing ones are False (park, housing, station features). The last step is to keep only valid values across columns, discarding irrelevant values, for example in age group feature.
These procedures lead to having a clean dataset having 6882147 complaint and 23 columns.

## 3. Modeling

After the data pre-processing, in order to classify three different types of crimes, based on their severity, several machine learning algorithms were applied in order to compare their results including, Random Forest (RF) [7], LightGBM [8], and XGBoost [9] classifiers. The classification is mainly used to recognize the labelled classes by knowing their attributes (features) in the dataset, thus predicting the class label for instances with known features. Hence, using the classifiers in crime prediction constructs a future-oriented model to identify the criminal type within a specific time. In this section, a description of all the used classification is illustrated in Table. 1

| Classifier | Description |
|---|---|
| Random Forest | RF [7] is a well-known machine learning algorithm; it has several decision trees since it is a tree-structured classifier. Then, each tree will give a vote or decision to assign the most proper class label for each input. In addition, it does not suffer from over-fitting. |
| LightGBM | LightGBM [8], is one of the most popular algorithms that are based on Gradient Boosted Machines. It offers a significant difference in the execution time for the training procedure as it is almost 7 times faster than XGBoost and is a much better approach when dealing with large datasets. |
| XGBoost | XGBoost [9] has been a very important machine learning approach in many areas. It is also a scalable Tree Boosting System, and has been reported to be very efficient in many research projects and other applications. |

Table 1: Classification methods used in this study, and their short description

## 4. Experimental Result

The classification task was done using three classifiers. For classification tasks, the confusion matrix and the ROC curve were an appropriate metrics to measure model performance.

1. **Confusion Matrix:** It is a performance measurement for machine learning classification problem where output can be two or more classes (3 classes in our case). It contains four importante measures:

   - *True Positive (TP), True Negative (TN):* Represents the number of the prediction, correctly predicts as a given class.
   - *False Positive (FP), False Negative (FN):* Represents the number of the prediction, falsely predicts as a given class.

2. **ROC curve (receiver operating characteristic curve):** Is a graph showing the performance of a classification model at all classification thresholds. This curve plots the True Positive Rate and the False Positive Rate.

| Classifier | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| Random Forest | 0.52 | 0.51 | 0.49 | 0.52 |
| LightGBM | 0.59 | 0.58 | 0.57 | 0.59 |
| XGBoost | 0.62 | 0.60 | 0.59 | **0.60** |

Table 2: Results of the classifiers accuracy

Hence, the f1 score was measured by calculating precision and recall values. Furthermore, the comparison was done among the three models, where accuracy was also measured for each of the models. Formulas for the confusion matrix and the ROC curve are shown below:

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

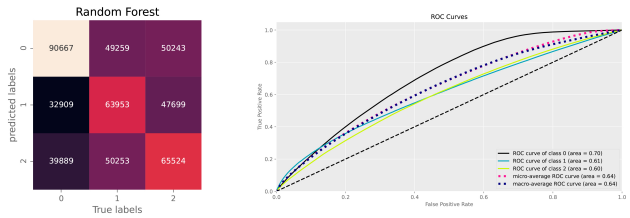
$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

$$Accuracy = \frac{TF + TN}{TP + FN + TN + FP} \tag{7}$$

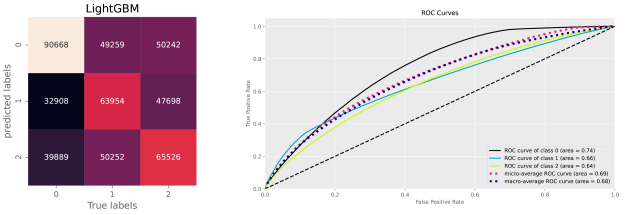
Hence, as shown in Table. 2 for the classifier scores using the confusion matrix. However, LightGBM and XGBoost models tend to have very close validation scores which are respectively 0.59 and 0.6, whereas is relatively low for the Random Forest model with accuracy scores of 0.52. However, XGBoost model tend to be the best classification model for correctly predicting the crime classes with accuracy scores of 0.6040. Furthermore, from the confusion matrix heatmaps we have a multi-class classification task with 3 classes from 0 to 2 where it represents class labels that replaced actual class names. From what we can see from the confusion matrix and the ROC curve plots, We can find that XGBoost model outperforms the other classifiers with the highest predictions in each class. Yet, they all tend to have a low numbers of true classifications for each class.

## 5. User Interface

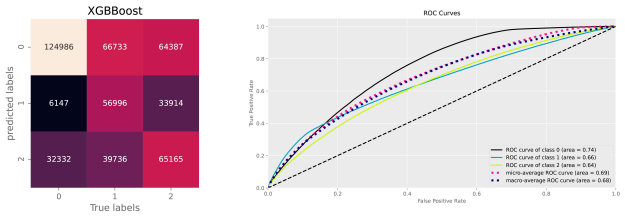
After training the model and saving the weights file, we built a web application (Figure 8) using Streamlit and Folium to allow the user to interact with the map and predict the type of crime that could happen. The user can enter




(a) Random Forest metrics.




(b) LightGBM metrics.




(c) XGBoost metrics.

Figure 7: Classification models metrics.

his gender, race, age, the date and hour in which he wants to predict the type of crime, the location on the map and finally, the place (In a park, In public housing or a station). This information is then transformed to fit the model input, and then, using the loaded model weights file, we predict the type of crime and send it back to the user along with the potential subtypes of that crime.

## 6. Conclusion

Crimes have been a significant problem in many cities. Hence, many researchers tried to solve it and predict the most criminal hot spots to increase the understanding of dangerous places at certain times. In this paper, we have analyzed the data of New York City in order to recognize the Spatio-temporal patterns for criminal incidents. Thus, from the data analysis, we can distinguish three major crime types that occurred from 2006 to 2019. Additionally, from the temporal visualization analysis on the scale of days, the highest rate of activity and recurrence of criminal incidents was on Saturdays and Sundays, specifically from 12 am until 6 am.

Moreover, we proposed a methodology to classify and predict crimes type by classifying the spatial-temporal


Figure 8: Streamlit User Interface.

locations using three machine learning algorithms; Random Forest, LightGBM, and XGBoost classifiers. As a result, LightGBM and XGboost classifiers were very close in prediction accuracy for 0.59 and 0.6, respectively. However, they fail in getting decent predictions in general. Finally, We have created a user interface to enable users to enter their information and get the class of the crime that can happen in a particular location at a specific time.

As future work, for better classification, it seemed that Deep Learning like Deep Artificial Neural networks or Deep Auto-Encoders might be used. Deep Learning has been in driving condition over traditional ML models in recent past years. Thus, it has good potential for better classification performance in predicting criminal types and hotspots.

## References


[1] New York Police Department (NYPD). Nypd announces citywide crime statistics for october 2021. https://www1.nyc.gov/site/nypd/news/pr1103/nypd-citywide-crime-statistics-october-2021, 2021.

[2] Hitesh Kumar Reddy ToppiReddy, Bhavna Saini, and Ginika Mahajan. Crime prediction monitoring framework based on spatial analysis. *Procedia Computer Science*, 132:696–705, 2018. International Conference on Computational Intelligence and Data Science.

[3] Abrar A. Almuhanna, Marwa M. Alrehili, Samah H. Alsubhi, and Liyakathunisa Syed. Prediction of crime in neighbourhoods of new york city using spatial data analysis. In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, pages 23–30, 2021.

[4] Niyonzima Ivan, Emmanuel Ahishakiye, Elisha T. Opiyo Omulo, and Danison Taremwa. Crime prediction using decision tree (j48) classification algorithm. 2017.

[5] Aziz Nasridinov, Sun-Young Ihm, and Young-Ho Park. A decision tree-based classification model for crime prediction. In *ITCS*, 2013.

[6] New York Police Department (NYPD). Nypd complaint data historic. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i, 2016.

[7] Wright MN Heuvelink GBM Hengl T, Nussbaum M and Gräler B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. 2018.

[8] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. volume 30. Curran Associates, Inc., 2017.

[9] Tianqi Chen and Carlos Guestrin. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016.