New York City crime prevention using Machine Learning

Mohamed Amine Toumi, Mohamed Aziz Tmar, Jassem Rabaoui and Imen Azzouz

Abstract— **Since the 1940s, ecological studies [1] of crime have shown that crime is not evenly distributed across cities, but is often concentrated in certain neighborhoods and at certain dates. In this paper, different visualization analysis of Spatio-temporal New York crime data are adopted to give the public authorities a way to evaluate the places at high risk and to predict the volume of crime committed in different places of New York. The work was divided into three stages: In the first stage, the raw dataset was processed, explored, and analyzed using various visualization techniques. Then, in the second stage, The user's information as well as their location was included into the machine learning algorithms that were used to make predictions about the various types of crime. Finally, in the third and final stage, a user interface was designed that implements the machine learning algorithm that yielded the best results out of the ones that were tested.**

I. INTRODUCTION

The crime rate is rising steadily. One of the most pressing problems that's becoming worse and worse is crime. The quality of life and the viability of societies are impacted by the prevalence of crime and, more importantly, the emotions of insecurity that may result from it. With the increasing rise in the number of crimes, an examination of crime has also become necessary. Analysis of crime is mostly comprised of processes and techniques that work toward the reduction of potential for criminal activity. It is a method used for identifying and analyzing patterns of criminal behavior. However, the most significant problem is to analyze the ever-increasing volume of crime data in an effective and precise manner. Therefore, a tool for crime prediction and analysis was required in order to efficiently detect patterns of criminal behavior. This paper provides various approaches for predicting where and when certain types of crimes are more likely to occur. Based on similarities, classification aids in the extraction of features and the prediction of future trends in crime data. Machine learning algorithms used in this study are LightGBM, Xgboost, Decesion tree and Random Forest Classifier. The paper organisation is as follows. The introduction of the study is described in Section one. Section II consists of the related works. Section III consists of the persued methodology. Section IV consists of Conclusion.

II. REALTED WORK

Numerous studies have been conducted in an effort to find solutions to the problem of lowering crime rates, and a great number of algorithms for predicting criminal behavior have been suggested. Some of the studies have been centered on a particular category of criminal behavior, others has focused on a given level such as in a city. One study looked at crime rates in different areas of London by analyzing cell phone data, demographic data, and court documents. Using a grid system, they determined whether or not each individual cell in London was a high-crime area[2]. The identification of spatial and temporal crime hotspots was the subject of another paper. Two crime datasets were compiled from Denver and Los Angeles and analyzed statistically. The Apriori algorithm was then used to identify recurring patterns in the hotspots. Location, time, and day were the recurring features. Then, classifiers like as Decision Trees and Naive Bayes were used to predict the nature of the crime. Data from Denver was coupled with demographic information to identify potential causes of crime.[3] An other paper[4] discusses the use of machine learning and data mining techniques for crime prediction and prevention. The authors provide an overview of the various approaches that have been used, including decision tree algorithms, artificial neural networks, and support vector machines.

III. METHODOLOGY

In this section, we will talk about the methodology that we have adopted to reach our objective,

which is the prediction of the types of crimes that a person can suffer according to specific data provided, such as age, sex, position, time, etc. Figure 1 illustrates the pipeline of our work.

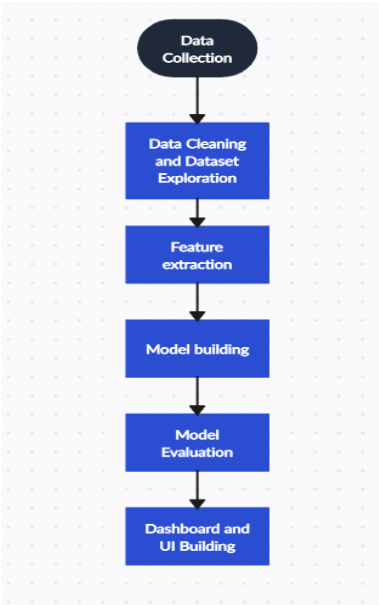

Fig. 1. Pipeline

A. Data Collection

We will use a database provided by the New York Police Department (NYPD). This dataset includes all valid crimes, misdemeanors, and violations reported to the New York City Police Department (NYPD) from 2006 to the end of 2019. It contains 7 million complaints and a documentation file that explains the meaning of each line provided. It provides categorical numerical and time series data

B. Data Cleaning and Dataset Exploration

We are limited to the period between 2006 and 2019. Then we process null values, rows with no name and we convert the dates into a specific format: extracting the day, month and time of the complaint into unique columns based on CMPLNT FR DT and CMPLNT FR TM provided in the dataset. Also we delete some columns that are not important for prediction, such as CMPLNT NUM, SUSP RACE, SUSP SEX, JURISDICTION CODE, because most of the rows in these columns are null. Then we encode some features such as a column named LAW CAT CD, which contains three categories and presents the severity of the crime: Violation, Misdemeanor and Felony , in 0, 1 and 2.

Figure 2 illustrates crime level by age. Certain age groups may be more vulnerable to certain types of crime. For example, children and teenagers may be more at risk of being victimized by bullying or other forms of victimization at school. Elderly individuals may be more at risk of financial scams and other types of fraud.

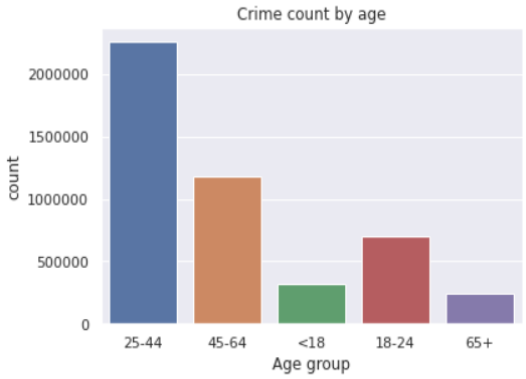

Fig. 2. Crime level by age

C. Feature extraction

We proceeded with different approaches like the correlation matrix between the different features of the dataset. figure 3 illustrates the correlation between the different features.

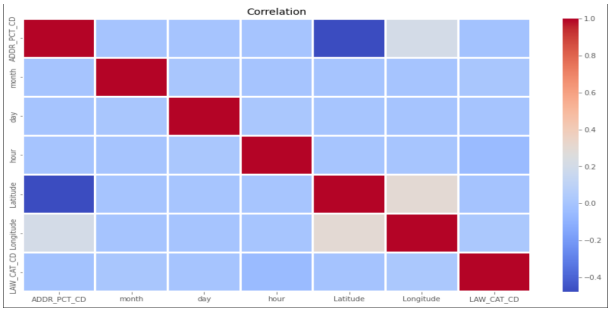

Fig. 3. Correlation matrix of the Features

Crime rates can vary significantly even within a single borough. Some boroughs, such as Staten Island and Queens, tend to have lower crime rates compared to other boroughs, such as Brooklyn and the Bronx.

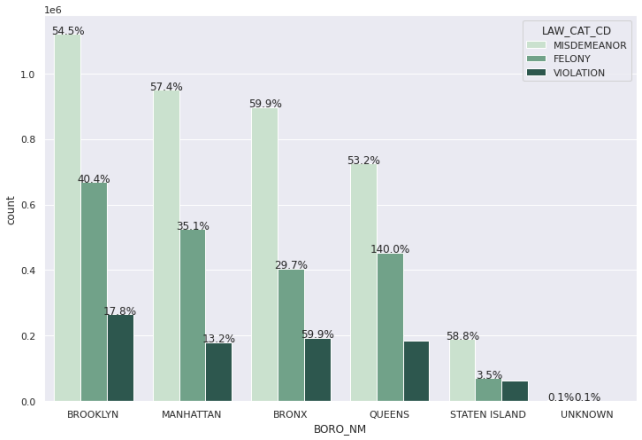

Fig. 4. Crime rates at each Borough

Crime rates tend to be higher at night and during the early morning hours. This pattern is often attributed to the fact that there are fewer people on the streets and in public places during these times, which can make it easier for criminal activity to go undetected.

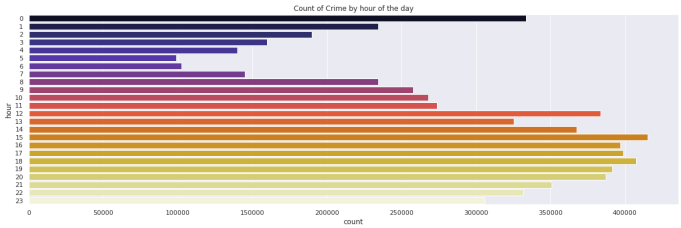

Fig. 5. Crime rates by hour of the day

Using folium plugins to display the map of New York City with the locations of the rapes based on the features of longitude, latitude and PREM TYP DESC. Figure 4 shows the map of New York City with the locations of the rapes.

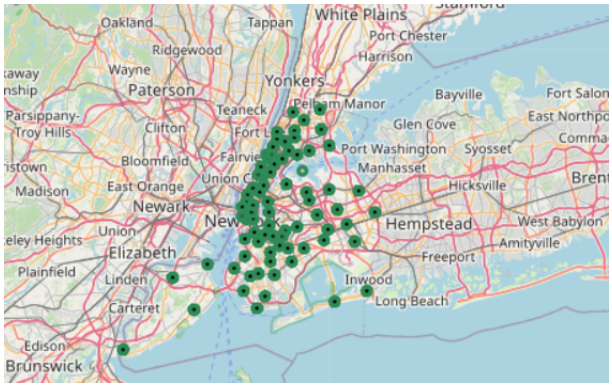

Fig. 6. The map of New York City with the locations of the rapes

Crime rates in the New York have been decreasing in recent decades, as shown in figure 7.

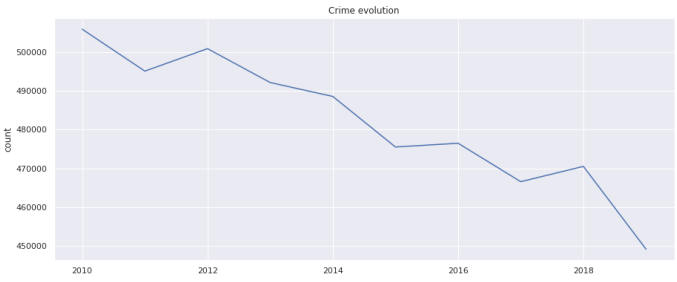

Fig. 7. Crime evolution by year

D. *Model building*

In this step, we implemented several classification methods and evaluated them with the different performance metrics. We used these models:

- LightGBM : short for light gradient-boosting machine, is a free and open source distributed gradient-boosting framework for machine learning. It is based on decision tree algorithms and used for grading, classification and other machine learning tasks. The development is focused on performance and scalability..
- XGBoost : which stands for Extreme Gradient Boosting, is a distributed and scalable gradient boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification and ranking problems. What makes it fast is its ability to perform parallel computations on a single machine..
- Decision tree : falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.
- Random forest : is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample.

E. Model Evaluation

For this study, we trained and tested all the above mentioned algorithms. From the results shown in figure 4, we observe that the accuracy of the Random forest model is the highest. Also this model reached an accuracy value equal to 0.58.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.58	0.54	0.58	0.49
Decision Tree_*	0.5778	0.53	0.58	0.48
Decision Tree_**	0.5780	0.52	0.58	0.47
LightGBM	0.5794	0.51	0.5	0.49
XGBoost	0.5778	0.5	0.52	0.47

* Decision Tree Gini
** Decision Tree Entropy

F. Dashboard and UI Building

As for the visualization of the final results of our research, we created a graphical user interface in the form of a web application using the Streamlit framework and web mapping techniques. In this interface, we have drawn a map of New York and given the user the option to select a location on the map they plan to visit. The user can also enter their age, gender, race, and the time they will visit the site. As a result, we display the most likely type of crime that may be committed against the user, along with its probability. On the server side of our application, we integrated our ready-to-use machine learning model in Pickle format and created an API. This API takes data provided by the user, applies necessary transformations, predicts the type of crime using the model, and returns the result to the client side of the application.

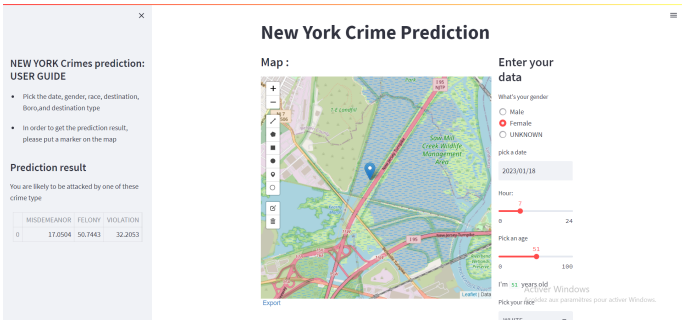

Fig. 8. User Interface

IV. CONCLUSION

Crime prediction is a complex task that involves analyzing various factors such as location, time, and demographic information. Accurate crime prediction can help authorities allocate resources and prevent crime. In this work, we used the database provided by NYPD. First, we understood the meaning of the data through the documentation provided with the dataset. Then, we cleaned and explored the data so that we could choose the features we would use as input to our model. Then we tested several machine learning algorithms to predict the types of crimes that a person can suffer according to the data that he will provide. And following the results provided we chose Random Forest as our final model due to its best metrics. Finally, we made a dashboard that helps the user to enter specific data to display as a result the most likely type of crime that will be committed against him.

REFERENCES

[1] Rafael Prieto Curiel, “Measuring the Distribution of Crime and Its Concentration” Journal of Quantitative Criminology volume 34, pages775–803 (2018).

[2] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, Alex Pentland, “Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data”, 16th International Conference on Multimodal Interaction, November 2014, Trento, Italy.

[3] Tahani Almanie, Rsha Mirza, Elizabeth Lor,”Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots", International Journal of Data Mining Knowledge Management Process (IJDKP), July 2015, Boulder,USA.

[4] "Combating crime with machine learning: A review of data mining and machine learning approaches for crime prediction and prevention" by D. J. Hand and J. F. Hunter.