Analysis and prediction of crime in New York City

Chebaane Zeineb
Higher School of Communication of Tunis
zeineb.chebaane@supcom.tn

Arfa Sirine
Higher School of Communication of Tunis
sirine.arfa@supcom.tn

Abbes Malek
Higher School of Communication of Tunis
malek.abbes@supcom.tn

Arij Flihi
Higher School of Communication of Tunis
arij.flihi@supcom.tn

Abstract

Crime is a serious and pervasive social issue that exists everywhere. The rate of crime has significantly increased in recent years. Advanced systems and fresh ideas are required for enhancing crime analytics in order to protect communities in response to this rise. Even though effective real-time crime prediction lowers crime rates, it is still a challenging topic for scientists because crime incidences depend on numerous complicated elements. In order to anticipate crimes in New York City, a variety of visualization approaches and machine learning algorithms are used in this study.

A raw dataset was processed in the first step, and several visualization techniques were used to better comprehend the data and the relationships between the various variables. After that, a machine learning system was employed to forecast different crime categories depending on user input and location of the users. The final stage is to create a user interface with folium and flask to simplify user interaction. This github repository contains the finished code: https://github.com/SirineArfa/New-York-City-Crime-Prediction-Project

Keywords: *Crime Analysis; Crime prediction; Data Visualization; Crime Maps; Machine Learning; Classification; Folium; Flask; Python; Random Forest Classifier*

1 Introduction

Proper analysis of past crime data aids in crime prediction and subsequently assists efforts to lower the crime rate. Investigating crime reports and immediately spotting fresh patterns, series, and trends are all part of the analysis process. This study aids in the quick preparation of statistics, queries, and maps. Because criminals are active and tend to operate in their comfort zones, it is possible to forecast the type of crime they will do next because, if they succeed in committing the first crime, they are likely to repeat it.

The following offense is typically attempted in a similar place and at a similar time. Studies indicate that there is a significant likelihood of repeat, even though this may not be true in all situations, making crimes predictable. This project suggests creating a web application and visual interface for a crime prediction tool in Python utilizing a variety of libraries, including Flask for the user interface, Folium for interactive leaflet maps, Pandas for data processing, etc. The suggested framework employs multiple visualization techniques to demonstrate the trend in crimes and various machine learning algorithmic approaches to forecast crimes.

Data preprocessing, data visualization, and model construction are the most crucial phases, and they are covered in more detail in the following sections. In a nutshell, the preprocessing stage entails data cleansing and transformation. Finally, in the model-building phase, we employed the Random Forest Classification algorithm to categorize the crimes

that may occur in a specific location. The visualization phase creates numerous reports and maps for the diagnostic and analysis process.

2 Related Work

Crimes are a common social problem affecting the quality of life and the economic growth of a society [1]. It is considered an essential factor that determines whether or not people move to a new city and what places should be avoided when they travel [2]. Today, a high number of crimes are causing a lot of problems in many different countries. In fact, scientists are spending time studying crime and criminal behaviors in order to understand the characteristics of crime and to discover crime patterns. Dealing with crime data is very challenging as the size of crime data grows very fast, so it can cause storage and analysis problems. In particular, issues arise as to how to choose accurate techniques for analyzing data due to the inconsistency and inadequacy of these kinds of data. These issues motivate scientists to conduct research on these kinds of data to enhance crime data analysis. The objective of this research is to apply suitable machine learning algorithm on crime data to predict the likelihood of a county having low, medium or high violent crimes

2.1 Crime analysis

Criminology is an area that focuses on the scientific study of crime and criminal behavior and law enforcement and is a process that aims to identify crime characteristics.[3] It is one of the most important fields where the application of data mining techniques can produce important results. Crime analysis, a part of criminology, is a task that includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Identifying crime characteristics is the first step for developing further analysis. The knowledge that is gained from data mining approaches is a very useful tool which can help and support police forces. 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

2.2 Why Crime is Predictible

There is a strong body of evidence to support the theory that crime is predictable (in the statistical sense) mainly because criminals tend to operate in their comfort zone . That is, they tend to commit the type of crimes that they have committed successfully in the past, generally close to the same time and location. Although this is not universally true, it occurs with sufficient frequency to make these methods work reasonably well. There are major theories of criminal behavior, such as routine activity theory, rational choice theory, and crime pattern theory. These theories are consolidated into what is referred to as a blended theory.

A previous work that inspired us to look more into crime prediction is crime prediction based on weather, crime data, and temporal data [4]. In the paper, the authors employed feature selection techniques to determine the most significant features mainly the most occurred crimes and the correlation between the features, in forecasting crime calculations and rates in New York City over 5 years. They used both machine learning and deep learning techniques and provided benchmarking based on the prediction accuracy. Another interesting work that motivated us is spatiotemporal crime forecasting using Amsterdam police Data [5] in which they focused on Crime history variables, Environmental variables, Demographic variables,Socio-economic variables, and Proximity variables to provide more detailed and reasonable comprehension and prediction that highlights the reasons of the committed crimes.

3 Methodology

In this section, we explain our methodology on how to build a machine learning model and cross-validate it on the New York crime data. We want to predict the different categories of crime based on : time , victim description and location, thus, we implement this workflow :

- Dataset extraction : This work relies on NYPD Complaint Data Historic dataset. This dataset includes different categories of 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.
- Data Preparation : To further understand the data in hand and analyze the different distributions and relations between features, we undergo an Exploratory data analysis in order to answer questions about what, where and when crimes occur, and dealing with null values, outliers and unnecessary caracteristics in order to clean the data and get the best possible accuracy for our machine learning model.
- Feature extraction: To select significant features we used the documentation provided with the data to select the features that we are going to need in our work then we used those features to create additional features such as correlation matrix , encoding techniques and detailing the time and dates like year , month , day and time zone.
- Modeling : After the data pre-processing, in order to classify three different types of crimes : felonies, misdemeanors and violations, based on their severity, we applied Random Forest model, The classification is mainly used to recognize the labelled classes by knowing their attributes in the dataset, thus predicting the class label for instances with known features. Hence, using the classifiers in crime prediction constructs a futureoriented model to identify the criminal type within a specific time.
- Model evaluation : we used the confusion matrix and the

ROC curve (receiver operating characteristic curve), as an appropriate metrics to measure model performance.

- Result visualization : we created a dashboard that presents the crimes prediction distributed in a map .

3.1 Data-Exploration and cleaning:

After studying the description of the features provided by our NYPD dataset we decided to keep just 12 features which are: CMPLNTFRDT, CMPLNTFRTM, OFNSDESC, LAWCATCD, XCOORDCD, YCOORDCD, Latitude, Longitude, PATROLBORO, VICAGEGROUP, VICRACE and VICSEX. Those features present the victim description date time location and the type of the crime which are the necessary information that we need to predict the type of the crime based on location and the person characteristics. We opted to fill missing values based on the distribution of the values in the data-set. As for the timestamp values we replaced all the nulls with the median value in each column and deleted outliers with unreasonable values in years and ages.

index	CMPLNT_FR_DT	CMPLNT_FR_TM	OFNS_DESC	LAW_CAT_CD	X_COORD_CD	Y_COORD_CD	Latitude	Longitude	PATROL_BORO	VIC_AGE_GROUP	VIC_RACE	VIC_SEX
0	09/04/2014	17:25:00	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	1005945.0	188862.0	40.685041	-73.921777	PATROL BORO BKLYN NORTH	25-44	WHITE	F
1	10/12/2016	07:40:00	GRAND LARCENY	FELONY	947034.0	171375.0	40.636991	-74.134093	PATROL BORO STATEN ISLAND	45-64	WHITE HISPANIC	F
2	09/28/2012	13:30:00	GRAND LARCENY	FELONY	1014179.0	239453.0	40.823876	-73.891863	PATROL BORO BRONX	45-64	WHITE HISPANIC	F
3	03/24/2015	15:45:00	PETIT LARCENY	MISDEMEANOR	1009041.0	247401.0	40.845707	-73.910398	PATROL BORO BRONX	<18	WHITE HISPANIC	F
4	05/20/2017	04:02:00	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	1031779.0	217663.0	40.763992	-73.828426	PATROL BORO QUEENS NORTH	25-44	BLACK	M
...	...	...	...	...	...	...	...	...	...	...	...	...
6959136	03/02/2014	20:30:00	ROBBERY	FELONY	1013232.0	236725.0	40.816392	-73.895296	PATROL BORO BRONX	45-64	BLACK	M
6959137	03/13/2013	12:00:00	GRAND LARCENY	FELONY	997737.0	151811.0	40.583360	-73.951446	PATROL BORO BKLYN SOUTH	25-44	WHITE	F
6959139	12/20/2012	13:30:00	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	993205.0	175595.0	40.648648	-73.967731	PATROL BORO BKLYN SOUTH	25-44	BLACK	F
6959140	04/04/2013	13:00:00	PETIT LARCENY	MISDEMEANOR	994185.0	200150.0	40.716045	-73.964163	PATROL BORO BKLYN NORTH	45-64	WHITE	M
6959142	08/19/2012	01:00:00	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	1009968.0	182502.0	40.667573	-73.907296	PATROL BORO BKLYN NORTH	25-44	BLACK	F

Fig1: NYPD dataframe after cleaning

3.2 Feature extraction:

For this part, we built derived values from our initial data which are more informative and non redundant. We started with generating year, month and day columns based on CMPLNTFRDT. Then from CMPLNTFRTM we categorized the different daytime into four classes: morning, afternoon, evening and night.The same way, having 14 types of crimes we grouped them into only 15 classes thus our prediction will be processed according to a fewer number of classes.

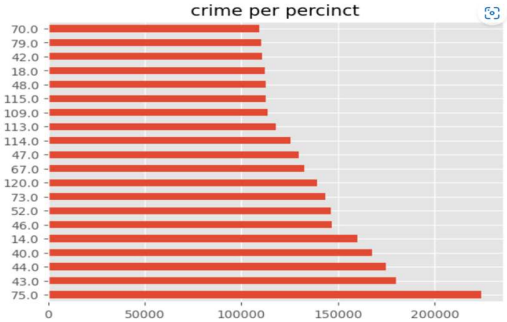

Fig2: The occurence of crime per percent

Fig3: The number of crimes per patrol borough

After that, we opted to convert the categorical features to numbers such that themodel is able to understand and extract valuable infor- mation. We used two types of encoding :

OneHotEncoding : All of PATROLBORO, VIC- SEX and VICRACE were encoded using this tech-nique as these variables don't present any natural order to take into consideration.
OrdinalEncoding : We encoded OFNSDESC, VICAGEGROUP and timezone columns.

3.3 Random Forest Classification Model:

which is a meta estimator that fits a number of de- cision tree classifiers on various subsamples of thedataset and uses averaging to improve the predic- tive accuracy and control over-fitting and we used sklearn python library to implement the model. After training, we evaluate this model which gave us : 0.519 as score of the accuracy.
To evaluate our model we used the confusion matrixand the ROC curve as shown by the figures below :

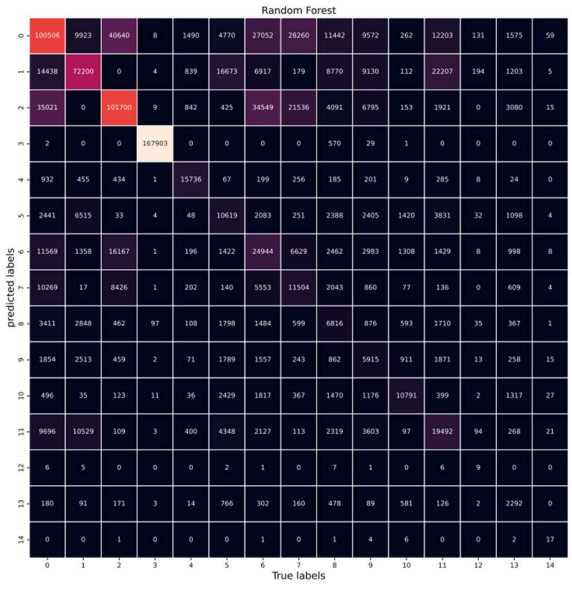

Fig4: Random Forest Confusion matrix

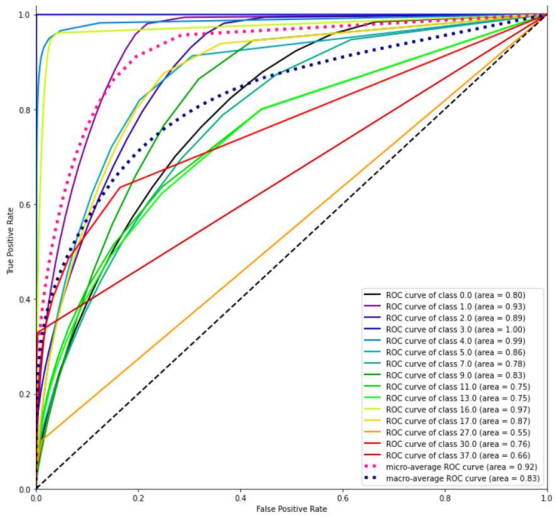

Fig5: Random Forest ROC curve

3.4 User Interface:

we built a web application using Flask and Folium to allow the user to interact with the map and pre- dict the type of crime that could happen. The user can enter his gender, race, age, the date and hourin which he wants to predict the type of crime, the location on the map and finally, the place. This in- formation is then transformed to fit the model in- put, and then, using the loaded model weights file, we predict the type of crime and send it back to theuser along with the potential subtypes of that crime.

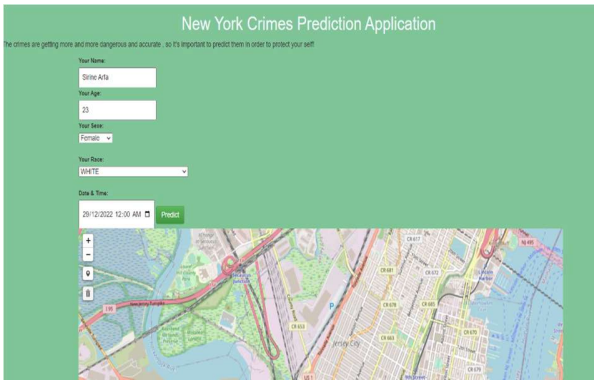

Fig6: User Interface

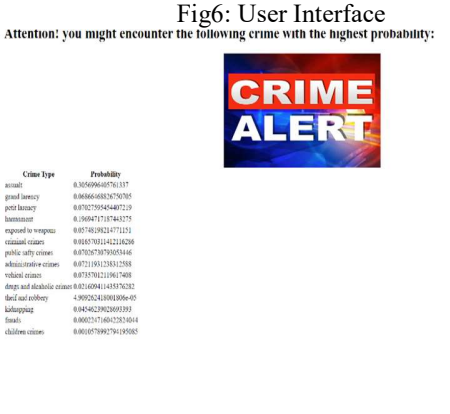

Fig7: Crime Prediction

4 Conclusion

The perception of a community as crime ridden can deter people from going there and induce residents to move away. This causes damage to the economy. Crimes affects the economy by placing a financial burden on taxpayers and governments because of increased needs for police, courts and corrections facilities, as well as intangible costs including psychological trauma and reduced quality of life for crime victims.

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 work, we used the data-set provided by the NYPD. As a first step we explored the data in order to understand it's pattern and produce insights.Then We kept the most essential features by performing techniques of feature selection and feature extraction, After that, we applied our Machine Learning Model Random Forest. As a result Random Fores was very close in prediction accuracy for 0.55. 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.

5 References

[1] Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pentland, 'Once Upon a Crime, Towards Crime Prediction from Demographics and Mobile Data', CoRR, vol. 14092983, 2014.

[2] R. Arulanandam, B. Savarimuthu and M. Purvis, 'Extracting Crime Information from Online Newspaper Articles', in Proceedings of the Second Australasian Web Conference - Volume 155, Auckland, New Zealand, 2014, pp. 31-38.

[3] Malathi A., Santhosh B.S,. Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters; Global Journal of Computer Science and Technology; Volume 11 Issue 11 Version 1.0 July 2011.

[4] Elluri, Lavanya Mandalapu, Varun Roy, Nirmalya. (2019). Developing Machine Learning Based Predictive Models for Smart Policing. 10.1109/SMARTCOMP.2019.00053

[5] Rummens, Anneleen Hardyns, Wim Pauwels, Lieven. (2017). The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. Applied Geography. 86. 10.1016/j.apgeog.2017.06.011.
[6]