New York City Crime Prediction

1st Chayma HAMDI
chayma.hamdi@supcom.tn

2nd Jawhar MARZOUGUI
jawhar.marzougui@supcom.tn

3rd Nada ZAHRA
nada.zahra@supcom.tn

4th Yosr ABID
yosr.abid@supcom.tn

***Abstract*—Crime is an extremely complex phenomenon in which many factors are involved. The rate of crimes in an area can significantly affect its inhabitants either directly by causing them harm or indirectly by driving away businesses, as investors search for safer environments. This, paired with the fact that poor neighborhoods are usually the most prone to crime in the first place, creates a feedback loop that traps these regions in poverty and generates more incidents. On a larger scale, countries suffering from high crime rates risk their international reputation as well as their ability to attract foreign investors and may be denied future loans.**

Although major cities are usually safer than rural areas due to a denser presence of law enforcement, that didn't stop certain categories of crime from rising. For example, the number of murders rose by more than 4% in the USA in 2021 compared to 2019 according to the FBI.

without no doubt,addressing crime requires sufficient resources. This paper aims to showcase the various machine learning techniques we have employed to develop models capable of predicting crime occurrences in New York based on multiple variables.

I. Introduction

A. Overview

Crimes have impacts on many sides of life both individually by limiting a persons freedoms suck as roaming the streets, risk on personal property, moving houses, etc.. as well as collective effects mainly on the economy and the countries reputation on the international scene. The reallocation of resources from economic growth to law enforcement places even more financial burden on the countries budget.

This paper aims to introduce methods of reducing the risk of citizens falling victim to crimes by collecting, mapping and analysing crime data in a region to extract useful insights, like the variables that affect crime rates the most, places and routes to avoid and finally build a model that is capable of predicting the probability and severity of a crime taking place in a predetermined scenario.

We built a web interface that can be used to input a scenario and query the model then display the models response/results in an illustrative and clear manner.

B. Related work

Due to the high relevance of analyzing and predicting crimes, numerous works and researches have tackled this issue. With such a complex issue at hand, there are many factors that come into play that can be exploited. Varun Roy [1] et al implemented a state-of-the-art method that relied on weather and its temporal features to identify relevance to crime. With feature selection techniques, they successfully identified the most important features in order to efficiently train machine learning models for forecasting crimes. Almuhanna [2] et al have opted for a spatial-temporal data analysis approach to localize crime hotspots and their types itself is a crucial element that would greatly help police in its investigation and in taking necessary precautions in certain areas.

II. Exploratory data analysis

To properly utilize the crime's data, Exploratory data analysis is a crucial step that helps in identifying data patterns, obvious errors and defines the methodology to face this problem.

A. New York Crime Dataset

This work exploited the NYPD Complaint Data Historic dataset, a dataset recording approximately 7 milion valid felony, misdemeanor and violation crimes reported to the police ranging from 2006 to 2019. With more than 25 different crime types and 35 spatial-temporal features, this dataset helps in creating a robust model for crime prediction.

B. Crime classification

Crimes in the New York Dataset are divided into 55 types that are classified into three big categories: Felony, Violation and Misdemeanor. Figure 1 shows that the most common type

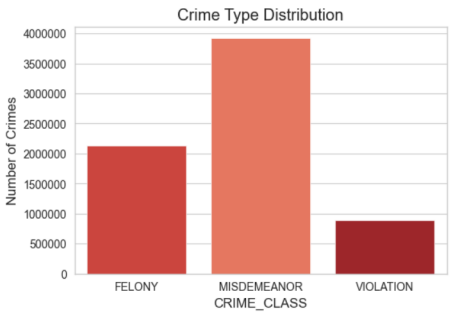

Fig. 1. Amount of crimes per type

is misdemeanor which portrays its frequent occurrence in New York with near double the amount of the other types.

It is also mandatory to examine the characteristics of the victims to determine the features of people who are most prone

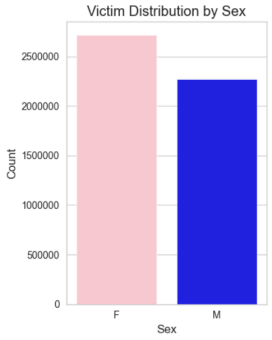

Fig. 2. Amount of crimes per gender

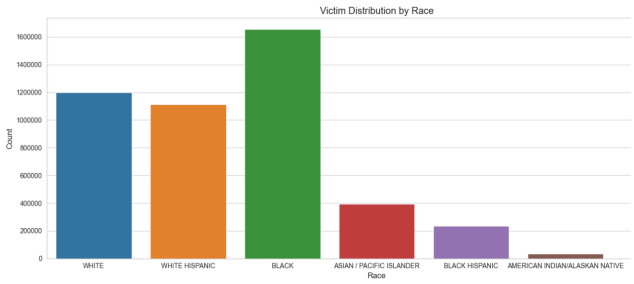

Fig. 3. Amount of crimes per race

to the occurrence of a crime. Figure 2,3 and 4 highlight the number of crime per age, race, and gender.

It is evident that victims that are women, between the age of 25 and 44 and of black race are the most susceptible to crimes.

C. Spatial analysis of the dataset

Localizing where the crime was commited greatly benefits the police to take precautions beforehand. The New York Crime Dataset provides the longitude and latitude of each crime. Therefore, we deduce the number of crimes commited by Borough.

According to figure 5, Brooklyn and Manhattan have the highest crime rate. These statistics can further be clearly illustrated by drawing a heatmap to showcase the crime hotspots.

D. Temporal analysis of the dataset

According to the U.S bureau of Justice Statistics [3], seasons and time undeniably have a huge influence on the crime rate. For that reason, it is necessary to examine the temporal distribution of crimes. Figure 7,8 and 9 displays an abnormal high rate of crimes in summer seasons, in the middle of the week and from 3 pm to 1 am.

This demonstrates the importance of the features of date in correctly predicting the crime.

E. Data cleaning

Before testing models, the dataset must undergo a data cleaning process. The first step is to identify the number of missing data as they are detrimental and can pose issues in

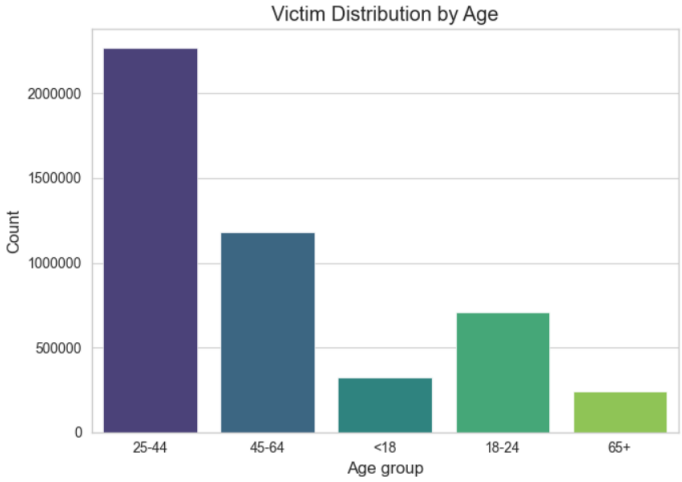

Fig. 4. Amount of crimes per age

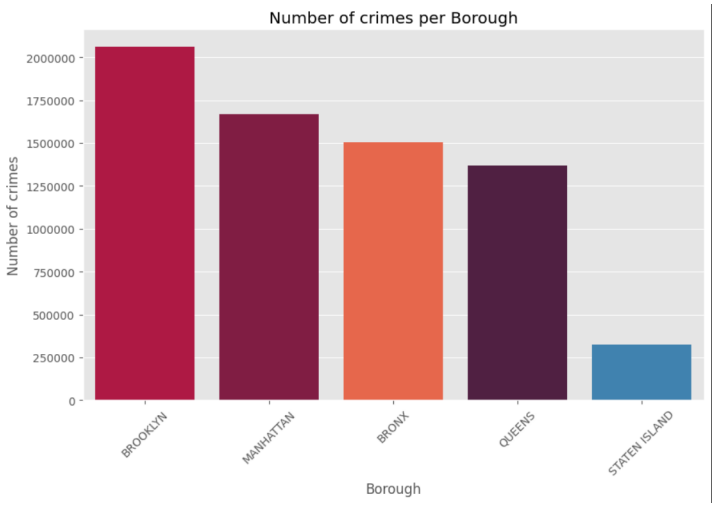

Fig. 5. Number of crimes in different Borough

the training phase. Figure 10 shows the percentage of missing data per column.

Columns 'TRANSIT_DISTRICT', 'HOUSING_PSA', SUSP_AGE_GROUP' and 'SUSP_SEX' have more than 50% missing data that could not be filled. Therefore, we dropped them as they do not have sufficient amount of data. Columns 'Y_Coord_CD', 'X_COORD_CD' and 'Lat_Lon' featured redundant data that are already present in columns 'Latitude' and 'Longitude'. Consequently, we dropped them. Columns 'VIC_RACE', 'VIC_SEX', 'VIC_AGE_GROUP' and 'BORO_NM' were crucial for the training of the model but contained 300 missing values. These values were replaced with the value "unknown", while for the column 'longitude' and 'latitude', some values were missing and the rows had to be deleted since there was no way to fill the empty value. Columns that provided data that does not contribute in the prediction of crimes such as 'PD_CD', 'KY_CD'

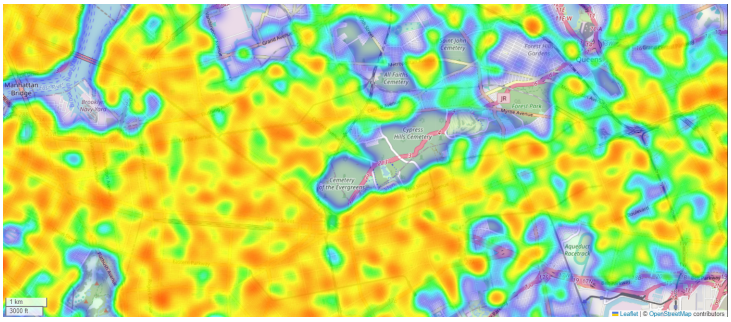

Fig. 6. New York crime heatmap

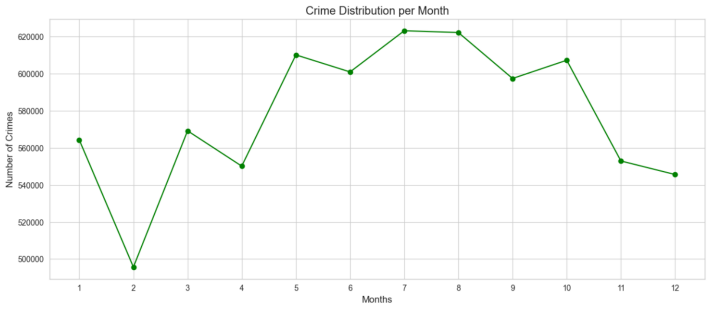

Fig. 7. Number of crimes per month

were additionally dropped as well. The date column was transformed to acquire years, months, days and hours separately. Columns containing categorical data such as the type of crime were encoded. This step is mandatory as the model cannot be fed categorical data. After this data cleaning process, the dataset contained 6958307 rows and 16 features.

III. Modeling

After data cleaning phase, the dataset can now be exploited for machine learning models in order to be able to accurately predict the crimes. Since there are many algorithms that are theoretically able to achieve that task, it is necessary to test many of them and benchmark the results to pick the most suitable one. Supervised classifers are a logical solution due to their ability to identify and classify a chosen feature given labeled attributes of the dataset and after sufficient training. The following classification models were chosen:

- RandomForest: RandomForest is composed of numerous decision trees that are fed sub-samples of the dataset. Then, each tree votes to assign the appropriate class and the final decision is based on the majority of those votes in the case of classification or the average in case of regression. This way, it prevents it from overfitting.
- LightGBM: Light Gradient-Boosting Machine is a tree-based algorithm. Compared to other gradient-boosting machine algorithms, LightGBM is much faster and is much more suitable for large datasets. However, it is much more prone to overfitting especially if given a small dataset.

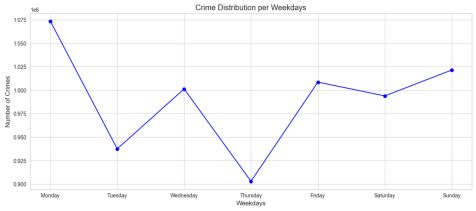

Fig. 8. Number of crimes per week

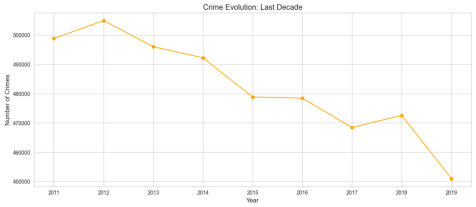

Fig. 9. Number of crimes in the last decade

- XGBoost: XGBoost is based on the Gradient Boosted Decision Tree algorithms with optimized features and methods for faster training and more accurate results.
- Multi-layer perceptron: MLP classifier is widely used to solve complex problems stochastically. It can solve non linear problems while handling a large input data well.

All of the above models were trained on the same dataset. In order to identify the best-performing model, the following evaluation metrics were used:

- Confusion Matrix: Confusion Matrix visualizes the performance of an algorithm in a table. The rows represent the instances of the actual class while the columns represent the predicted class; hence, the confusion matrix showcases the number of correct and false predictions for each class in a clear manner. With the confusion matrix, we can calculate the Accuracy, Precision, Recall and F1 score.
- Accuracy: Accuracy represents the number of correct predictions the model achieved.
- Precision: The precision metric indicates the proportion of correct positive predictions of the model.
- Recall: Recall evaluates the number of correct positive predictions out of all the actual positives.
- F1 score: F1 score assesses the predictive skills of a model by combining precision and recall to evaluate its performance class-wise.

The table below illustrates the metrics values recorded by the chosen models:

TABLE I
Chosen models performance

Model	Precision	Accuracy	F1 score	Recall
RandomForest	0.51	0.48	0.62	0.57
XGBoost	0.51	0.49	0.60	0.73
LightGBM	0.51	0.48	0.56	0.56

From the results displayed above, all models were close

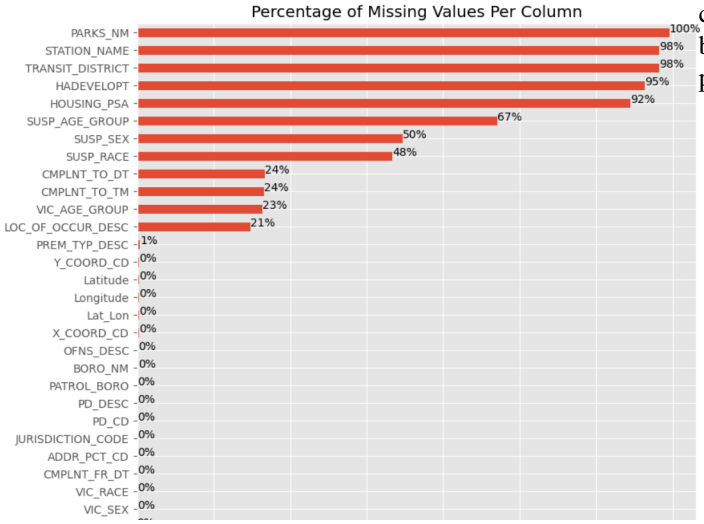

Fig. 10. Missing values per column

in accuracy with XGBoost scoring the highest with 0.49 accuracy. However, XGBoost clearly outperformed the other models in F1 score and Recall. Even though all models had low numbers of true classifications, XGBoost was the best performer; hence, it was the best model for this task. The model's weights after training were saved for further use.

IV. User Interface

To visualize the results obtained and facilitate the utilization of the model for users, we built a web application using Streamlit for the visualization of the map. The user taps in his age, race, gender and date. He can also click on the map on the location he wishes to go to and the coordinates are retrieved. The data is transmitted to the model and the user receives the model's prediction on which type of crime is likely commited there on a person with his characteristics. This helps in spreading awarness and taking necessary precautions.

V. Conclusion

Crime is a notoriously difficult phenomenon to predict due to the countless factors behind it; thus, it has been studied extensively to understand the reasons and variables behind it in order to forecast it and reduce its threat. In this work, we conducted an exploratory data analysis to identify patterns for criminal activity. We discovered the features related to the occurrence of crimes that can be used for the prediction. We then utilized many techniques to clean the dataset for the training process. Moreover, we have implemented four machine learning algorithms and benchmarked them in order to decide the suitable one for our case. Although the results were close, XGBoost managed to outperform the other algorithms in every category. Finally, we have built a graphical web interface for users to give their information as input and acquire the type of threats they are prone to. For future improvements, deep learning models have recently been gaining success, outperforming their traditional machine learning algorithms counterparts. These models have potential to yield a much better result than the ones tested in this work if utilized properly.

References

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

[2] 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.

[3] How Seasons and the Weather Affect Crime Trends https://esfandilawfirm.com/do-seasons-weather-affect-crime-trends