Analysis and prediction of crime in New York City using a Machine Learning approach and Spatio-temporal data

Sirine Kochbati
Higher School of Communication of Tunis
sirine.kochbati@supcom.tn

Anoir Chabchoub
Higher School of Communication of Tunis
bechiranoir.chabchoub@supcom.tn

Eya Baklouti
Higher School of Communication of Tunis
eya.baklouti@supcom.tn

Mohamed Bellakhal
Higher School of Communication of Tunis
mohamed.bellakhal@supcom.tn

***Abstract*—The Urban population is exponentially increasing nowadays, having a large number of people living in a really small surface can present a danger, that's why paying attention to the security of those people is very important. New York City is one of the most crowded cities in the world that presents a huge rate of crime with around 364.7 violent crimes per 100,000 population in the year 2020.**

The Police databases hold a large amount of crime data that can be translated into insights in order to model the current and future crime trends and patterns. Therefore, the predictive analysis aims to optimize the use of this data to anticipate criminal events. To achieve this, we developed a machine learning model for Spatio-temporal prediction that is specifically adjusted for an imbalanced distribution of the class labels and test them in an actual setting with state-of-the-art predictors. For the plotting of our results, we used Web Mapping techniques To present the New York City map and visualize the danger label of each zone as well as the probability of being a potential crime victim for a single individual.

***Index Terms*—Crime prediction, Machine learning, Spatio-temporal modeling, Web Mapping**

I. Introduction

Security is an essential aspect of strengthening the roots of a country. In fact, crimes can make a significant impact on the economic growth of a country. Therefore, countries are spending a substantial amount of their (GDP) on law enforcement agencies to control crimes. Besides, advancement in technology and especially geographical information systems (GIS), assisted the researchers in presenting numerous crime detection and prediction techniques.

The enormous amount of data being available in the past few years have been a great motivation for the scientists to pursue research activities in the field of crime and criminal investigations. Studying as well as trying to understand and predict the crime trends and patterns have been the priority of the law enforcement agencies to make an effective policy by taking advantage of the historical data to lower the crime rates and make a peaceful community. Based on historical data, forecasting crimes has been a subject of interest that recently gained much attention in research, which resulted in proposing a significant number of different methods for the discovery of different aspects related to crime prediction. Crime can be considered as a location-oriented feature as some places can exhibit greater risk of crime to be committed than others due to several factors such as the degree of urbanisation and populated density, the greater rates of migration and population growth in urban populations or the variation of the demographic structure in different areas. It is an understood fact that in a particular area, no matter the size, crime is not distributed evenly, uniformly, or even randomly within that area or city.

Different types of crimes and the full consideration of the protection and safety of citizens in any society are significant components that play a vital and direct role in the quality of the life of residents.Numerous types of crimes can occur in an area with different frequencies. An area may be flagged for higher misdemeanor events while the other for felony.

The inclusion of spatial and temporal information in the crime data-sets using GIS has revolutionized the crime prediction systems. The spatio-temporal information helps the researchers to present more credible and accurate crime prediction systems that can be reliable in crime prevention.

II. RELATED WORK

A. Smart Policing

Smart Policing represents an emerging paradigm in American policing that stresses crime reduction and promotes improvement of the evidence base for policing. Smart Policing emphasizes effectively using data and computational intelligence as well as improving analysis, performance measurement, efficiency and evaluation research along with

Identify applicable funding agency here. If none, delete this.

encouraging innovation.

This introduction defines Smart Policing in historical and contemporary contexts and discusses several important and emerging characteristics in the local Smart Policing sites, namely, the need to improve the evidence base for policing, the police agency-research partnerships that are emerging in Smart Policing, the type of problems identified and approaches undertaken by the SPI sites, and future issues for Smart Policing.[3]

B. Crime Prediction

A previous work that inspired us to look more into crime prediction is crime prediction based on weather, crime data, and temporal data [1]. In the paper, the authors have proposed the utilization of weather information for crime forecasts. They 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 spatio-temporal crime forecasting using Amsterdam police Data [2] 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.

III. METHODOLOGY

In this section, we explain our methodology on how to build and compare different machine learning models and cross-validate them on the New York crime data. We want to predict the crime category based on time , victim description and location. Figure 1 represents the pipeline of our work.

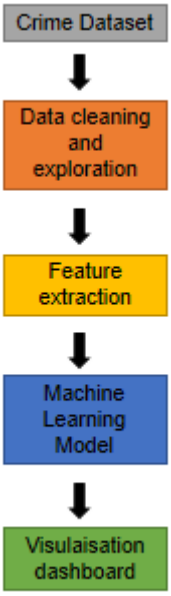

Fig. 1. Project pipline

- Data-set Extraction: We have downloaded the New York Police Department (NYPD) available on Kaggle with all the crime types with approximately 7 million complaints.
- Data Cleaning and exploration: dealing with null values , outliers and unnecessary columns based on the requirement of the project and the documentation provided with the data in order to get the best possible accuracy for our models.

- Feature extraction: To select and extract 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 .

- Machine learning models: In this step, we performed multiple classification methods like RandomForest , SVM , Multi-layer Perceptron and K-NN

- Model evaluation : we used the confusion matrix and observed various performance metrics to evaluate our models

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

A. Data-set Extraction and cleaning :

In this step we have downloaded the data set provided by NYPD website containing 35 columns in which we find 55 type of crimes and a documentation file explains the meaning of each provided feature . After studying the description of the features we decided to keep just 10 Features which are :CMPLNT-FR-DT, CMPLNT-FR-TM,OFNS-DESC, PD-DESC, Latitude, Longitude ,PATROL-BORO, VIC-AGE-GROUP, VIC-RACE, VIC-SEX

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.

B. Feature extraction :

For feature extraction, 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 CMPLNT-FR-DT. Then from CMPLNT-FR-TM we categorized the different daytime into four classes: morning, afternoon, evening and night.The same way, having 55 types of crimes we grouped them into only ten classes thus our prediction will be processed according to a fewer number of classes.

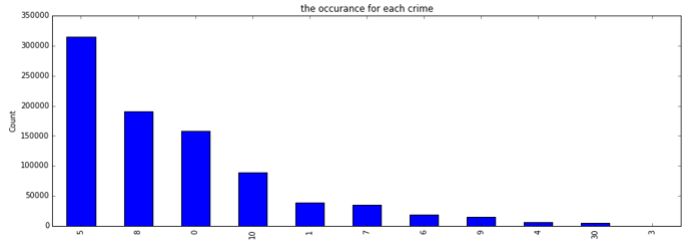

Fig. 2. the occurance for each type of crime

After that, we opted to convert the categorical features to numbers such that the model is able to understand and extract valuable information. We used two types of encoding.

- One hot encoding : All of PATROL-BORO, VIC-SEX and VIC-RACE were encoded using this technique as these variables don't present any natural order to take into consideration.

- Ordinal encoding : We encoded OFNS-DESC, VIC-AGE-GROUP and time-zone columns using this type of encoding.

Finally, we kept only the most relevant features for use in our model construction so we dropped both of CMPLNT-FR-DT and CMPLNT-FR-TM columns.

C. Machine learning models:

Predicting the type of the crimes based on location, timing and the person description can be achieved by developing a classification solution and more precisely a supervised classification solution since the data is labeled. So we tried the following models :

- RandomForest Classification model: which is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting and we used sklearn python library to implement the model.as a result the model gave us

- Multi-layer Perceptron classifier: This model optimizes the log-loss function using LBFGS or stochastic gradient descent.To implement this model also we used sklearn python library.

- K-nearest neighbors algorithm (kNN): Which is used for classification in our case. The class membership of the output is resulted from a plurality vote of its neighbors. In our study we set k to 3.

- Decision Tree Classification algorithm: Which is a tree-structured classifier, where internal nodes represent the features of a data-set, branches represent the decision rules and each leaf node represents the outcome.

- Support Vector Machine (SVM): This model tries to find a hyper-plane in an N-dimensional space (N is the number of features) that distinctly classifies the data points.

After training the previous models and evaluate them we adopted Random Forest model that gave us the following evaluation metrics values

Accuracy	0.713
Recall	0.731
Precision	0.76
F1 Score	0.713

FINAL RESULTS

As to visualize the final results of our research, we created a graphical user interface in the form of a web application using web mapping techniques in which we plotted the New York map and we gave the user the possibility to choose a spot in the map for which he intends to visit, to enter his data such as his age,gender and the time in which he will be visiting the spot and then as an output we displayed the most likely type of crime that will be commited against him.
Behind the scenes or as we may call it in the server-side of our application, we integrated our ready to use Machine Learning model and we created an API that takes the data provided by the user, apply the necessary transformations on it, predicts the type of crime using our model and then sends back the result to the client side of our application.

CONCLUSION

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 various Machine Learning models and compared their performances in order to predict the type of crime expected. as our random forest model gave the best results, we opted to use it as our final model. In the future and as a next stage for this research, we want to include population density based on the location with the current features and observe if this factor plays a significant role in predicting crimes.

REFERENCES

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

[2] 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.

[3] Coldren JR, Huntoon A, Medaris M. Introducing Smart Policing: Foundations, Principles, and Practice. Police Quarterly. 2013;16(3):275-286. doi:10.1177/1098611113497042

[4] Umair, Areeba and Sarfraz, Muhammad Shahzad and Ahmad, Muhammad and Habib, Usman and Ullah, Muhammad Habib and Mazzara, Manuel,Spatiotemporal Analysis of Web News Archives for Crime Prediction,Applied Sciences,2020;8220.doi:10.3390/app10228220