New York City Crimes Prediction using Machine Learning

Afli Ramzi Oumayma Mahmdi

ramzi.afli@supcom.tn, oumayma.mhamdi@supcom.tn

Abstract

A crime is a deliberate act that can cause physical or psychological harm, as well as property damage or loss, and can lead to punishment by a state or other authority according to the severity of the crime. The number and forms of criminal activities are increasing at an alarming rate, forcing agencies to develop efficient methods to take preventive measures. In the current scenario of rapidly increasing crime, traditional crime-solving techniques are unable to deliver results, being slow paced and less efficient. Thus, if we can come up with ways to predict crime, in detail, before it occurs, or come up with a "machine" that can assist police officers, it would lift the burden of police and help in preventing crimes. To achieve this, we suggest including machine learning (ML) and computer vision algorithms and techniques.
Firstly, in order to gain deep understanding of the data we suggest multiple visualization techniques to understand better the data and the relations between the different variables.

After that ,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 interfaces using basic html,css with separate Middleware which is based on fast api (python).

Keywords : Crime prediction; spatial analysis; Feature engineering; Prediction; ML model;

Introduction

Crime is one of the dominant and alarming aspects of society. Everyday a huge number of crimes are committed. These frequent crimes have made the lives of common citizens restless.
So, it is obvious to comprehend the patterns of criminal activity to prevent them.
Furthermore, the capability to predict any crime on the basis of time, location.. can help in providing useful information to law enforcement from a strategical perspective. And they can work effectively and respond faster if they have early information and pre-knowledge about criminal activities of the different points of a city.

However, predicting the crime accurately is a challenging task because crimes are increasing at an alarming rate. Thus, the crime prediction and analysis methods are very important to detect future crimes and reduce them.

In recent times, artificial intelligence has shown its importance in almost all fields and crime prediction is one of them .

In this paper, a supervised learning technique is used to predict criminal activity.

Our proposed solution consists of creating a web application used to predict crimes by analyzing New York city criminal activities data set for 10 years from 2006 to 2015. This data contains records of previously committed crimes and their patterns.

Methodology

In this section, we describe our methodology. to build machine learning models and validate them on New York crime data.
Our objective is to predict the types of crimes that can suffer a person according to very specific data. such as age, sex, position etc...
We are now going to present the approach of our solution in fig1.

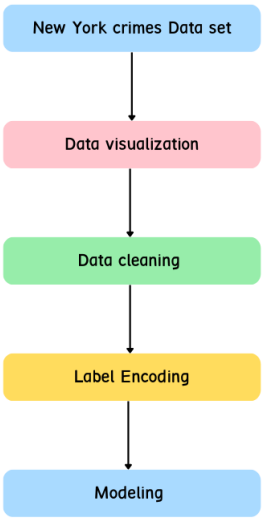


Figure 1: Project workflow

New York crimes Data Set

NYPD Complaint dataset has been published since 2016 and is updated every year. It contains 35 columns where we found 55 types of offences, 7 million lines where each submits a complaint and a documentation file explains ´ the meaning of each column provided.

For example, she includes the date and time of occurrence of the crime and even spatial data of the place of occurrence, data on victim and suspect such as age, race and gender. . . It includes numerical, categorical and time data.

Data Visualization

In order to answer questions about what, where and when crimes occur, we must firstly understand information in our data by visualizing it by plotting graphics , statistics etc ... That's why we have firstly to understand what each column in our data set explains after that we see that data and by that we will be able to detect what are blank fields we can also detect what are the most significant columns

```
crimes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6500870 entries, 0 to 6500869
Data columns (total 35 columns):
 #   Column              Dtype
---  ------              -----
 0   CMPLNT_NUM          int64
 1   CMPLNT_FR_DT        object
 2   CMPLNT_FR_TM        object
 3   CMPLNT_TO_DT        object
 4   CMPLNT_TO_TM        object
 5   ADDR_PCT_CD         float64
 6   RPT_DT              object
 7   KY_CD               int64
 8   OFNS_DESC           object
 9   PD_CD               float64
 10  PD_DESC             object
 11  CRM_ATPT_CPTD_CD    object
 12  LAW_CAT_CD          object
 13  BORO_NM             object
 14  LOC_OF_OCCUR_DESC   object
 15  PREM_TYP_DESC       object
 16  JURIS_DESC          object
```

Figure 2: Data visualization

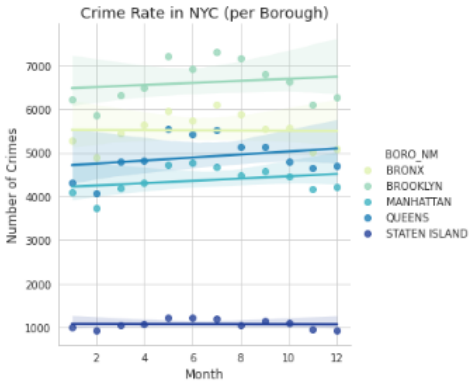


Figure 3: Crime rate by borough

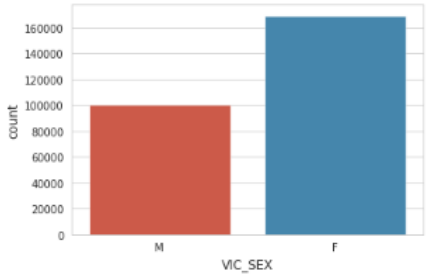


Figure 4: 'VIC$_SEX'Column$

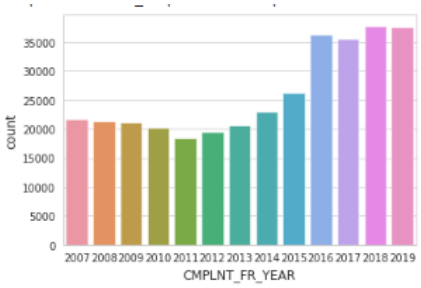


Figure 5: Violations per Year

Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a data set. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. That's Why in this phase we were able to delete more than 50% of blank lines(PARKS NM, STATION NAME, TRANSIT DISTRICT, HADEVELOPT, HOUSING PSA).

Also data related to the suspect person are not available to the user who ' wants to know if he is going to be a victim of crime thats why he must delete these data.

Similarly, we have decided to delete the lines ´ with NaN values for columns and replace these NaN values by the term UNKNOWN in others according to the percentage of empty.

Finally, we chose to fill in the other values missing according to the distribution of values in the time data set. As for timestamped values, we have replaced all null values with the median value in each column.

```
(crimes.isna().sum()/crimes.shape[0]*100).sort_values
```

```
PARKS_NM              99.677597
STATION_NAME          97.783220
TRANSIT_DISTRICT      97.783220
HADEVELOPT            95.021220
HOUSING_PSA           92.268389
SUSP_AGE_GROUP        70.620748
SUSP_SEX              51.619799
SUSP_RACE             49.569088
VIC_AGE_GROUP         25.203411
CMPLNT_TO_DT          24.721322
CMPLNT_TO_TM          24.655177
LOC_OF_OCCUR_DESC     21.455728
PREM_TYP_DESC          0.579969
OFNS_DESC              0.289269
Y_COORD_CD             0.266718
```

Figure 6: Data cleaning

Label Encoding

One of the most important steps for data decision-making is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within our large data set .

So first of all, by computing missing values in our features we noticed that the features [PARKS_NM, HADEVELOPT, HOUSING_PSA, TRANSIT_DISTRICT, STATION_NAME] are almost null values so it's better to delete them to keep just relevant features.

Then we also dropped data with negative time duration Since the data set contain text and there are only a few algorithms such as CATBOAST, decision trees can handle categorical values very well but most of the algorithms expect numerical values to achieve state-of-the-art results.
There are many ways to convert categorical values into numerical values. Each approach has its own trade-offs and impact on the feature set. So we used Label Encoding to convert the labels into a numeric form so as to convert them into the machine-readable form.

Modeling

The quality of the model's results is a fundamental factor to take into account when choosing a model. choosing a model is based on many factors :

- Complexity : A complex model can find more interesting patterns in the data, but at the same time, it will be harder to maintain and explain.
- Dataset size : Going beyond the amount of available data, a related consideration is how much of it you truly need to achieve good results.
- Dimensionality : It's useful to look at dimensionality in two different ways: the vertical size of a dataset represents the amount of data we have. The horizontal size represents the number of features.

We decided to use a Neural Network model, as its hidden layers, and connections give the most accurate results, given the number of classes and the complexity of the data and it is really good at processing and synthesizing tons of data.

Results

We construct a neural network , we trained it and then test it . We obtain an occuracy of 0.52 for training and testing

```
[ ] model.summary()
```

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_15 (Dense)            (None, 500)               5000

 dense_16 (Dense)            (None, 500)               250500

 dense_17 (Dense)            (None, 200)               100200

 dense_18 (Dense)            (None, 64)                12864

 dense_19 (Dense)            (None, 6)                 390

=================================================================
Total params: 368,954
Trainable params: 368,954
Non-trainable params: 0
_________________________________________________________________
```

Figure 7: model structure

```
[ ] # Evaluate model on test data
    scores = model.evaluate(X_test, y_test)
    print("\n%s: %.14f%%" % (model.metrics_names[1], scores[1]*100))
```

```
1687/1687 [==============================] - 4s 3ms/step - loss: 1.0790 - accuracy: 0.5175

accuracy: 51.74669623374939%
```

Figure 8: Training accuracy

```
Epoch 49/50
844/844 [==============================] - 3s 3ms/step - loss: 1.0799 - accuracy: 0.5190
Epoch 50/50
844/844 [==============================] - 3s 3ms/step - loss: 1.0796 - accuracy: 0.5194
<keras.callbacks.History at 0x7fa52d6f3b20>
```

Figure 9: Testing accuracy

I n order to visualize the final results of our research, we have created a graphical user interface in the form of a a web application using the Flask framework and web mapping techniques. In this interface, we have drawn the map of New York and given the user the possibility to choose a location in the map for which he intends to visit, enter his data such as his age, gender, race and the time of day he will visit the site, then as a result, we displayed the most likely type of crime that will be committed against him. In the server side

of our application, we have integrated our machine learning model in h5 format and we have created an API that takes the data provided by the user, apply the necessary transformations on it, predict the type of crime using our model then it returns the result to the user

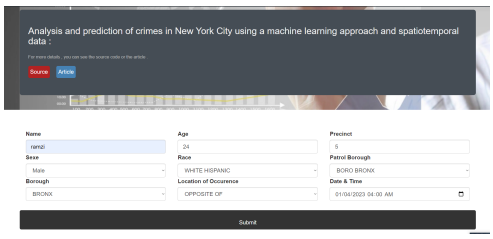


Figure 10: Filling out the crime prediction form

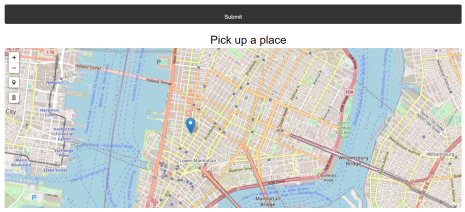
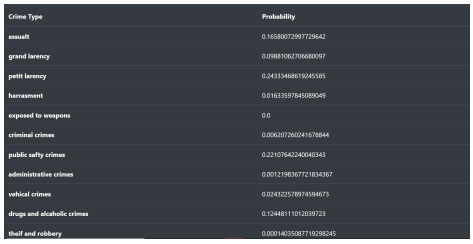


Figure 11: Selecting user location



Crime Type	Probability
assualt	0.16580072997729642
grand larency	0.09881062706680097
petit larency	0.24333468619245585
harrasment	0.016335978450890049
exposed to weapons	0.0
criminal crimes	0.006207260241678644
public safty crimes	0.22107642240040343
administrative crimes	0.0012198367721834367
vehical crimes	0.024322578974594673
drugs and alcaholic crimes	0.12448111012039723
theif and robbery	0.0001403508771929824

Figure 12: result

Conclusion

In this work, we used the database provided by NYPD. . First, we understood the meaning of the data then we did the cleaning of the data as well as the coding so that they are ready to be used by machine learning algorithms. we chose only the most relevant columns as input to our model. Then, we applied a neural network model to the data in order to predict the types of crimes that a person can suffer according to data provided by the user. Finally, we have created a web application that helps users to enter data into the system as a result we obtain the most likely type of crime that will be committed against him. In the next works, we aim to schematize the densities of crimes in a given area and also improve our model by adding some data that may be significant.

.

References

[1]https://dl.acm.org/doi/fullHtml/10.1145/3325112.3328221
[2]https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9383227
[3]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8529125/