Predicting Manhattan Crime Using Naive Bayes Classifier

Refi Ghozi

October 2020

# Introduction

A crime can happen anytime and anywhere without any warning. Sometimes, it can happen at the same place and at the same recurring time. I think it would be great if we can predict where and when the crime is going to be and we can prevent it before it happened. By using artificial intelligence and machine learning from historical data, we could build a classifier to make a prediction out of it. So in this project I will try to train a model using naive bayes classification algorithm to predict a crime that is going to happen around Manhattan on January 2020. Naive bayes is a classification algorithm that using a concept of likelihood a posterior event may happened after a prior event. Hopefully, someone will be interested on this matter of research and conduct any further research using more powerful algorithm and state of the art method. It will be very helpful for the law enforcer to handle the crime in the area.

# Data Description

* nyu-geojson data to retrieve the neighborhood latitude and longitude [1]
* Foursquare API to retrieve the venues around the neighborhood [2]
* NYPD complaint data on year range from 2015 – 2019 [3]

# Methodology

**Venues data**

First I load the nyu-geojson data to get neighborhood’s longitude and latitude. I only use the Manhattan borough for this project.

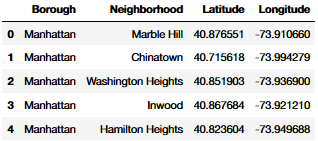


Figure . nyu-geojson dataframe

Then, I’m using folium to visualize map of Manhattan using the latitude longitude coordinates from that geojson data before.

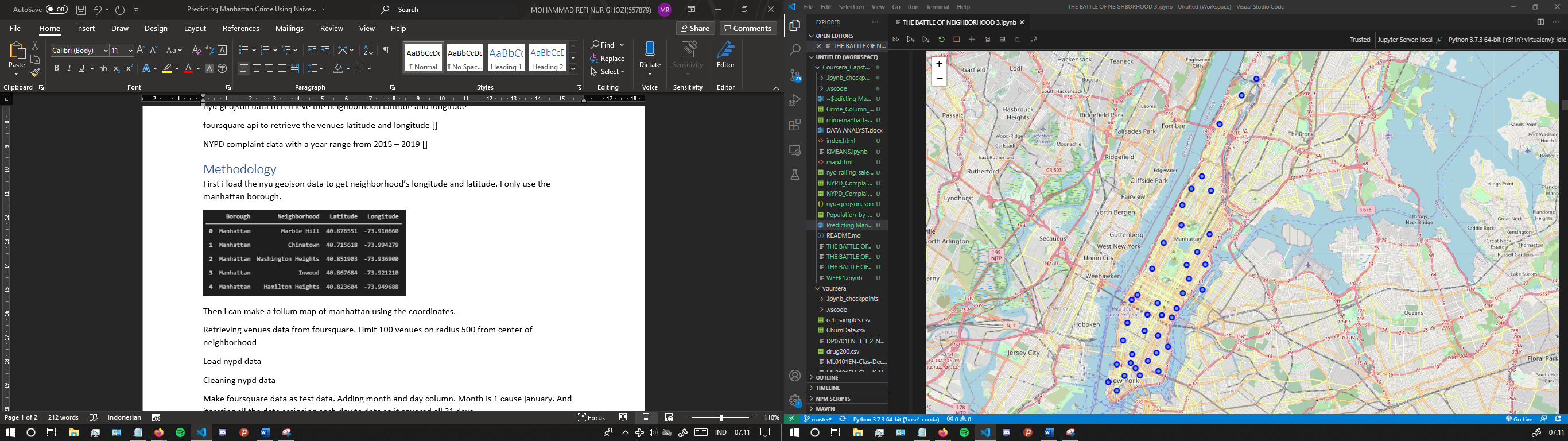


Figure . Folium map of Manhattan Neighborhood

After that, I’m retrieving venues data using foursquare API. I’m using parameter limit 100 venues on radius 500 from center of neighborhood’s longitude and latitude. Then, I merge both manhattan neighborhood dataframe and venues dataframe, the table can be seen on the image below.

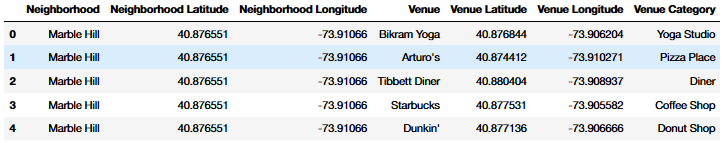


Figure . Merged venues dataframe

This dataframe is going to be the data that we are going to predict. The next I’m gonna do is adding month and day column. I put 1 for all rows on month column which means January and using iteration assigning 31 day to all the data so now the size of the data is 31 times than before.

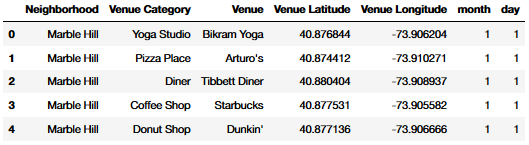


Figure . Final venues dataframe

**NYPD data**

The NYPD data contains a lot of attributes so I simplify it by selecting only the attributes that i thought are useful.

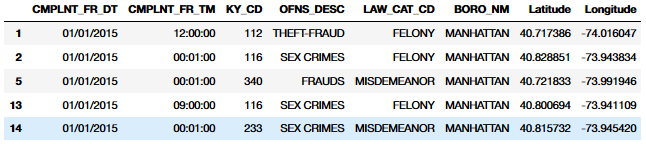


Figure . NYPD complaint data

CMPLNT\_FR\_DT means for date the crime happened and CMPLNT\_FR\_TM implies for the time the crime happened. KY\_CD is a code for crime classification, which is we’re gonna use it as a target label for classification. OFNS\_DESC is description of KY\_CD. LAW\_CAT\_CD is level of crime offense, it consists of felony, misdemeanor, and violation. BORO\_NM is borough name which in this case is only Manhattan. And the last one is latitude and longitude.

From this data, majority of the crime that reported are under misdemeanor category, then felony and violation on the 2nd and 3rd highest respectively.

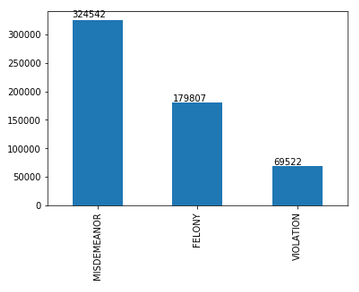


Figure . Bar plot number of reported crime by category

As we know, the target label which is KY\_CD is having more than 2 class, so this is going to be a multiclass classification task. But, i’m gonna simplify it because the number of class or KY\_CD unique value is 67 number in total. That is quite a lot. That’s why I decide to choose classes from just felony category, decreasing number of class into 23 class in total now.

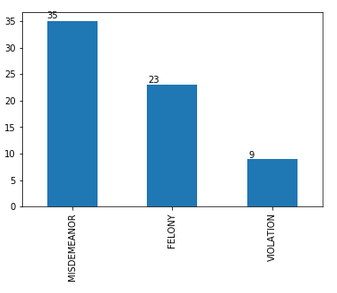


Figure . Bar plot number of crime classification by category

Because we only use the felony category, I’m gonna change the label class of both misdemeanor and violation category into 0. We’re gonna use this as the negative class, so its either felony crime or not.

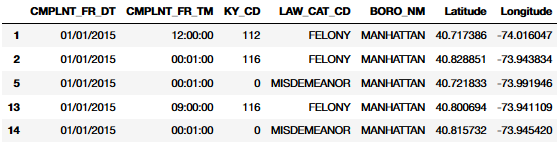


Figure . Processed NYPD dataframe with only felony category crime class

After that, we’re going to extract the date format to get month and day information. And finallly, the data is ready to be processed into the next step with the column month, day, latitude, longitude, and target label KY\_CD.

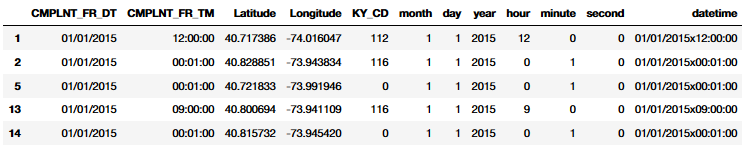


Figure . Datetime processed NYPD dataframe

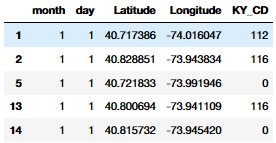


Figure . Final predict data

Because we only use the felony category and we changed other category label into 0 before, now we got an imbalanced data. The felony category is 179807 number in total, and the rest is 394064. So we try to balance it by downsampling the majority data, decreasing it into the same amount of the felony category.

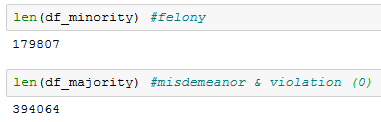


Figure . The processed crime class leaving just felony category made the data imbalanced

And finally we’re going to train the model, but first, i split the data into train data and test data with ratio train/test is 70%/30%. Then we fit the gaussian naive bayes model and analyze the performance of the trained model.

Results

After training model and then testing the model using test data, the model yield the accuracy score of 0.49 and the f1 score 0.33.

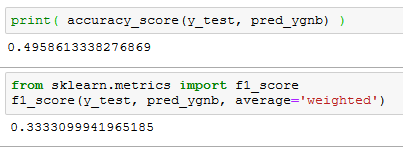


Figure . The evaluation metrics result of model

After that, I’m making the prediction using the predict data, and it produces a result predicted 225 total crime of INTOXICATED/IMPAIRED DRIVING is going to happen on January 5th around 6 distinct neighborhood.

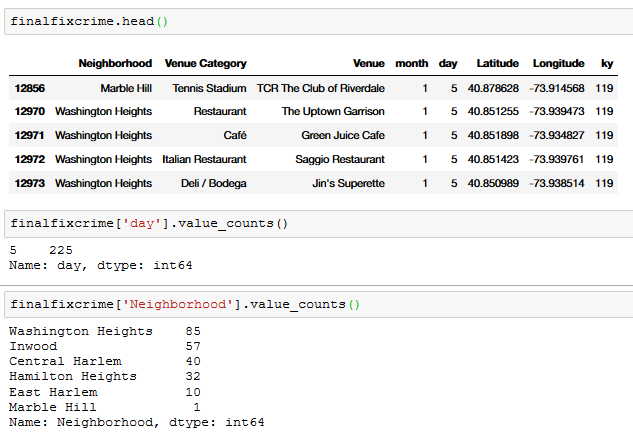


Figure . Prediction dataframe of a crime in Manhattan that is going to happen on January 2020

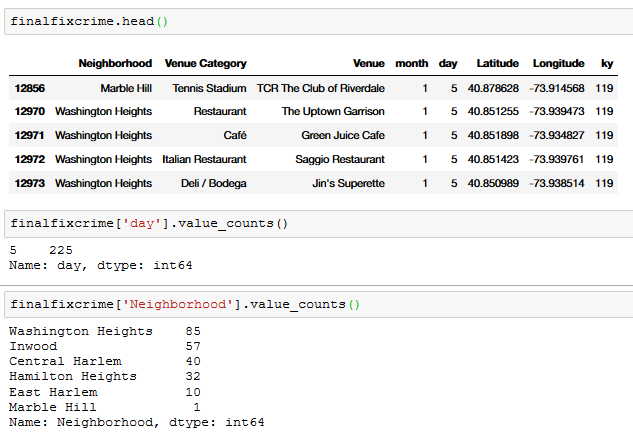
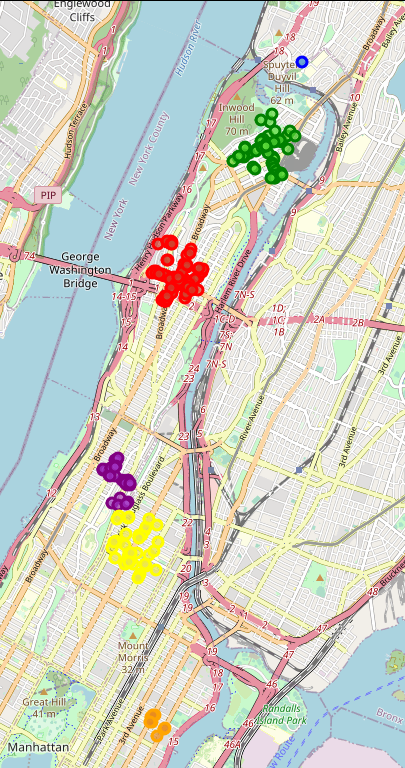


Figure . Six neighborhood that are going to be a location of a predicted crime



Blue : Marble Hill

Green : Inwood

Red : Washington Heights

Purple : Hamilton Heights

Yellow : East Harlem

Orange : Central Harlem

Figure . A folium map of predicted crime

# Discussion

The naive bayes classifier algorithm can somewhat predict the crime in manhattan with accuracy about 50% in this project. The prediction made from the model is there’s gonna be 225 intoxicated/impaired driving crime happening on january 5th around 6 neighborhood as depicted on a map in figure 15. However, in the real situation we can’t one hundred percent trust this prediction. Besides from the accuracy that aren’t very high, even if the model generated prediction with 100% accuracy, there are a lot of factor that can that directly or indirectly indicate the happening of a crime in the first place. So, the furthest thing this project is trying to do is only implementing the classifier and look how the model works by predicting the venue.

50% accuracy means the model is still can be improved. We can improve the model performance in the technical side, like adjusting the parameter of algorithm or using other resampling method for the data. We can add the venues data retrieved from the foursquare api by increasing the limit or radius amount. Also, maybe we need to add more data source to add features that are affecting the happening of a crime, like I said earlier that a lot of other factor can affects a crime action, not only just time and location.

We can also try to use other algorithm aside from naive bayes, maybe we can try to use logistic regression, support vector machine, neural network or any other classifier algorithm and see how they performs a prediction.

# Conclusion

This project is able to make a prediction of crimes that is going to happen in manhattan by implementing naive bayes classifier algorithm. Even though, there are many flaws and still need a lot of improvement but hopefully this project can brings insight for the reader and any further project in the future.

References

[1] https://geo.nyu.edu/catalog/nyu\_2451\_34572

[2] https://developer.foursquare.com/

[3] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i