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.

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Figure 1. nyu-geojson dataframe

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

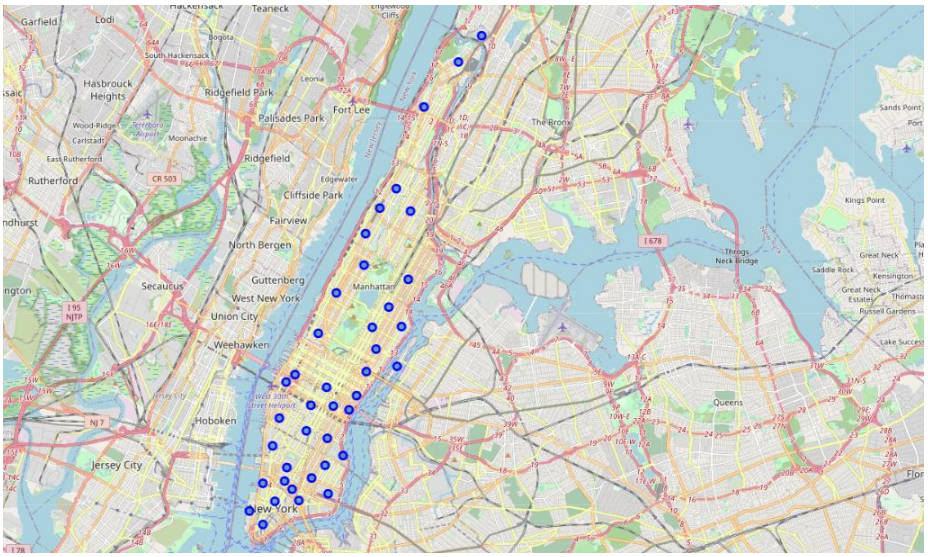

Figure 2. Folium map of Manhattan Neighborhood

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.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

Figure 3. 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.

	Neighborhood	Venue Category	Venue	Venue Latitude	Venue Longitude	month	day
0	Marble Hill	Yoga Studio	Bikram Yoga	40.876844	-73.906204	1	1
1	Marble Hill	Pizza Place	Arturo's	40.874412	-73.910271	1	1
2	Marble Hill	Diner	Tibbett Diner	40.880404	-73.908937	1	1
3	Marble Hill	Coffee Shop	Starbucks	40.877531	-73.905582	1	1
4	Marble Hill	Donut Shop	Dunkin'	40.877136	-73.906666	1	1

Figure 4. 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.

	CMPLNT_FR_DT	CMPLNT_FR_TM	KY_CD	OFNS_DESC	LAW_CAT_CD	BORO_NM	Latitude	Longitude
1	01/01/2015	12:00:00	112	THEFT-FRAUD	FELONY	MANHATTAN	40.717386	-74.016047
2	01/01/2015	00:01:00	116	SEX CRIMES	FELONY	MANHATTAN	40.828851	-73.943834
5	01/01/2015	00:01:00	340	FRAUDS	MISDEMEANOR	MANHATTAN	40.721833	-73.991946
13	01/01/2015	09:00:00	116	SEX CRIMES	FELONY	MANHATTAN	40.800694	-73.941109
14	01/01/2015	00:01:00	233	SEX CRIMES	MISDEMEANOR	MANHATTAN	40.815732	-73.945420

Figure 5. 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.

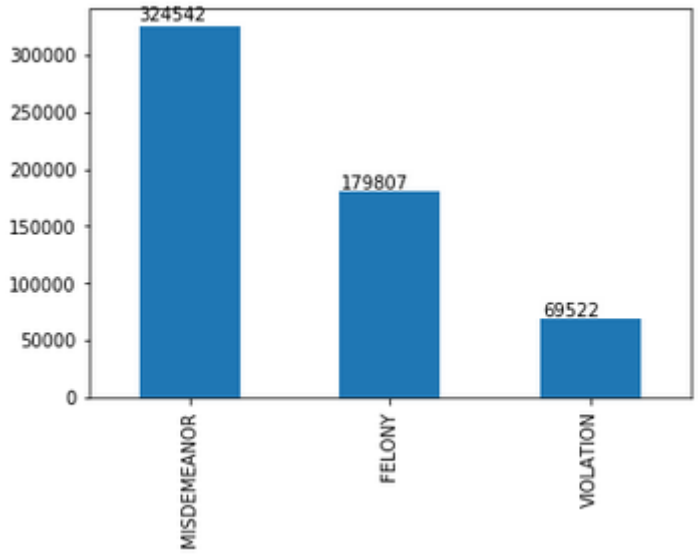

Figure 6. Bar plot number of reported crime by category

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.

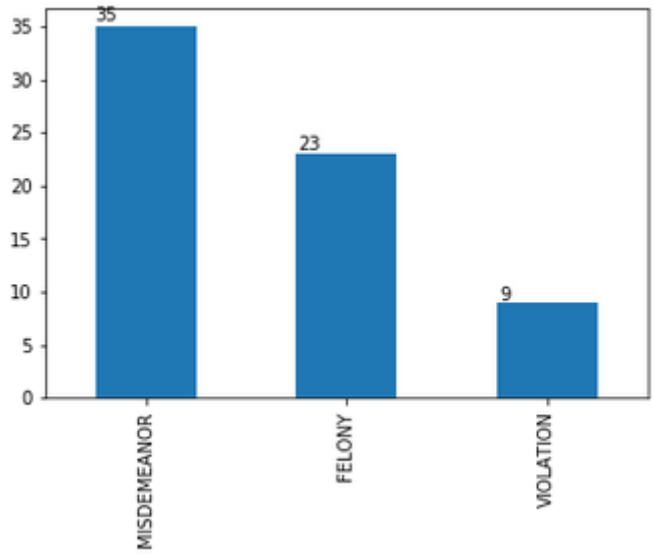

Figure 7. Bar plot number of crime classification by category

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.

	CMPLNT_FR_DT	CMPLNT_FR_TM	KY_CD	LAW_CAT_CD	BORO_NM	Latitude	Longitude
1	01/01/2015	12:00:00	112	FELONY	MANHATTAN	40.717386	-74.016047
2	01/01/2015	00:01:00	116	FELONY	MANHATTAN	40.828851	-73.943834
5	01/01/2015	00:01:00	0	MISDEMEANOR	MANHATTAN	40.721833	-73.991946
13	01/01/2015	09:00:00	116	FELONY	MANHATTAN	40.800694	-73.941109
14	01/01/2015	00:01:00	0	MISDEMEANOR	MANHATTAN	40.815732	-73.945420

Figure 8. 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.

	CMPLNT_FR_DT	CMPLNT_FR_TM	Latitude	Longitude	KY_CD	month	day	year	hour	minute	second	datetime
1	01/01/2015	12:00:00	40.717386	-74.016047	112	1	1	2015	12	0	0	01/01/2015x12:00:00
2	01/01/2015	00:01:00	40.828851	-73.943834	116	1	1	2015	0	1	0	01/01/2015x00:01:00
5	01/01/2015	00:01:00	40.721833	-73.991946	0	1	1	2015	0	1	0	01/01/2015x00:01:00
13	01/01/2015	09:00:00	40.800694	-73.941109	116	1	1	2015	9	0	0	01/01/2015x09:00:00
14	01/01/2015	00:01:00	40.815732	-73.945420	0	1	1	2015	0	1	0	01/01/2015x00:01:00

Figure 9. Datetime processed NYPD dataframe

	month	day	Latitude	Longitude	KY_CD
1	1	1	40.717386	-74.016047	112
2	1	1	40.828851	-73.943834	116
5	1	1	40.721833	-73.991946	0
13	1	1	40.800694	-73.941109	116
14	1	1	40.815732	-73.945420	0

Figure 10. 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.

```
len(df_minority) #felony
```

```
179807
```

```
len(df_majority) #misdemeanor & violation (0)
```

```
394064
```

Figure 11. 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.

```
print( accuracy_score(y_test, pred_ygnb) )
```

```
0.4958613338276869
```

```
from sklearn.metrics import f1_score
f1_score(y_test, pred_ygnb, average='weighted')
```

```
0.3333099941965185
```

Figure 12. 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.

	Neighborhood	Venue Category	Venue	month	day	Latitude	Longitude	ky
12856	Marble Hill	Tennis Stadium	TCR The Club of Riverdale	1	5	40.878628	-73.914568	119
12970	Washington Heights	Restaurant	The Uptown Garrison	1	5	40.851255	-73.939473	119
12971	Washington Heights	Café	Green Juice Cafe	1	5	40.851898	-73.934827	119
12972	Washington Heights	Italian Restaurant	Saggio Restaurant	1	5	40.851423	-73.939761	119
12973	Washington Heights	Deli / Bodega	Jin's Superette	1	5	40.850989	-73.938514	119

Figure 13. Prediction dataframe of a crime in Manhattan that is going to happen on January 2020

```
Washington Heights    85
Inwood                57
Central Harlem        40
Hamilton Heights      32
East Harlem           10
Marble Hill            1
```

Figure 14. Six neighborhood that are going to be a location of a predicted crime

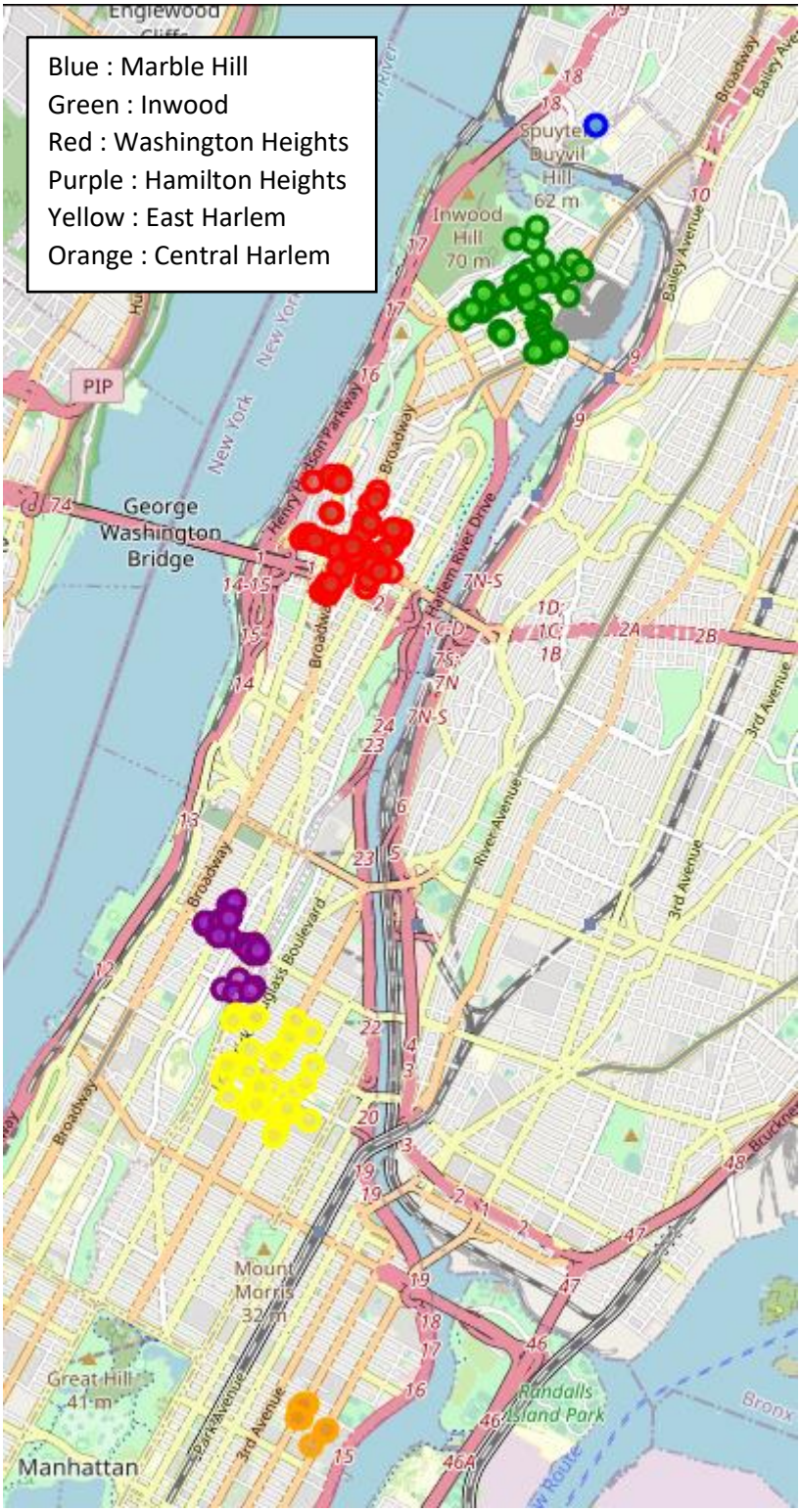

Figure 15. A folium map of predicted crime

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