Predicting High Risk Crime Areas in Chicago

Ibnu Santoso
May 5, 2019

1. Introduction

1.1. Background

Chicago is the most populous city in Illinois, as well as the third most populous city in the United States. As common as other populous cities, Chicago has crime issues as well. As each area in Chicago has different crime rate in certain time, police department needs to allocate different number of resources in certain area and time. Therefore, it is advantageous for police department to accurately predict which area and venue types has high risk crime in certain time. For example, this informations can be used by police department to allocate more resources or spend more time in the particular area.

1.2. Problem

Data that might contribute to determining high risk crime areas might include history of incidents in certain day, time, police district, and community area. This project aims to predict number of crimes in certain community area in certain time based on these data. Nearby venues when the incident happened before would also be used to predict venue types where the crime may happen, such as park, sports bar, cafe, etc.

1.3. Interest

Police department would be very interested in accurate prediction of high risk crime rate areas in certain time, for helping in planning. Others community organization who works in lowering crime rate may also be interested.

2. Data acquisition and cleaning

2.1. Data Sources

Chicago crime incidents dataset from 2001 to present can be found Chicago Data Portal. This dataset has details information on incidents, such as incident coordinates (latitude and longitude), time, community area where it happened, crime type, etc. I also used Foursquare Explore API to get nearby venues based on incident coordinates.

2.2. Data Cleaning

The crime data from Chicago Data Portal has huge number of records more than 6 millions records. Therefore, I decided to slice the data and only used the last 2 years of data (January 2017 - April 2019).

There are some incident records which do not have latitude and longitude values, for those records I decided to drop those records, because the number is not significant (~6000 records) compared to 2 years dataset (~600,000 records). We need latitude and longitude values to get nearby venues using Foursquare API and get the most common venue categories applied in the model.

There is a daily limit for calling Foursquare API to get incident nearby venues, in this project I only used Foursquare API for populating incidents nearby (within radius 100 metres) venues for the last 50 incidents per community area which has highest predicted number of incidents in particular district..

2.3. Feature Selection

After data cleaning, there are ~596,000 incident records and 21 features. As the goal of this project is to predict number of incident of each community area in certain time and to get common nearby venues where incident took places, I chose some of features which useful for the prediction : *case_number, date, district, community_area, latitude, and longitude.*

For the incident nearby venues records from Foursquare API I only kept *case_number* and *Venue Category.* Case number will be used to join with the incident data, and I am only interested on venue category, instead of the specific venue name, e.g. park, instead of Magnolia Park. The reason is some incidents might have similarity happened in the park but it happened in different location.

3. Exploratory Data Analysis

3.1. Calculation of Target Variable

Number of incidents in certain time was not a feature in dataset. Hence, I did a calculation by grouping the data by some features (date, district, community_area, day_of_the_week, hour_category) and got the count. The following sections will explain about the features selection that used to group the data.

3.2. District Feature

Police departments in Chicago consists of 25 police districts. It would be beneficial if the prediction can group the high risk crime community areas per district. Totally there are 25 districts in dataset, hence the dataset has data for all districts.

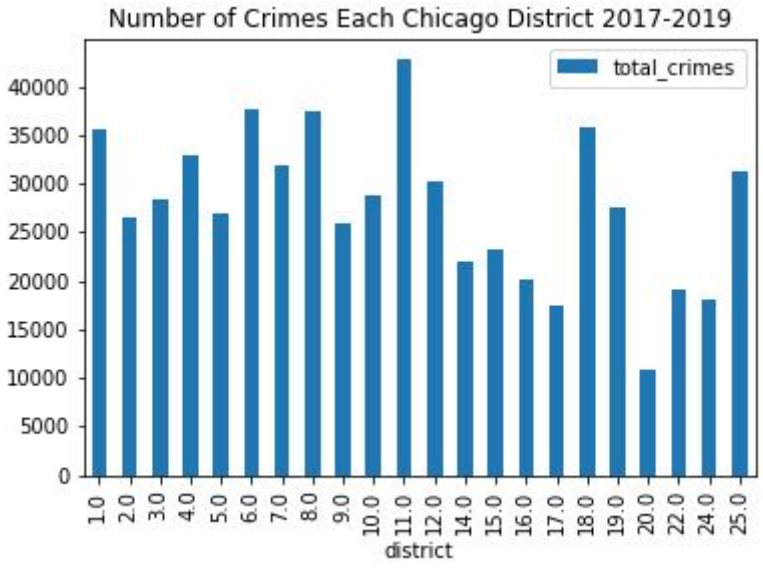

Figure 1. Number of Crimes Each District (January 2017-April 2019)

3.3. Community Area Feature

Community area has smaller size than district, hence it is a good feature for the prediction, so we can have more precise prediction. Totally there are 78 community areas in dataset. To take note, one community area may belong to multiple police districts.

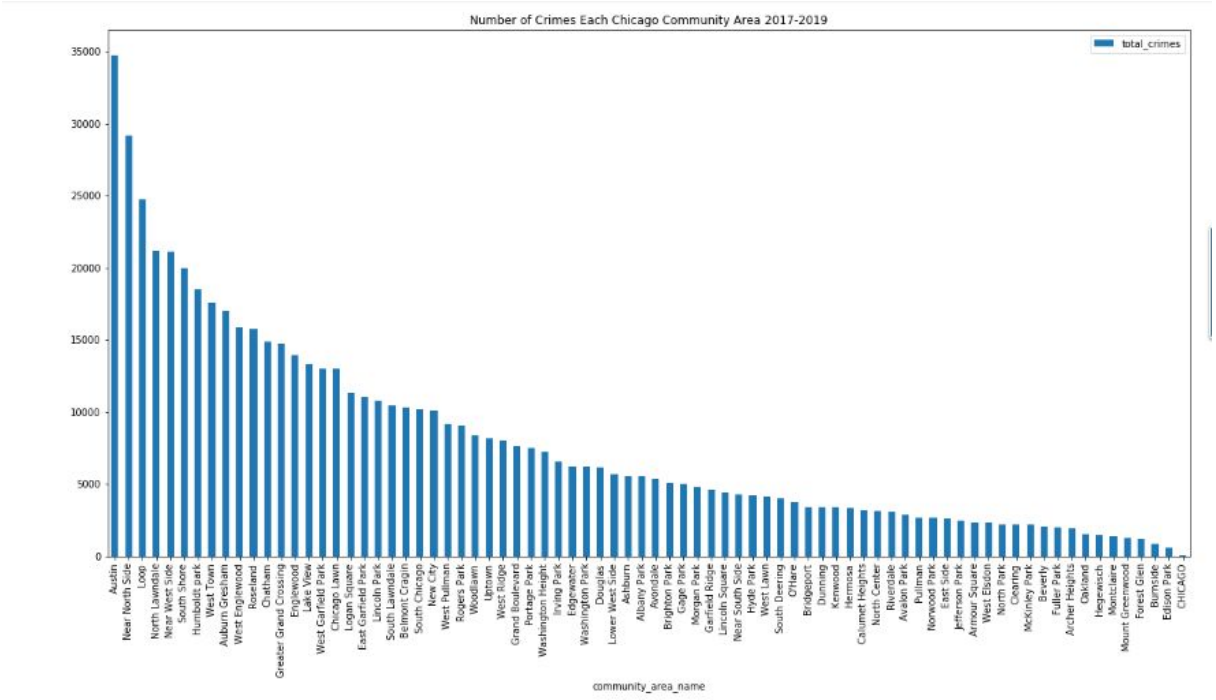

Figure 2. Number of Crimes in Each Community Area (January 2017-April 2019)

3.4. Day of Week Feature

I hypothesized that day of week may determine the number of crimes. Example: on Friday evening people usually go out for drink with friends, hence it will cause crowds in public area and it may increase the risk of crime rate. Compared to Sunday evening when people usually prepares for first day of work (Monday) after weekend, and just stay at home, then it should have lower crime rate. The following the number of crime incidents for each day of week.

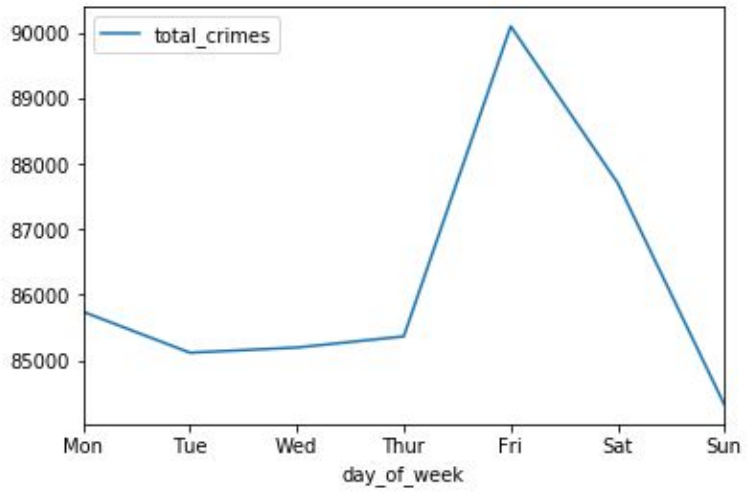

Figure 3. Number of Incidents for Different Days in a Week (January 2017 - April 2019)

The dataset does not have day of week feature, hence I had to extract the day of week from the incident date. I used pandas library to get the day of week from the incident date and put the value into a new column in pandas dataframe.

3.5. Hour Category Feature

I hypothesized that different hour may have different number of crimes. Example: at night while most of people are sleeping, the number of crimes should be less than in the afternoon while most people are active. The following the number of crime incidents for different hours within a day. Based on the figure 4 the peak was at 12 PM and it dropped at 5 AM.

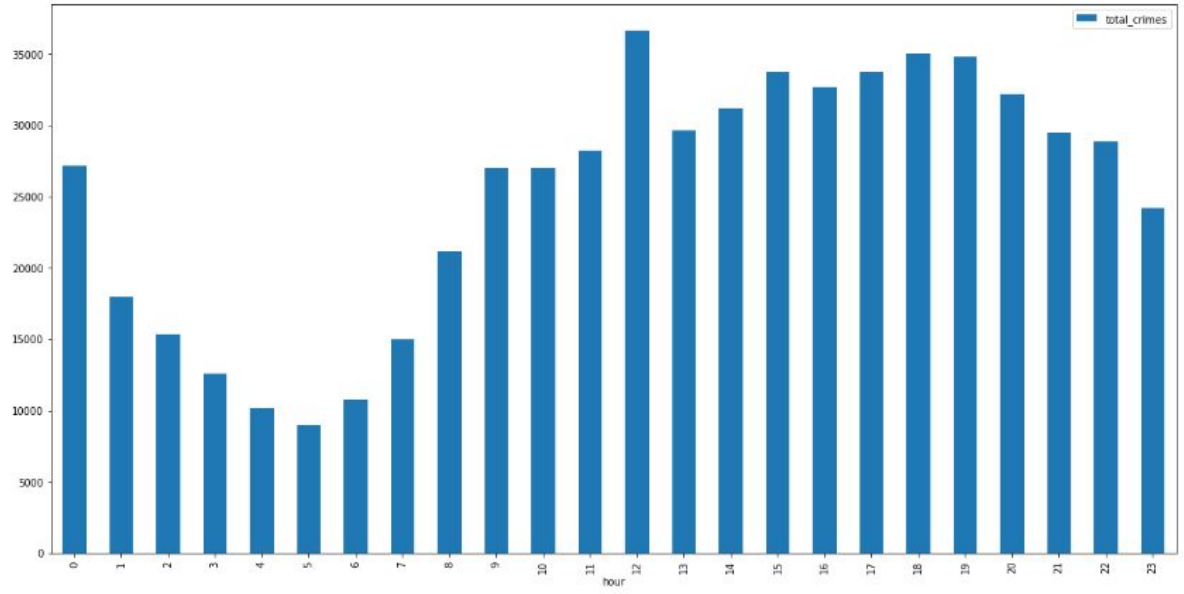

Figure 4. Number of Crimes for Different Hours within a Day

As this project's objective is to predict the number of incidents for each community area in a district within a day, hour feature may produce very low number of crime incidents. Hence I groped the hours into 6 hourly: 0-5, 6-11, 12-17, and 18-23 (4 categories per day). Here is the number of crimes for different hour category. The peak is between 12 to 17.

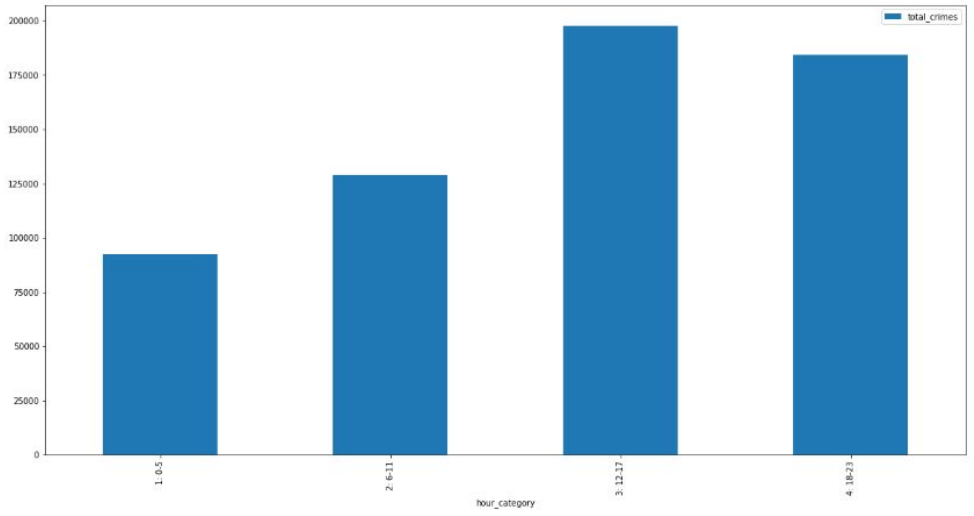

Figure 5. Number of Crimes for Different Hour Category within a Day

4. Predictive Modelling

I used regression models for this project because the objective is to predict number of incidents in each community area of district in certain time. I applied some regression models (Linear SVR, KNN Regressor, SGD Regressor, and Elastic Net) and calculated the Root Mean Squared Error (RMSE) for each model.

KNN regressor had the best performance (lowest RMSE) among other models. The downside of KNN regressor is its speed performance which is the slowest among other models which I used. SGD Regressor has the second best performance and its speed is faster than KNN regressor. Hence I did choose SGD regressor for my model.

Linear SVR	KNN Regressor (5 neighbors)	SGD Regressor	Elastic Net
1.83	1.69	1.77	2.31

Table 1. RMSE for Each Model

5. Nearby Incident Venues Retrieval Using Foursquare API

I used Foursquare API to get nearby venues (radius 100 metres) of last 50 incidents for each highest crime rate community areas in each district.

	Case Number	Incident Latitude	Incident Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	JB574395	41.892396	-87.70768	GameStop	41.896158	-87.707951	Video Game Store
1	JB574395	41.892396	-87.70768	Wingstop	41.896637	-87.706980	Wings Joint
2	JB574395	41.892396	-87.70768	Baskin-Robbins	41.896095	-87.707916	Ice Cream Shop
3	JB574395	41.892396	-87.70768	Subway	41.895667	-87.707406	Sandwich Place
4	JB574395	41.892396	-87.70768	ALDI	41.896815	-87.708491	Grocery Store

Figure 6. Nearby Venues within 100 Metres for Each Incidents

The venue categories were converted to one hot encoding to be used to find the frequency for each category.

	Case Number	ATM	African Restaurant	American Restaurant	Art Gallery	Automotive Shop	BBQ Joint	Bakery	Bank	Bar	...	Train Station	Vacation Rental	Video Game Store	Video Store	Warehouse	Warehouse Store	Wine Bar	Wings Joint	Women's Store	Y Stu
0	JB574395	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	
1	JB574395	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	
2	JB574395	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
3	JB574395	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	
4	JB574395	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	

Figure 7. One Hot Encoding for Venue Categories

6. Model Application

The model uses day of the week (Mon - Fri), hour category, district number, and community area to predict number of incidents. In order to make it more applicable I created a python function which receives 2 parameters only : date and hour category. In the function the day of week is extracted from the date input and iteratively predicts all available district numbers and community areas. For each highest community area in a district, it calls Foursquare API to get last 50 incidents in that community area and get the top 5 most common nearby venue categories..

```
[ ]: # Hour Category options : 0-5, 6-11, 12-17, 18-23

     # Friday evening
     predict('2019-05-10', '18-23')
```

Figure 8. The Model Calling by Passing Date and Hour Category

Here is the sample of prediction output for District 1 to 5 only. For complete prediction result can be found in here.

```
Prediction of incidents number for each community area, date: 2019-05-10 , hour: 18-23
District:  1
---Community Area: Armour Square, predicted number of incident : 2.3235137810103677
---Community Area: Douglas, predicted number of incident : 2.5395488840941756
---Community Area: East Garfield Park, predicted number of incident : 0
---Community Area: Loop, predicted number of incident : 8.855660551750738
---Community Area: Near South Side, predicted number of incident : 3.156256121354368
---Community Area: Near West Side, predicted number of incident : 3.710297850544273
Predicted community area with highest incident number : Loop, with number of incidents: 9
Top 5 common nearby venues (radius 100 metres)
                venue  freq
0               Hotel  0.08
1         Coffee Shop  0.07
2      Sandwich Place  0.06
3  Italian Restaurant  0.03
4  Seafood Restaurant  0.03
```

```
District:  2
---Community Area: Armour Square, predicted number of incident : 1.0
---Community Area: Douglas, predicted number of incident : 2.5395488840941756
---Community Area: Englewood, predicted number of incident : 3.758803833759994
---Community Area: Fuller Park, predicted number of incident : 2.0397036960388135
---Community Area: Grand Boulevard, predicted number of incident : 3.1977206675991323
---Community Area: Greater Grand Crossing, predicted number of incident : 1.0
---Community Area: Hyde Park, predicted number of incident : 2.345991274454128
---Community Area: Kenwood, predicted number of incident : 2.233168247631974
---Community Area: Logan Square, predicted number of incident : 0
---Community Area: Oakland, predicted number of incident : 1.8432874641351147
---Community Area: Riverdale, predicted number of incident : 0
---Community Area: Washington Park, predicted number of incident : 2.189733421845695
---Community Area: Woodlawn, predicted number of incident : 2.947916725402425
Predicted community area with highest incident number : Englewood, with number of incidents: 4
Top 5 common nearby venues (radius 100 metres)
                  venue  freq
0           Water Park  0.43
1             Bus Line  0.14
2          Bus Station  0.14
3  Fast Food Restaurant  0.14
4    Light Rail Station  0.14
```

```
District:  3
---Community Area: Englewood, predicted number of incident : 3.758803833759994
---Community Area: Greater Grand Crossing, predicted number of incident : 2.2231948151217336
---Community Area: Hyde Park, predicted number of incident : 2.345991274454128
---Community Area: South Shore, predicted number of incident : 3.774397683246899
---Community Area: Washington Park, predicted number of incident : 2.189733421845695
---Community Area: Woodlawn, predicted number of incident : 2.947916725402425
Predicted community area with highest incident number : South Shore, with number of incidents: 4
empty !
Top 5 common nearby venues (radius 100 metres)
                  venue  freq
0     Currency Exchange  0.09
1  Fast Food Restaurant  0.07
2          Bus Station  0.07
3   Fried Chicken Joint  0.07
4        Sandwich Place  0.07
```

```
District:  4
---Community Area: Avalon Park, predicted number of incident : 2.459281603727085
---Community Area: Burnside, predicted number of incident : 2.135502397363075
---Community Area: Calumet Heights, predicted number of incident : 2.565590479976358
---Community Area: Chatham, predicted number of incident : 1.0
---Community Area: East Side, predicted number of incident : 2.4430102365612503
---Community Area: Greater Grand Crossing, predicted number of incident : 1.0
---Community Area: Hegewisch, predicted number of incident : 2.270134973624275
---Community Area: South Chicago, predicted number of incident : 4.205115213358251
---Community Area: South Deering, predicted number of incident : 2.3763445500418365
---Community Area: South Shore, predicted number of incident : 3.774397683246899
Predicted community area with highest incident number : South Chicago, with number of incidents: 4
Top 5 common nearby venues (radius 100 metres)
                  venue  freq
0   American Restaurant  0.08
1  Fast Food Restaurant  0.08
2           Pizza Place  0.08
3                Lounge  0.08
4          Liquor Store  0.05
```

Figure 9. Model Output Sample

7. Conclusion

In this study, I analyzed the relationship between day of the week, hour category (0-5, 6-11, 12-17, 18-23), and number of incidents. It shows that day of the week and hour category affect number of incidents, hence they are good features for prediction. I built a regression model using day of the week, hour category, district number, and community area name to predict the incident number for each community area in all police districts in Chicago in certain day and time. This model can be very useful for police department in Chicago to plan resource number would be assigned in certain area and how much time to spend in specific area.

8. Future Direction

This model uses day of week and hour category to predict the incident number, which are 2 good features to use. However, month of year or week of month may also a good feature to use for prediction. There are still some features can be explored to get better prediction result. In order to explore month of year feature, it would need more data than the data which I used for this project.

To be more intuitive for the users it may be good to use a map and put a circle on each areas which have high risk crime in certain time. Besides that predicted crime types for each areas may be useful as well.

9. Python Codes

I shared my Jupyter notebook for this project in Github which can be found here: https://github.com/ibnuws/Coursera_Capstone/blob/master/Week%204/Crime-incidents-prediction.ipynb