Crime in the City (Chicago)

Usama Maabed

May 10, 2019

1. Introduction

1.1. Background

City crimes are one of the main concerns for public and authorities' decision makers. Today's life in a big cities with high population consider safety as major factor. The safety of the community creates challenges for authorities and police departments. Crimes in big cities are different with diverse rates and might happen during any time, which in its turn need proper planning and resources to understand its trend and where and when to allocate the required resources in certain areas.

Crime in Chicago has been tracked by Chicago Police Department's Bureau of Records since the beginning of the 20th century. The city's overall crime rate, especially the violent crime rate, is higher than the US average. Chicago was responsible for nearly half of 2016's increase in homicides in the US. The reasons for the higher numbers in Chicago remain unclear, which needs new approach and techniques that can help in promoting the overall police operations with focus in improving the crime investigation process.

1.2. Problem

Forecasting used to be a challenge for both the planners and decision makers. Proper planning for critical service such as police operations needs clear inputs to answer critical questions such as:

- What are the major crimes in the city and its future trend?
- How the crime types distributed across the city?
- Where are such crimes concentrated in each city's district?
- With limited budget, which districts should get more focus and financial support to improve its operations' capabilities?
- What are the main geo characteristics used to be repeated or found for a specific crime?
- Can such characteristics be used in a proactive approach to improve the policy investigation process?

This project aimed to develop a model in which the historical data utilized to provide the required answers for such concerns and put insight on where are the areas of improvements.

1.3. Interest

Government authority and police department would be very interested in accurate indicators and forecasts that promote the overall operations and empower the crime fighting process.

2. Data acquisition and cleaning

2.1.Data sources

This project depends on different information resources.

- Chicago crime incidents dataset from 2001 to present that can be found at Chicago Data Portal. This dataset has detailed information on the daily cases, such as where it happened, when, case coordinates (latitude and longitude), crime type, etc.
- Foursquare, which is a geo information platform that powers leading business solutions and consumer products through a deep understanding of location. Foursquare API used to get nearby venues based on case coordinates.

2.2. Data cleaning

During our data review we found some issues and manged as follows:

- There were some records with “nan”, so the first step was to review the missing data and decide on how to deal with it.
- We dropped the rows with “nan” at the Latitude and Longitude columns.
- We found more than 80,000 record was no policy district details, so for those with Crime case having Latitude and Longitude, we used KNN classifier to complete the missing details with accuracy 99.9%
- We found some duplicated rows, so we dropped the duplicated row as well
- We found year 2109’s data not complete then we dropped its data
- The 'Police Districts' column data type has been changed to integer.
- The “Date” column type has been changed to date in place of string
- Also, added one column “Month”, which extracted from the “Date” column.

2.3.Feature selection

After data cleaning, there were 6,705,122 samples and 18 attributes. Upon examining the meaning of each feature, it was clear that there were some redundancy in the features.

The following attributes have been selected to be utilized through our project.

['ID', 'Case Number', 'Date', 'Block', 'Primary Type', 'Description', 'Arrest', 'Community Area', 'Year', 'Latitude', 'Longitude', 'Zip Codes', 'Community Areas', 'Census Tracts', 'Wards', 'Boundaries - ZIP Codes', 'Police Districts']

Latter we added the “Month” attribute that has been extracted from “Date” column, which we use during our analysis.

3. Exploratory Data Analysis

3.1. Calculation of target variable

During our project server grouping and filtering option have been used. Also, many relation between data have been used e.g.

- Relation between the Policy District and its yearly load. The Load calculated as the total number of the registered crimes. – **Figure 1**
- Relation between the arrested vs. non-arrested cases per year. - **Figure 2**
- The total number of Crimes per year - **Figure 3**
- The top 5 Crimes Distribution per year - **Figure 4**
- The Theft Crimes distribution per month - **Figure 5**
- Total number of Theft Crimes per Police District - **Figure 6**
- The venues nearby a registered crime case - **Figure 7**

In addition, the K nearest neighbor (KNN) classifier used to complete the Police District incomplete data.

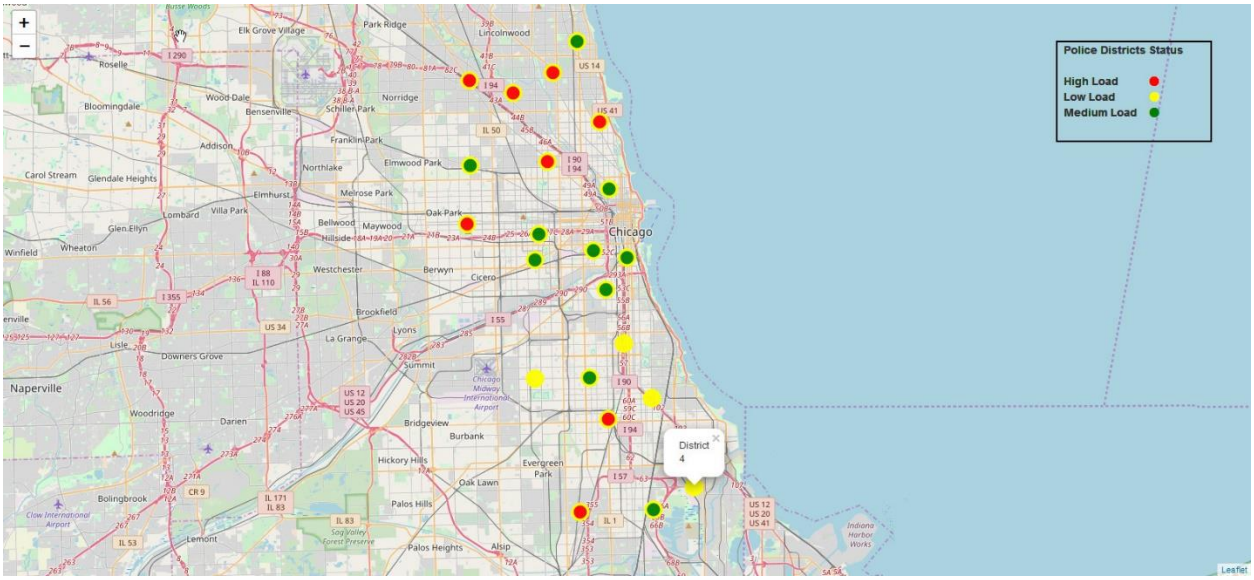

Figure 1: Categorizing the Police Districts per its load

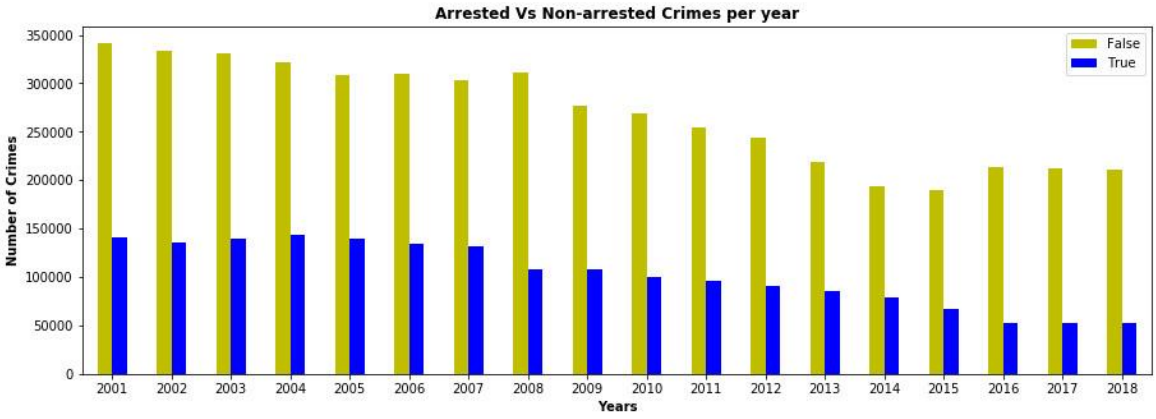

Figure 2: Arrested Vs Non-arrested Crimes per year

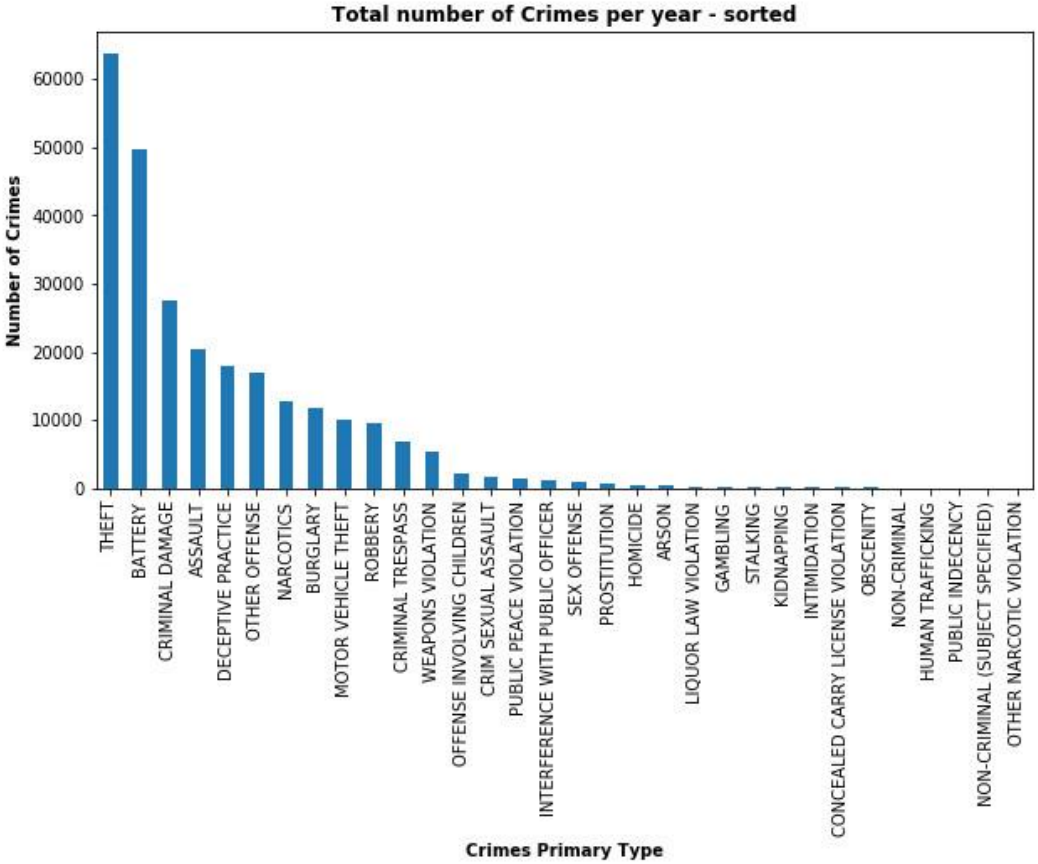

Figure 3: Total number of Crimes per year

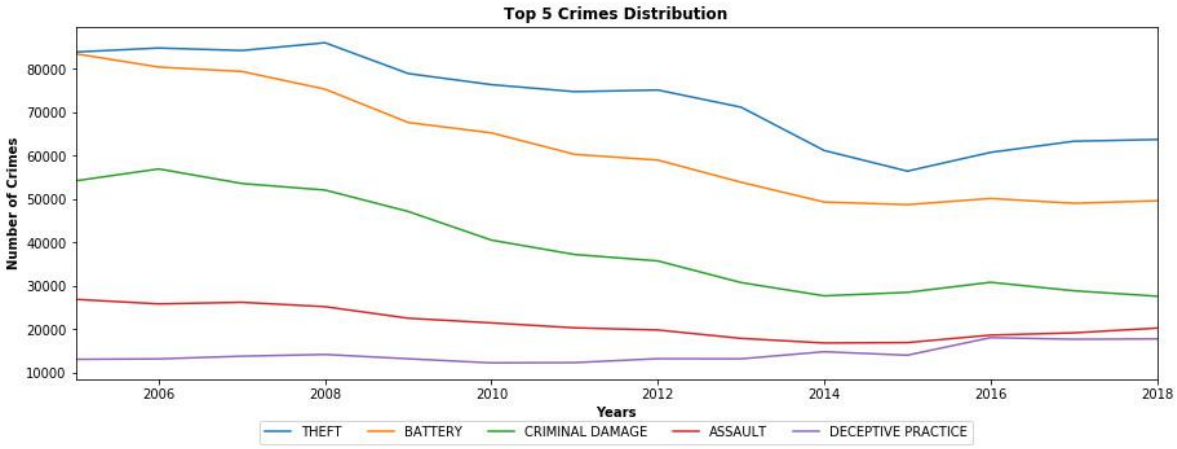

Figure 4: Top 5 Crimes Distribution

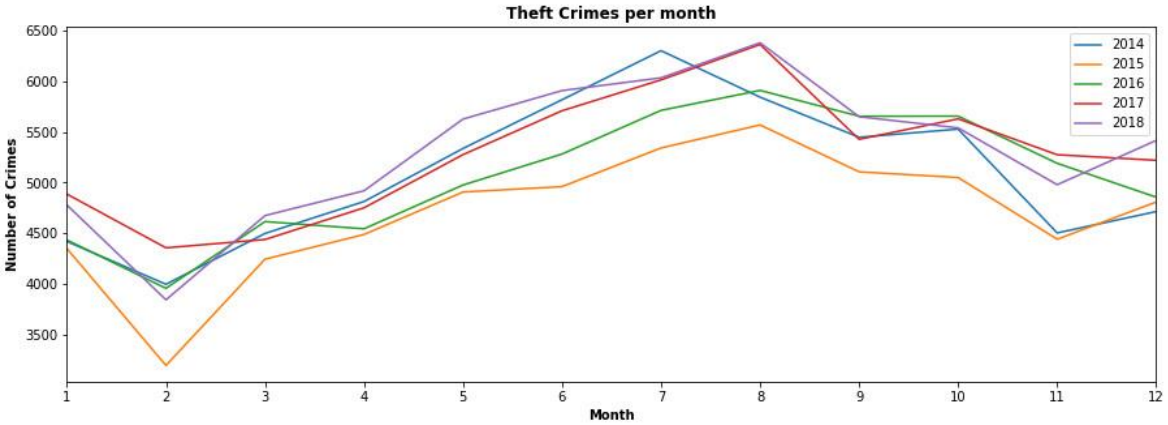

Figure 5: Theft Crimes per Month

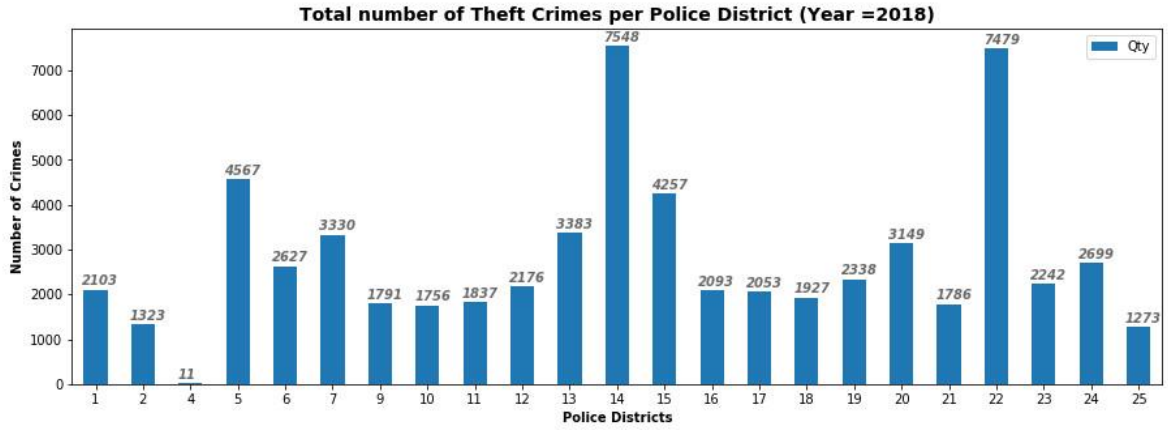

Figure 6: Total Theft Crimes per Police District

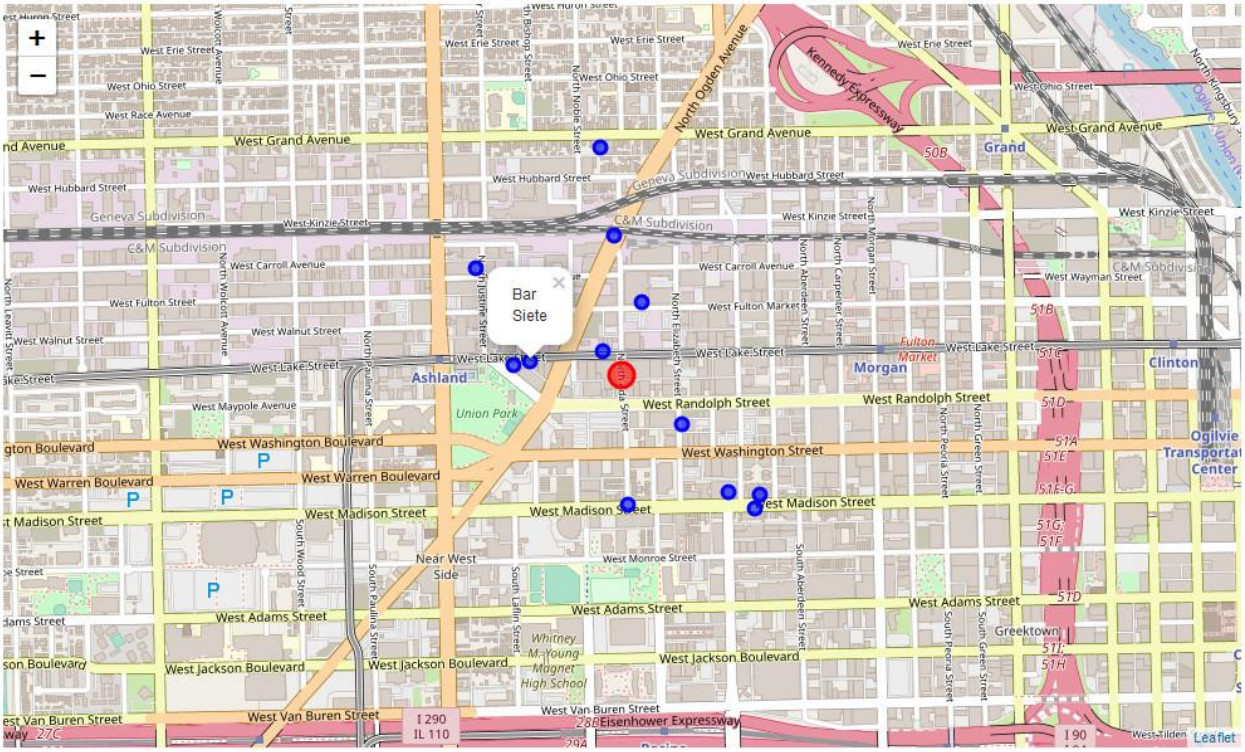

Figure 7: The Venues that are nearby to Case No. JB410653

4. Classification model as a Predictive Modeling

The application of classification models was straightforward, I used to predict the missing police district details. I divided the samples into three classes using Binning. I used KNN as my classifier and segregate the training and test sample. The Train set had 4,500,584 records, while the Test set had 1,125,146 records.

K has been estimated and found that the best accuracy was with 0.999 with k= 1, **Figure 8**

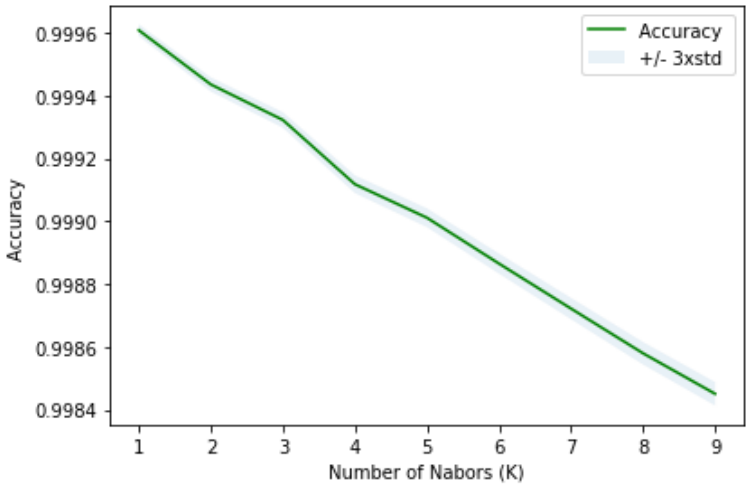

Figure 8: The best K estimation

5. Conclusions

In this study, the main objective was helping the government authorities and police department in improving the planning and the quality of the provided service. The primary objective was to understand the crimes and the police load trends and where and when it is necessary to allocate resources in certain area. The project come to provide proper answers to some challenging questions such as:

- What are the major crimes in the city and its future trend?
- How the crime types distributed across the city?
- Where are such crimes concentrated in each city's district?
- With limited budget, which districts should get more focus and financial support to improve its operations' capabilities?
- What are the main geo characteristics used to be repeated or found for a specific crime?
- Can such characteristics be used in a proactive approach to improve the policy investigation process?

The project succeeded to cover these main concerns and provide the required inputs that can help the decision makers to have good insight into what is going and the areas that need improvements. Final decision on optimal actions will be made by the stakeholders based on specific characteristics and locations, taking into consideration additional factors like budget constraints of each location, levels of load per each districts, social dynamics of every area etc.

6. Future directions

I was able to achieve the objective of this study, however, there was still some variance that could not be predicted by the models. I think the models can be enhanced by capturing other attributes and extend the use of "Foursquare" tips. For example, considering the number of police resources and current budget allocation per district. Also to cluster the crimes based on specific symptoms. Also the social life surrounding the crime location can provide inputs for better investigation process.