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, 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 resource allocation. 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 (2017 - 2019). For nearby incident venues, I created a simple python codes to call Foursquare API iteratively by passing incident coordinates (latitude and longitude).

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).

There are 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 250 metres) venues in District 11 which is the district with the largest number of crimes in Chicago.

2.3. Feature Selection

After data cleaning, there are ~596,000 incident records, ~120,000 incident nearby venues 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 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.