Project Proposal

NYC Crime Type Prediction

Background:

When 911 is called, to what extent police should react depends on the type of a crime. And we think the crime type is closely related to the location and time of the crime occurrence. Thus, to predict crime type given the reported location and time, we first focus on living environment in this project. We wonder if the location description of the street appear in daily conversation can predict or evaluate crime type and rate. In this project, we will find several location features in NYC to verify this assumption. Also, we are interested in the time when crime occurs and we want to figure out whether there is any pattern with regard to time.

Potential Question(s):

(a) Is there any pattern that some types of crime happens more often than others in certain area? If so, what is the pattern?
(b) Is there anything common in terms of the house price, neighborhood restaurant type related to crime type? If so, which featues are related?
(c) Is there any temporal pattern on the crime type distribution? If so, what is the pattern?

Target Variable(s):

(a) Correlation between ranked zip codes in NYC by number of crime of each type.
(b) Seasonal components of time series in number of crime of each crime type

Data Sourced from:

- NYC Crime Data: http://www1.nyc.gov/site/nypd/stats/crime-statistics/crime-statistics-landing.page
- Craigslist:https://newyork.craigslist.org
- NYC Restaurant Inspection Data: https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/xx67-kt59
- NYC School/Library/Museum locations:
  - Colleges and Universities: https://data.cityofnewyork.us/Education/Colleges-and-Universities/4kym-4xw5
  - schools: https://data.cityofnewyork.us/Environment/schools/v9yc-xnt7
  - NYC Museums: https://data.cityofnewyork.us/Recreation/New-York-City-Museums/ekax-ky3z
  - Library: https://data.cityofnewyork.us/Business/Library/p4pf-fyc4
- 24 hours store(Target, Walmart...) locations:

- ○ Legally Operating Businesses: https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh
- ○ DCA license by zip: https://data.cityofnewyork.us/Business/DCA-license-by-zip/eugb-mqmz

Analysis Approach:

(a) Data Mining

- Utilize web crawler to extract house information like house location, house price etc.
- Convert latitude and longitude locations into zip code

(b) Data Cleaning

- Combine several dataset we found based on features we want to generate our dataset for further working. Because we will use several dataset from different sources, the shape of each dataset is incompatible. Thus, we need resize the data by sample it. A bootstrap will apply while choose sample rate.
- In our project, lots of location information will be used which can not be used directly. Convert it into numeric value by distance or area density is one method we will apply. New features add in and drop old one. Since the location dataset has different density wide around NYC, our dataset need rescale at last.

(c) Correlation & Covariance

- Some of our feature selected might be useless for our questions or dependent with each other. We will analyze covariance to filter exist features. Some features might be combined by Principal Component Analysis(PCA). Some trivial features will be drop.
- After dimensionality reduction, our data set will has less noise and more suitable for further model fitting

(d) ARIMA time series

- The NYC Crime Data under an obvious time series form. Some information can be obtained by analyze in time domain.
- We will analyze whether crime record has some periodic feature like ‘more crime in winter’ or ‘what type of crime more likely happened in winter’. Also, stationary the data can reveal if total crime records increasing cross years, and give us a better view to calculate feature like ‘special type crime rate’ or ‘crime density’. By fitting an ARIMA model to several kinds of crimes, we even can predict a probability of specific type of crime in an area.

(e) Hyperparameters fine-tuning, model selection and evaluation

- Fine-tune the hyperparameters to make the model perform well.
- Try different model to solve our questions, evaluate them and find best solution.