Crime Frequency Prediction

http://130.211.140.62:5000
Yuanqing Hong, Weiyi Li, Claire Lee
APMA 4990 Final Project
Dorian Goldman

Motivation

The goal of our project is to provide a prediction on the crime frequency of different precincts in Manhattan with varying time to help people plan a safe trip.

Audience

The project is aimed at residents in Manhattan so they can enter 1) Day, 2) Time, and 3) PCT to look up the predicted crime frequency to anticipate potential risk.

Dataset

We used two sets of public data from NYC Open Data: 1) NYPC Complain Data Current YTD (479,000x24) and 2) NYPD Complain Data Historic (5,100,000x24) to train our Random Forest Regression model. Specifically we selected the four independent features out of the total 24 columns from the datasets: CMPLNT_NUM, CMPLNT_FR_DT, CMPLNT_FR_TM, and ADDR_PCT.

Methods

Random forest regression algorithms were performed on the dataset due to the nonlinear pattern of crime occurrences in Manhattan.

Part I. Data Cleansing

- Feature Selection: Complaint Number, Complaint From Date, Complaint From Time, and Location
- Variable Definition:
    - "Frequency " is defined by the number of crimes per six hours in a region
    - Four Compartmentalization of a day - to train our model, we divide a day (24h) into four parts, with each part taking six hours. (Morning[0] 6:00-12:00 , Afternoon[1] 12:00-18:00, Evening[2] 18:00-0:00, and Late Night[3] 0:00-6:00)
    - "PCT" - NYPD divided into 22 precincts and assigned regional codes.

```
df_=df[['CMPLNT_NUM', 'CMPLNT_FR_DT', 'CMPLNT_FR_TM', 'ADDR_PCT_CD']]
df_.head()
```

	CMPLNT_NUM	CMPLNT_FR_DT	CMPLNT_FR_TM	ADDR_PCT_CD
0	394278393	1/1/15	9:00:00	41
1	722254800	1/1/15	0:01:00	77
2	530381050	1/1/15	0:00:00	47
3	429186741	1/1/15	12:00:00	50
4	388921549	1/1/15	0:01:00	88

Table 1. Sample of the data frame used in training the model

Part II. Random Forest Regression

The purpose of regression was to determine the relationship between [time, location] and the crime frequency in the 22 precincts of Manhattan for each day of the week. Initially, we ran random forest regression using the hourly crime report for each day of the week. However, we found that the random forest regression model did not perform well due to the high variability of the data. Therefore, we aggregated the crime reports and divided 24 hours into the four parts: late night(midnight~6am), early morning (6am~noon), afternoon (noon~6pm), and evening (6pm~midnight), which yielded a significantly improved performance.

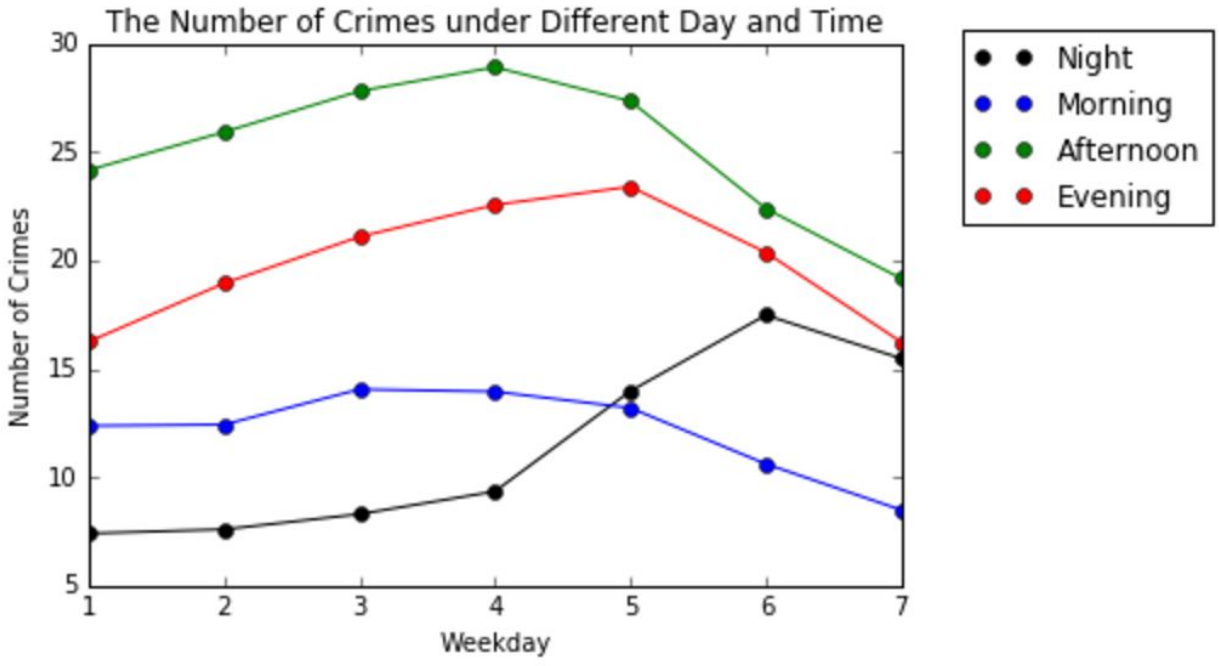

Figure 1. The number of crime with varying day and time

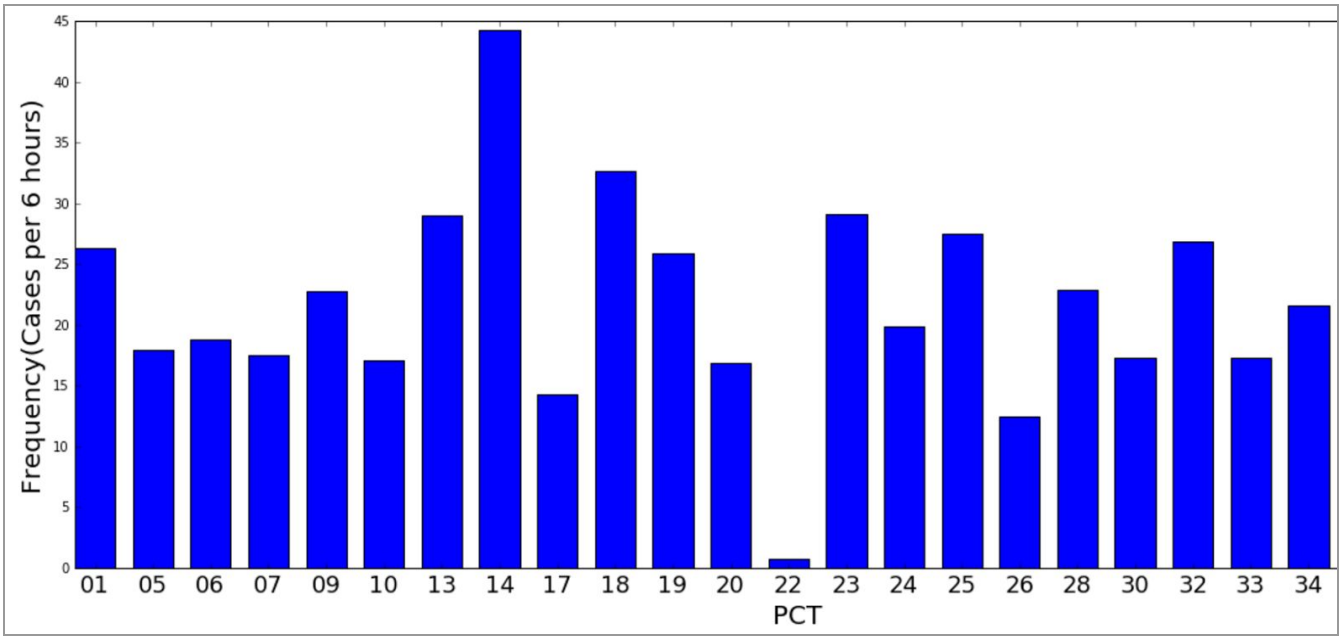

Figure 2. The number of crime of the 22 PCTs in Manhattan

Tree Depth	# of Trees	Score
100	100	0.861789557
100	90	0.86174069
100	80	0.861600828
100	70	0.862144633
100	60	0.861712505
100	50	0.860983779
100	40	0.863501545
100	30	0.859851868
100	20	0.85793175
100	10	0.861375598
100	5	0.85263743
100	1	0.808536164
90	100	0.862624469
90	90	0.860959332
90	80	0.861448649
90	70	0.861316454
90	60	0.859619117
90	50	0.860674133
90	40	0.862067039
90	30	0.859096748
90	20	0.857316391
90	10	0.780753348
90	5	0.780753348
90	1	0.780753348
80	100	0.86202573
80	90	0.860726948
80	80	0.862260609
80	70	0.860914372
80	60	0.860233691
80	50	0.861992429
80	40	0.860297896
80	30	0.860321298
80	20	0.85793175
80	10	0.857563057
80	5	0.8559895
80	1	0.800278211
70	100	0.862676641
70	90	0.861489422
70	80	0.86204288
70	70	0.862801232
70	60	0.862760879
70	50	0.862752289
70	40	0.859881125
70	30	0.859772009
70	20	0.856515952
70	10	0.857808597

Tree Depth	# of Trees	Score
50	20	0.878759303
100	40	0.863501545
30	80	0.862976156
70	70	0.862801232
70	60	0.862760879
70	50	0.862752289
60	100	0.86269666
60	70	0.862691309
50	60	0.862679188
70	100	0.862676641
90	100	0.862624469
20	100	0.862494258
20	80	0.862361955
20	50	0.862317762
80	80	0.862260609
40	70	0.862251238
30	60	0.862248366
30	90	0.862237119
100	70	0.862144633
90	40	0.862067039
70	80	0.86204288
50	90	0.862039636
80	100	0.86202573
80	50	0.861992429
60	90	0.861939748
30	100	0.861876217
50	80	0.861812668
100	100	0.861789557
40	40	0.861747233
100	90	0.86174069
100	60	0.861712505
100	80	0.861600828
20	30	0.861535022
70	90	0.861489422
90	80	0.861448649
50	70	0.86144258
40	90	0.861397454
40	60	0.861397048
100	10	0.861375598
30	70	0.861355321
90	70	0.861316454
20	40	0.86118275
50	100	0.861095957
40	80	0.861015578
20	60	0.861006033
100	50	0.860983779

Table 1&2: Performance evaluation on the algorithm with the varying number and depth of tree

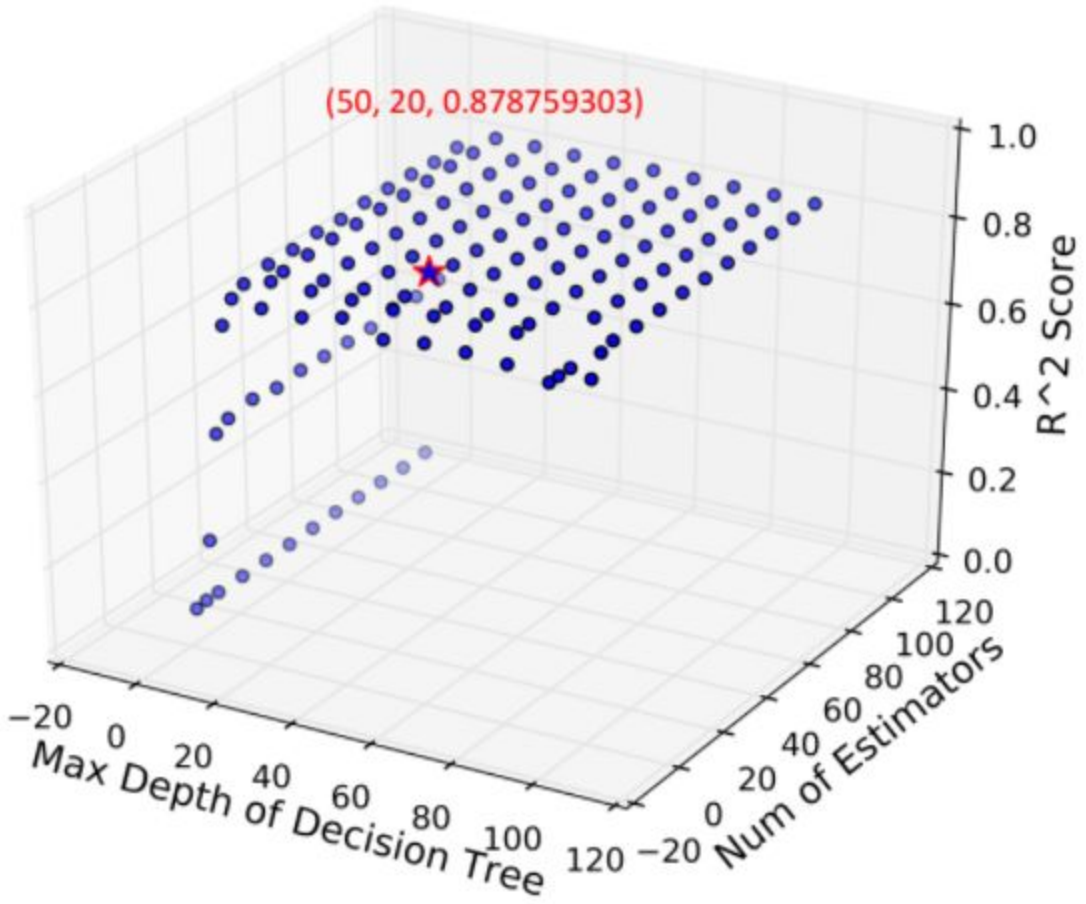

Figure 2. R^2 Score to determine the optimal number of decision trees and estimators

We tested on various pairs of the different values of decision tree and depth. And the result above shows that our prediction model performs the best when the number of trees is 20 and the depth of the tree is 50. Based on this observation we set the design of our random forest model to yield the most accurate prediction.

Part III: Front-End Engineering

Our web app was deployed using Google Cloud services. The user inputs weekday and time of day. Then the crime frequencies in different precincts of Manhattan are shown in the page. Also we used Google Map API to display the boundaries of PCT with the opacity corresponding to the predicted frequency.

Limitations and Future Work

There are two limitations in the project due to the nature of dataset. First, the crime reports are not evenly distributed which can potentially distort the prediction model. Therefore, we decided to only look at the data collected over the last one and a half year from January 2015 to April 2016 to gain a better understanding of the pattern in the crime occurrence in Manhattan. Lastly, our Random Forest algorithm was designed with only few features (time of day, weekday, and location) and it did not consider other factors. Thus it might not be able to give a comprehensive outlook on criminal activities in Manhattan.

Regarding application of our project, it can be used to alert Manhattan residents on a potential crime incident throughout day and week by sending out a notification to ensure their safety.