Crime Analytics and Prediction in Los Angeles

Hye Lim Park
University of California, San Diego
Spring 2015 CSE 190
Assignment 2
A97010359
hlpark@ucsd.edu

ABSTRACT

Crime can be considered as difficult subject for analytics and prediction because people could commit crimes at random; and actual crime prediction model could never be faultless as "Minority Report." However, analyzing previous criminal data and predict possible crimes would definitely help reduce crime rate and prevent future loss. In this paper, I analyze crimes occurred in 2014 and predict possible crimes in future in Los Angeles area using LAPD Crime and Collision Raw Data – 2014, which includes all crimes and collisions reported in 2014. For minimizing confusion in the dataset, I exclude crimes occurred before 2014 and all collisions reported in 2014. By analyzing the dataset, I sort the safest area and the most dangerous area, the safest month and the most dangerous month of the year, the safest date and the most dangerous date of the month, and the safest hour and the most dangerous hour of the day based on the number of crimes occurred at each feature. I build two models using linear regression and popularity algorithm (baseline) for crime prediction based on location, month, date, and hour. I randomly choose 50,000 test set include location, month, date, and hour from training set and evaluate two models with mean square error and classification accuracy in order to distinguish how appropriate these models are for crime prediction.

Keywords

Crime, Prediction, Analytics, Los Angeles, Linear Regression, Algorithm

1. DATASET

LAPD Crime and Collision Raw Data – 2014 contains all crimes and collisions reported in 2014. I exclude crimes occurred before 2014 and all collisions reported in 2014 to eliminate possible confusion when processing the dataset. The data includes features like date reported, unique crime id, date occurred, time occurred, area code, area name, rd, crime code, crime description, status, status description, location, cross street, latitude and longitude. Among these features, I use date occurred, time occurred, area code (with area name), crime cd (with crime description), latitude and longitude when analyze the dataset.

There are total 188205 crimes occurred, 21 area codes, and 126 crime codes in the dataset. In order to analyze the dataset based on month, date, and hour, I relate the number of crimes occurred with its month (January to December), date (1 to 31), and hour (0 to 23). The most occurred crime overall is 'battery - simple assault', occurred 18297 times, and the least occurred crimes are 'dishonest employee attempted theft', 'purse snatching - attempt', 'replica firearms (sale, display, manufacture or distribute)', 'bike - attempted stolen', 'till tap - grand theft ($950.01 & over)', 'bribery', 'theft, coin machine - attempt', and 'grand theft / insurance fraud', all occurred 1 time.

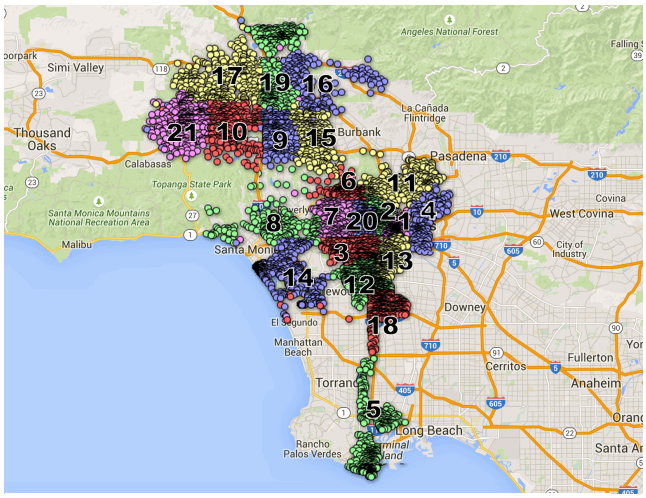

Figure 1. Area Codes Map

Figure 1 shows the area codes map that indicates each area with corresponding code numbered 1 to 21. Each point represents crime occurred within the area and each color is for distinguish boundary for the area. Figure 2 shows heat map of crime occurred in Los Angeles area. Assuming the area is divided into three groups (top, middle, and bottom), crime is more likely

to be occurred in the middle of each group than exterior side.

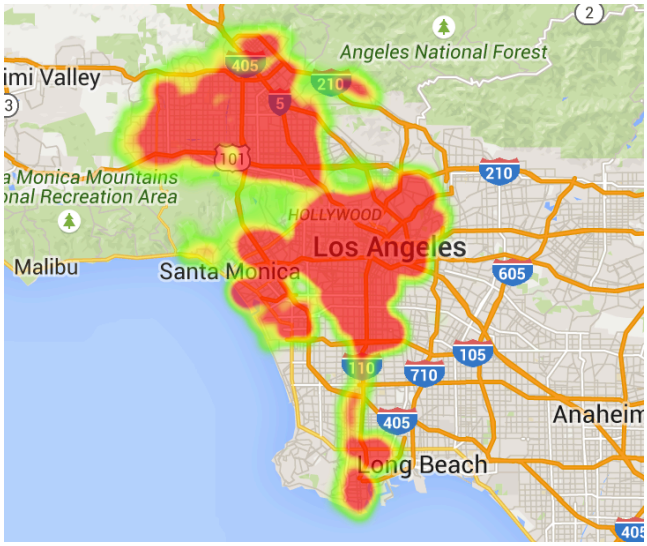

Figure 2. Crime Occurred Heat Map

Figure 3 shows the total number of crimes of each area. As two most dangerous areas, total 13633 crimes occurred in Area 12 and 12567 crimes occurred in Area 3. On the other hand, as two safest areas, 6671 crimes occurred in Area 4 and 6801 crimes occurred in Area 16.

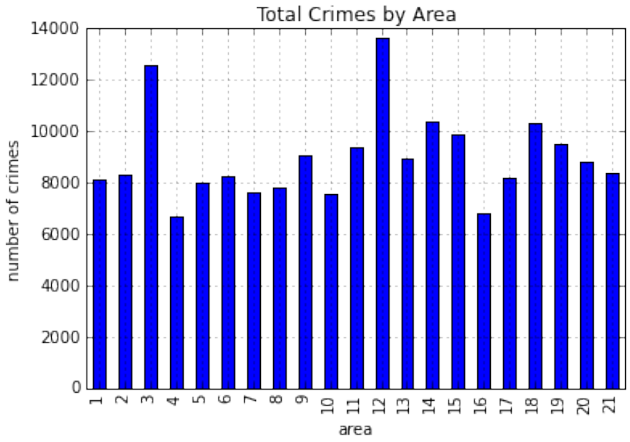

Figure 3. Total Number of Crimes by Area

Figure 4. Total Number of Crimes by Month

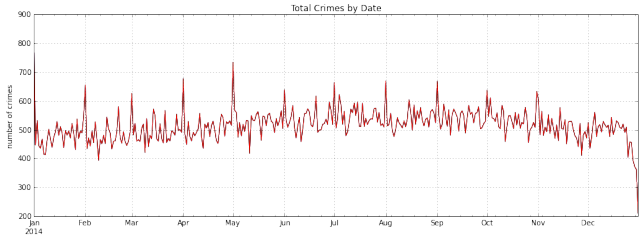

Figure 5. Total Number of Crimes by Date

Figure 4 shows the total number of crimes of each month. I count the number of crimes occurred from the first day of the month until the last day of the month. As the most dangerous month of the year, total 16935 crimes occurred in July and as the safest month of the year, total 13431 crimes occurred in February.

Figure 5 shows the total number of crimes of each date of the each month. More crimes are likely to be occurred on the first day of the month. For example, on the first day of May, 734 crimes occurred while about 500 crimes occurred on the rest of days.

Figure 6 shows the total number of crimes of each hour of the day. As the most dangerous hour of the day, 14467 crimes occurred at 17:00 and as the safest hour of the day, 2303 crimes occurred at 8:00. More crimes are likely to be occurred during the nighttime than the daytime.

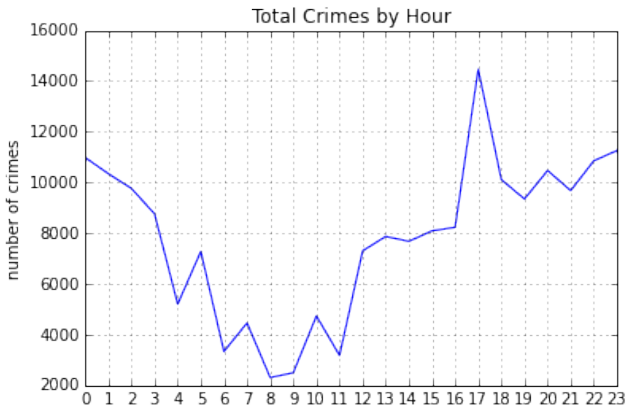

Figure 6. Total Number of Crimes by Hour

2. PREDICTIVE TASK

Using four features, location, month, date, and hour, I predict the number of possible crimes with given location, month, date and hour. Also, I predict the possible crime by its crime code like recommendation system. Rather than suggest recommended movies or foods based on user's characteristics, my prediction model warns possible crimes based on user's location, month, date, and hour.

To predict the number of possible crimes, I build prediction model with linear regression. Area code, 1 to 21 is represented with 0 and 1 such as 1 is represented as [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] and 15 is represented as [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0]. In the same way, month (1 to 12), date (1 to 31), and hour (0 to 23) is represented with 0 and 1. I use linear regression to build the model because, as shown from the first section, all features (location, month, date, hour) are significantly associated with the number of crimes. Thus, by using thetas corresponding to these features, the model predicts more accurate number. Accuracy of the model is evaluated with mean square error.

In order to predict possible crimes based on location, month, date, and hour, first I intuitively build the baseline using either location, month, or date. Assuming location, month, date, and hour are given, the baseline predicts the possible crime by counting the number of crimes occurred based on its crime code. First, the model outputs crime code of the most occurred crime in the area, then crime code of the most occurred crime of the month, crime code of the most occurred crime of the date, and lastly, crime code of the most occurred crime of the hour. I randomly choose 50,000 test set from training set, predict possible crime, and calculate classification accuracy for comparing the baseline.

To improve the accuracy, I not only combine all four features but also increase number of possible crimes (originally 1) as an improved baseline. First, the model retrieve each 5, 10, 15, 20 possible crimes from the dataset based on the most occurred crimes of each of four features and assume prediction is correct if actual crime is in the intersection of those four possible crime sets.

3. LITERATURE

LAPD Crime and Collision Raw Data – 2014 is intended for public access and use [1]. Not many researches have been done with this dataset; however, data mining and predictive analytics with criminal data has become one of the most important and fastest growing research areas. In August 2014, Los Angeles Mayor has hired "the city's first chief data officer, Abhi Nemani [2]. Most of researches with criminal data are focusing on analyzing crime patterns based on various factors. For example, Cung relates the record of crimes with weather data and demographic information. Her research doesn't include any predictive task, however, she uses clustering approach to analyze crime patterns. While my analytics is focusing on features directly from criminal data such as location, occurred month, date, and time, Cung focuses on referenced data such as weather, population, and holidays [3].

4. RESULTS

First, I build two models predicts the number of possible crimes based on location, month, date, and hour. In order to check whether mean square error for linear regression model is reasonable, I build baseline model, which divides test set into five groups based on its average of 4 values (number of crimes of each factor: location, month, date, hour) in descending order. If given features belong to first group, the model predicts 5, second group predicts 4, third group predicts 3, fourth group predicts 2, and fifth group predicts 1. As shown in Table 1, for same test set (50,000 randomly chosen data), mean square error of baseline model is 3.00 while linear regression model is 1.74.

Table 1. Mean Square Error for Number of Crimes Prediction

	Baseline	**Linear Regression**
MSE	3.00	1.73944884853

Then I also build two models predicts the crime code of possible crime based on location, month, date, and hour. As shown in Table 2, the baseline has very low accuracy rate (average 0.10835) regardless to what feature is used for prediction. Table 3 shows classification accuracy based on number of possible crimes when all four features are considered. As number of possible crimes increases, accuracy is getting higher. However, as number of possible crimes increases, false positive rate is also getting higher because the improved baseline ignores rest of possible crimes in the intersection set but only consider whether the set contains actual crime occurred or not.

Table 2. Classification Accuracy of the Baseline (Each feature)

	Location	Month	Date	Hour
Classification Accuracy	0.1129	0.09772	0.1002	0.12258

Table 3. Classification Accuracy of the Baseline (All features)

	5 Crimes	10 Crimes	15 Crimes	20 Crimes
Classification Accuracy	0.25964	0.56206	0.7976	0.84708

From these results, I conclude that linear regression model performs better than baseline model and its mean square error is reasonable in order to predict the number of possible crimes. Since all features (location, month, date, hour) are significantly related with the number of crimes, model based on linear regression model succeeds while baseline model (grouping by its average) fails. Also, I conclude that in order to predict the crime code of possible crime, the model using all features performs better than the model using one of features. The model using all features succeeds while the model using each feature fails because the crime is dependent to various features rather than only one feature. Since improved baseline model achieves high accuracy rate by ignoring remaining predicted crimes in output set, in future, the model needs to be improved to reduce false positive rate. Overall, both models demonstrate that crime is relevant to features like location, month, date, and hour. In future, I would want to build the prediction model with more various features like characteristics of victim and suspect (while maintain confidentiality) with bigger dataset like all crimes occurred in last 10 years.

5. REFERENCES

[1] http://catalog.data.gov/dataset/lapd-crime-and-collision-raw-data-2014-db997

[2] Ellingson, Annlee, 2014, L.A. appoints first chief data officer, L.A. Biz, http://www.bizjournals.com/losangeles/news/2014/08/21/l-a-appoints-first-chief-data-officer.html

[3] Cung, B. 2013. Crime and Demographics: An Analysis of LAPD Crime Data. University of California, Los Angeles