Crime and commercial activities in Toronto neighborhoods

Richard Hou

Executive summary

In this project, we intend to study the relationship between crime rate/types and commercial activities in Toronto neighborhoods. In particular, we want to inspect the relationship/association between certain commercial activities in a neighborhood, represented by the presence as well as density of certain commercial venues in a neighborhood, and the types of crime and overall crime rates in the underlying neighborhood. In order to do so, we have obtained the historical crime statistics from Toronto Open Data Portal, Toronto neighborhood data from Wikipedia, and also commercial venue information from Foursquare. The data were first thoroughly cleaned and then transformed and merged together for further analysis. Different data analytical techniques have been employed to analyze the data, for example, data visualization, descriptive statistics, and more advanced machine learning techniques, such as logistic regression, decision tree, support vector machine (SVM) and neural network (NN). Predictive models have been developed to predict the probability of occurrence of certain crimes with a specific neighborhood, given presence and density of all the commercial venues, and their performances are compared. Overall the model performance is not satisfactory with the best accuracy ratio (achieved by using RandomForest method) being around 62%. However, we did gain extra insight with regard to what might be important factors deciding the types of crimes.

1 Introduction

Toronto is the capital city of Ontario as well as the largest city by population in Canada. As of 2016 it had a population of over 2,700,000 (wikipedia). Being the most populous city in the country, Toronto is of great importance both domestically and globally.

1.1 Business

It is an international centre for business and finance. Generally considered the financial capital of Canada, it has a high concentration of banks and brokerage firms on Bay Street. The Toronto Stock Exchange is the world's seventh-largest stock exchange by market capitalization. In particular, the city accounted for nearly one quarter of the country’s employment in the finance and insurance industry (wikipedia).

Also the city is an important centre for the media, publishing, telecommunication, information technology and film production industries; it is home to Bell Media, Rogers Communications, and Torstar. Other prominent Canadian corporations in the Greater Toronto Area include Magna International, Celestica, Manulife, Sun Life Financial, the Hudson's Bay Company, and major hotel companies and operators, such as Four Seasons Hotels and Fairmont Hotels and Resorts.

1.2 Living

In 2018, Toronto ranked 7th (tied with Tokyo) for the world's most livable cities (3rd in North America) according to the Economist Intelligence Unit Opens in new window (toronto.ca), based on . Cities are rated across five categories; stability, healthcare, culture and environment, education and infrastructure. In each of the categories, cities are given a score between 1 and 100, where 1 is considered intolerable and 100 is considered ideal. Once all category scores are compiled and weighted, an overall score is given between 1 and 100. Toronto received an overall score of 97.2. Toronto received high scores (100) in stability, healthcare and education. Toronto's lowest score (89.3) was in infrastructure. (Toronto.ca)

1.3 Crime

Although being the most populous city in Canada, the overall crime rate in Toronto is below than national average, based on Statistics Canada data.

1.4 Project description

In this project, we intend to study the geo-distribution of crime rate in Toronto and particularly its relation/correlation with commercial activities in the neighborhood. The results would be of great interest for both business persons, who would naturally be interested in doing business in a safe and prosperous neighborhood, and government, who is keenly interested in making Toronto a more livable city.

As it is an exploratory study, we do not foresee what the outcome would look like, but ideally we hope to get some insight regarding interactions between crime activities and business activities. More specifically, we want to answer the following questions:

1. How is the crime rate distributed in Toronto neighborhood? Where are neighborhoods with the highest/lowest crime?

2. What do the crime rate distributions differ for different crime types?

3. How is the business activity (venue density) distributed in Toronto neighborhood? Where are neighborhoods with the highest/lowest, e. g., restaurant concentration and so on?

4. Are there any relationship/correlation/association between crime rate (or a specific type of crime rate) and particular type of business venue or vice versa?

1.5 Target Audiance

1. Business persons, who are interested in a perfect location for their businesses

2. Municipal government, who is interested in making Toronto a more livable city.

2 Source Data Description

In this project, we are planing to use mainly three data sources, namely, Foursquare data, City of Toronto Crime Statistics data and also the neighbourhood data on [Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M).

2.1 Foursquare API data

Foursquare provides different endpoints, from which information about venues and/or users can be retrieved. We are particularly interested in the "explore" endpoint to get basic information of venues in a

neighborhood their categories and so on. This information is then going to be combined with Crime Statistics data above.

Table 1 Foursquare API data option

Endpoint	Usage
search	Search for Venues
explore	**Get Venue Recommendations**
select	Report Venue Selection
trending	Get Trending Venues
suggestcompletion	Suggest Completions
likes	Get Users Who Liked a Venue
categories	Get Venue Categories
similar	Get Similar Venues
nextvenues	Get Next Venues
listed	Get Lists a Venue is On

2.2 City of Toronto Crime Statistics

This data is available via City of Toronto Open Data Portal and is shared by [TorontoPoliceService](http://data.torontopolice.on.ca/datasets/98f7dde610b54b9081dfca80be453ac9_0). This dataset can be downloaded as CSV or JSON file. It includes all Major Crime Indicators (MCI) 2014 to 2018 occurrences by reported date and related offences. Below is a list of variables that are included in the datasets.

Table 2 Crime data columns/fields

Fields	Field_Description
Index	Unique ID
event_unique_id	Occurrence number
occurrencedate	Date of occurrence
reporteddate	Date occurrence was reported
premisetype	Premise where occurrence took place
ucr_code	URC Code
ucr_ext	URC Code Extension
offence	Offence related to the occurrence
reportedyear	Year occurrence was reported
reportedmonth	Month occurrence was reported
reportedday	Day occurrence was reported
reporteddayofyear	Day of week occurrence was reported
reporteddayofweek	Day of year Occurrence was reported
reportedhour	Hour occurrence was reported
occurrenceyear	Occurrence year
occurrencemonth	Occurrence month
occurrenceday	Occurrence day
occurrencedayofyear	Occurrence day of year
occurrencedayofweek	Occurrence day of week
occurrencehour	Occurrence hour
MCI	Major Crime Indicator related to the offence
Division	Division Assigned to occurrence after offsetting X and & Coordinates to nearest intersection node
Hood_ID	Nieghbourhood Name Assigned to occurrence after offsetting X and & Coordinates to nearest intersection node
Neighbourhood	Nieghbourhood ID Assigned to occurrence after offsetting X and & Coordinates to nearest intersection node
Long	Longitude of point extracted after offsetting X and & Coordinates to nearest intersection node
Lat	Latitude of point extracted after offsetting X and & Coordinates to nearest intersection node

2.3 Postal codes for Toronto neighborhoods from Wikipedia

The last data source is a table from Wikipedia that contains postal codes for all Toronto neighborhoods. The first a few rows of the table are shown below just as an example.

Table 3 Wikipedia data: Toronto Neighbourhood (only the first a few rows are shown here for examples.)

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M5A	Downtown Toronto	Regent Park
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor
M7A	Queen's Park	Not assigned
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue

We'll need it to get the geo-codes of the neighborhoods. The Crime dataset does contain the geo-codes of all crime incidents and corresponding neighborhood, however, as one can see, there is no geo-code for neighborhoods. Therefore we shall take a little detour to get it.

2.4 Data usage

The two data sources will eventually be combined on ***Neighbourhood***. Further data analysis and modelling will be based on the final dataset.

3 Data Exploration

In this section, we are going to explore the data. There are a few purposes here.

1. To check the data quality, such as missing values, outliers and other apparently data quality issues;
2. To see what each variable is, its distribution, pattern and trend and so on. This step is usually called univariate analysis;
3. To check some simple correlations/associations within the data. This step is usually called bi-variate analysis; (The discussion of the results will be postponed till next section when we have the final dataset.)
4. To perform necessary transform and manipulation of data;
5. To prepare the data for final analysis.

Selected results are shown below. The selection is based on both intuition as well as feature importance that is later revealed in RandomForest method. For most of the selected features, we are going to inspect its effect on number of crimes and types of crimes.

The following chart shows the distribution of number of crimes by day of the week. One can see Friday has the highest number of crimes during a week, followed by Saturday and Sunday. Other week days seem to be rather steady, though Tuesday seems to have the least number of crimes.

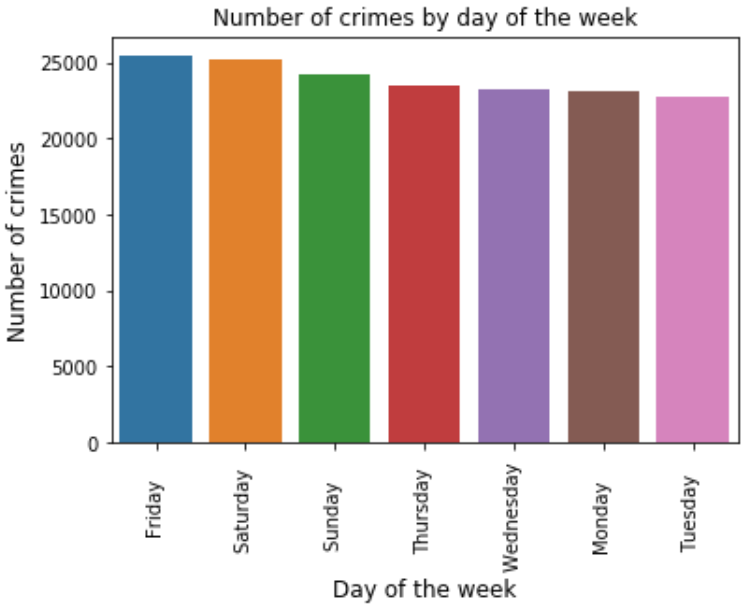

Figure 1 Original data: Number of crimes by day of the week.

The following chart shows distribution of number of crimes over by the month of the year. We see that temperature seems to have an effect on crime rate. Crime rate appears to be higher in the summer seasons, extending to late spring and early autumn, with October enjoying the highest number of crimes and February the least. However, February being the least needs to be interpreted with cautions. It is probably due to the fact that February has 2 - 3 days less than other months.

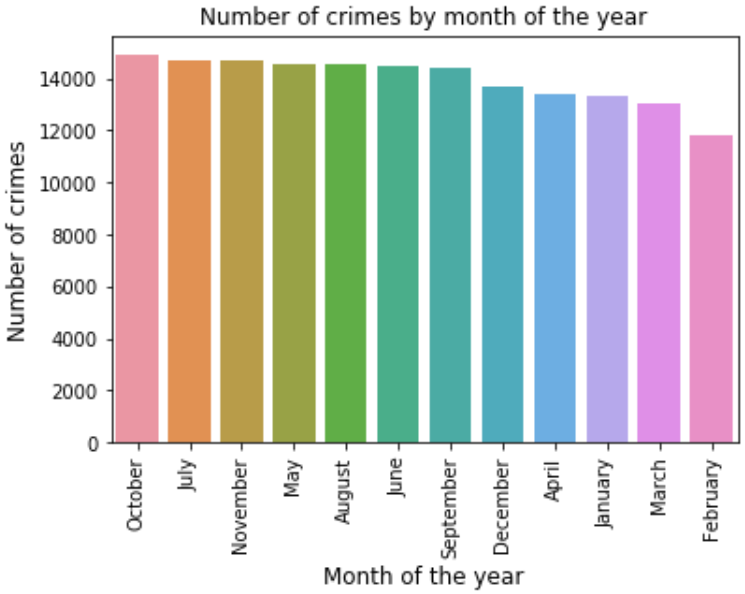

Figure 2 Original data: Number of crimes by day of the week.

The following chart shows the number of crimes by premise type, and the numbers are sorted from the highest on the left-hand-side (Outside) to the lowest on the right-hand-side (Other).

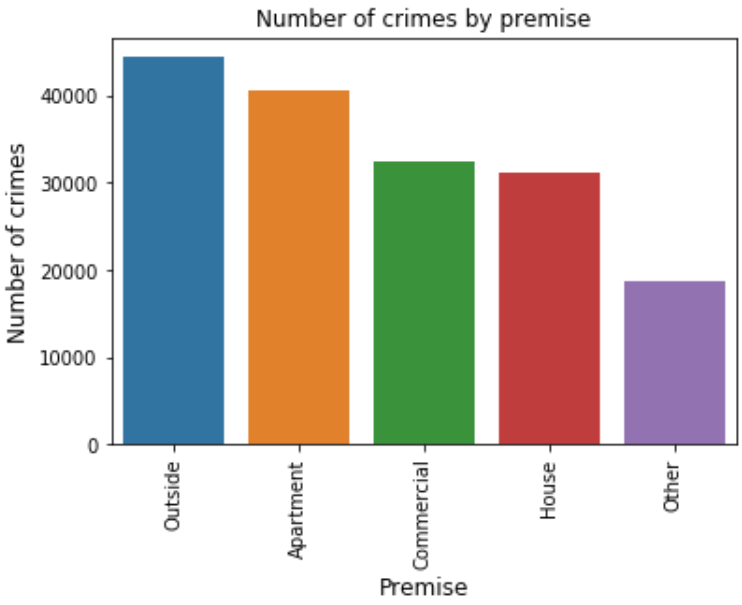

Figure 3 Original data: Number of crimes by premise.

The chart below depicts the number of crimes by MCI (Major Crime Indicators). “Assault” ranks the highest in terms of number of crimes, while “Theft Over” the lowest. As we’ll see later on in the final dataset that most of the assaults happened outside. That probably partly explains why “Outside” ranks the highest in the premise.

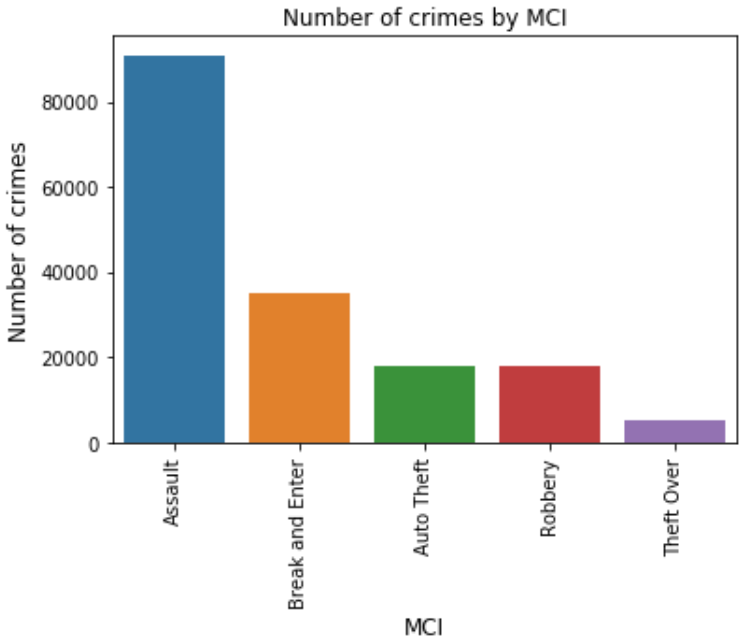

Figure 4 Original data: Number of crimes by MCI.

4 Final Dataset

Due to data quality concern, the final dataset we have focus on only 5 neighbourhood, as is shown below:

Table 4 List of neighbourhoods in the final dataset.

	Postcode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront,Regent Park
3	M6A	North York	Lawrence Heights,Lawrence Manor
4	M7A	Queen's Park	Queen's Park

And in the final dataset, the top 5 categories of venue are shown below in the list. As one can see that coffee shop leads the list, followed by Italian Restaurant, Park and Bakery.

Table 5 List of top 5 venues in the final dataset.

	Venue value counts	Venue Category
Coffee Shop	159	Coffee Shop
Café	123	Café
Italian Restaurant	71	Italian Restaurant
Park	66	Park
Bakery	63	Bakery

The following chart depicts the distribution of types of crimes in the final dataset.

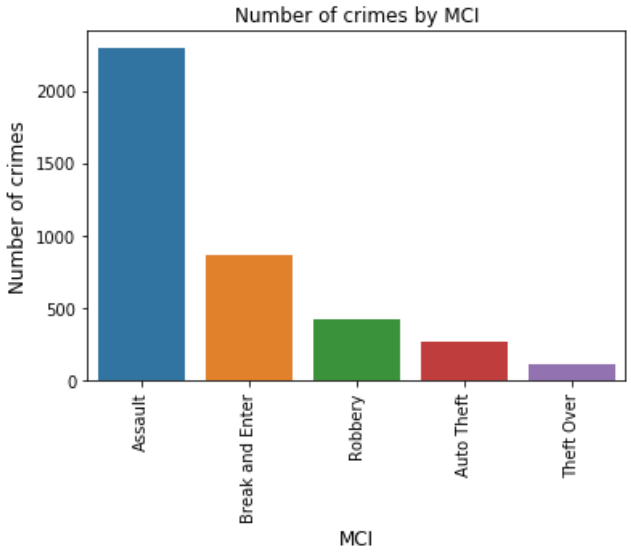

Figure 5 Final dataset: Number of crimes by MCI.

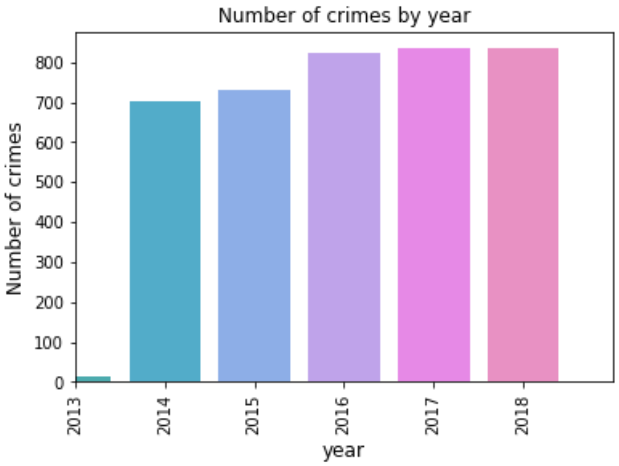

Figure 6 Final dataset: Number of crimes by year. There is a slight tendency of increase in the number of crimes over the years.

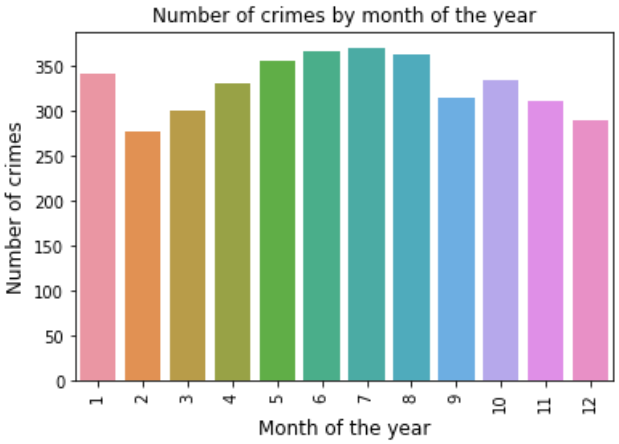

Figure 7 Final Dataset: Number of crimes by month of the year. We see very similar pattern to the one we observed in original data. Again, February has the lowest number mainly due to the fact that it has the least days.

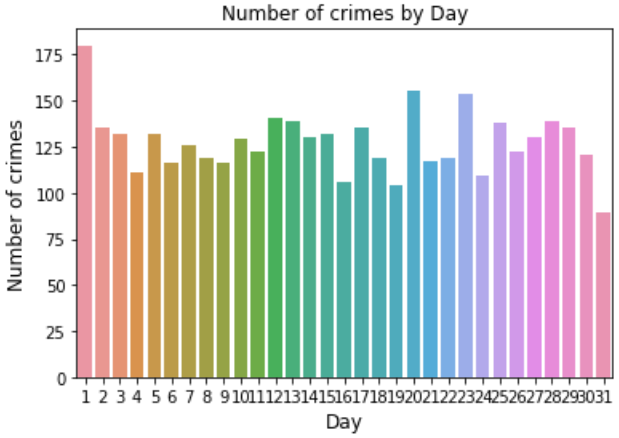

Figure 8 Final dataset: Number of crimes by day of the month. We see very similar pattern to the one we observed in original data. Again, 31 has the lowest number mainly due to the fact that only 7 out of 12 months have 31 days.

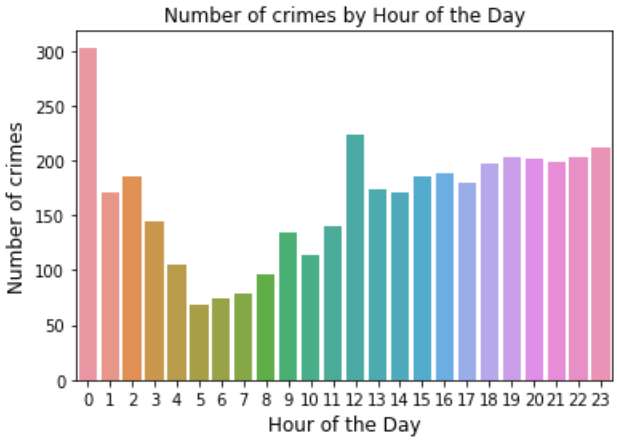

Figure 9 Final dataset: total number of crimes by Hour of the Day. We see very similar pattern to the one we observed in original data. The number of crimes spikes at the mid of the night, and also mid of the day.

It should be noted that different crimes may have different patterns in the hour of the day. The following chart shows the patterns for different crime types. Some very interesting observations, such as, Robbery happens mostly in the evening, and Theft Over happens mostly during the noon.

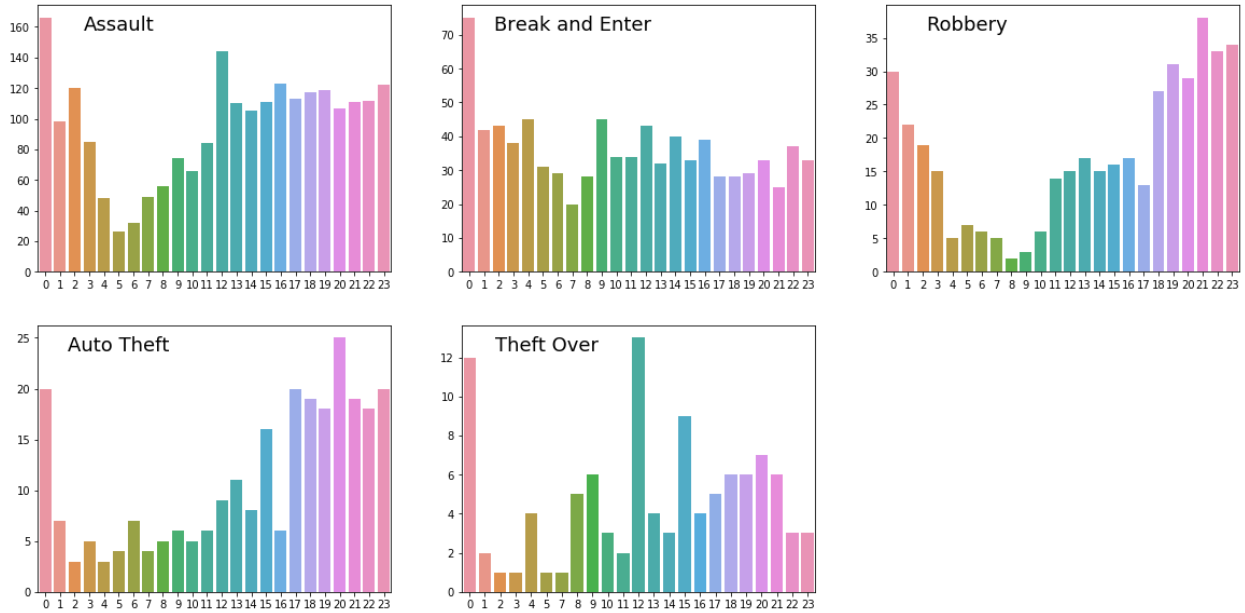

Figure 10 Final dataset: Number of crimes by hour of the day for different crime types. Some very interesting observations, such as, Robbery happens mostly in the evening, and Theft Over happens mostly during the noon.

The following chart shows the distribution of number of crimes by premise. We see very similar pattern to the one we observed in original data. Outside is leading in the number of crimes, followed by Apartment, commercial, house and then other.

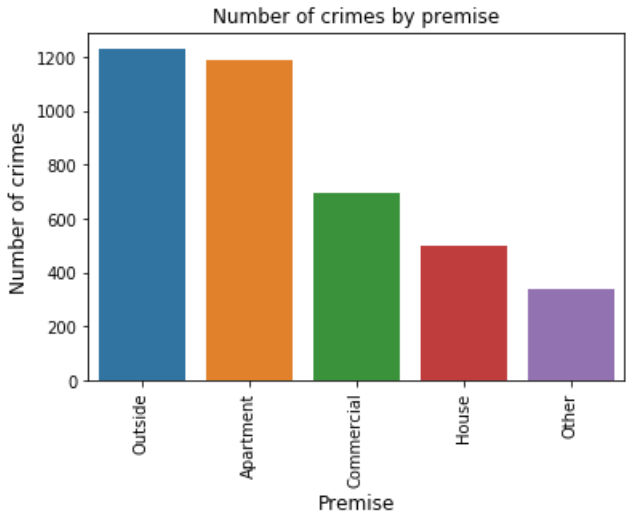

Figure 11 Final dataset: Total number of crimes by premise. We see very similar pattern to the one we observed in original data.

It is not surprise to see below that the order of the most frequent premises depends on the type of crimes. As one can see that most of the Assault happens in Apartment; most of Break and Enter happens in Commercial places; most of Robbery and Auto Theft happen Outside; and most of Theft Over happens in House.

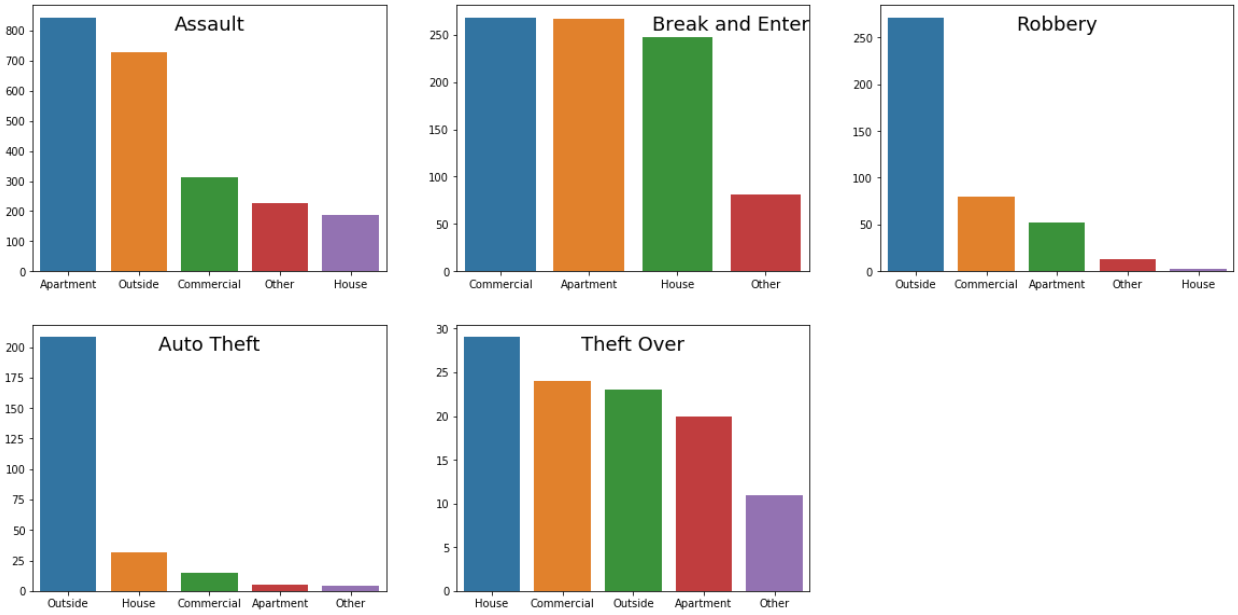

Figure 12 Final dataset: Number of crimes y premise for different crime types. As one can see that most of the Assault happens in Apartment; most of Break and Enter happens in Commercial places; most of Robbery and Auto Theft happen Outside; and most of Theft Over happens in House.

5 Machine Learning

After thorough data exploration, we have much better understanding of the data we have and a better definition of the problem we are try to solve.

In this section, we are going to apply various machine learning methods to study the relationship of type of crimes and all the other variables. So our target variable is MCI, which can take: “Assault”, “Break and Enter”, “Robbery”, “Auto Theft” and “Theft Over”, and our predictive/independent variables or simply predictors are anything else in the data. We have removed certain variables, such as the specific longitude and latitude and so on. Remaining variables are mainly time of the crime incident, and number of specific category of the venue in the corresponding neighborhood.

For those who are not so familiar with different Machine Learning algorithms, the following chart from SAS gives a very good illustration.

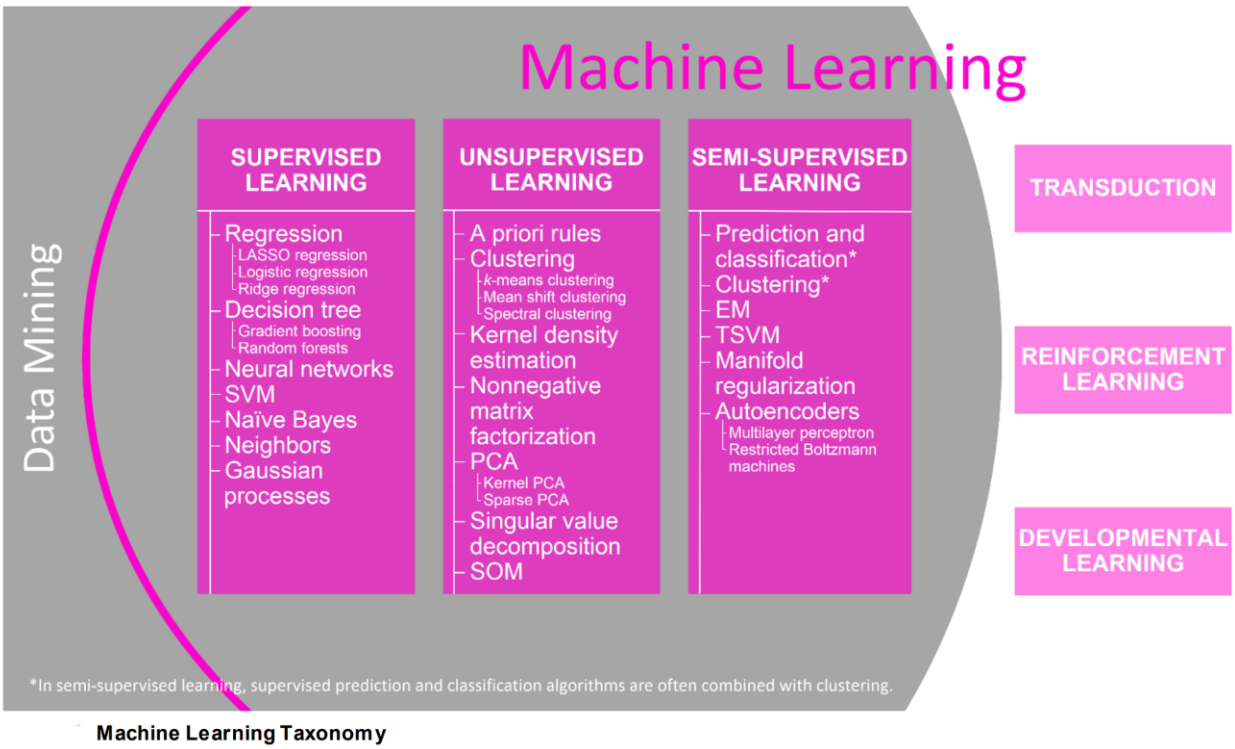

Figure 13 Machine Learning Taxonomy. Source: An Overview of SAS® Visual Data Mining and Machine Learning on SAS® Viya, SAS Institute Inc.

There is also a very nice machine learning cheat-sheet from SAS available to help determine which algorithms can and should be used.

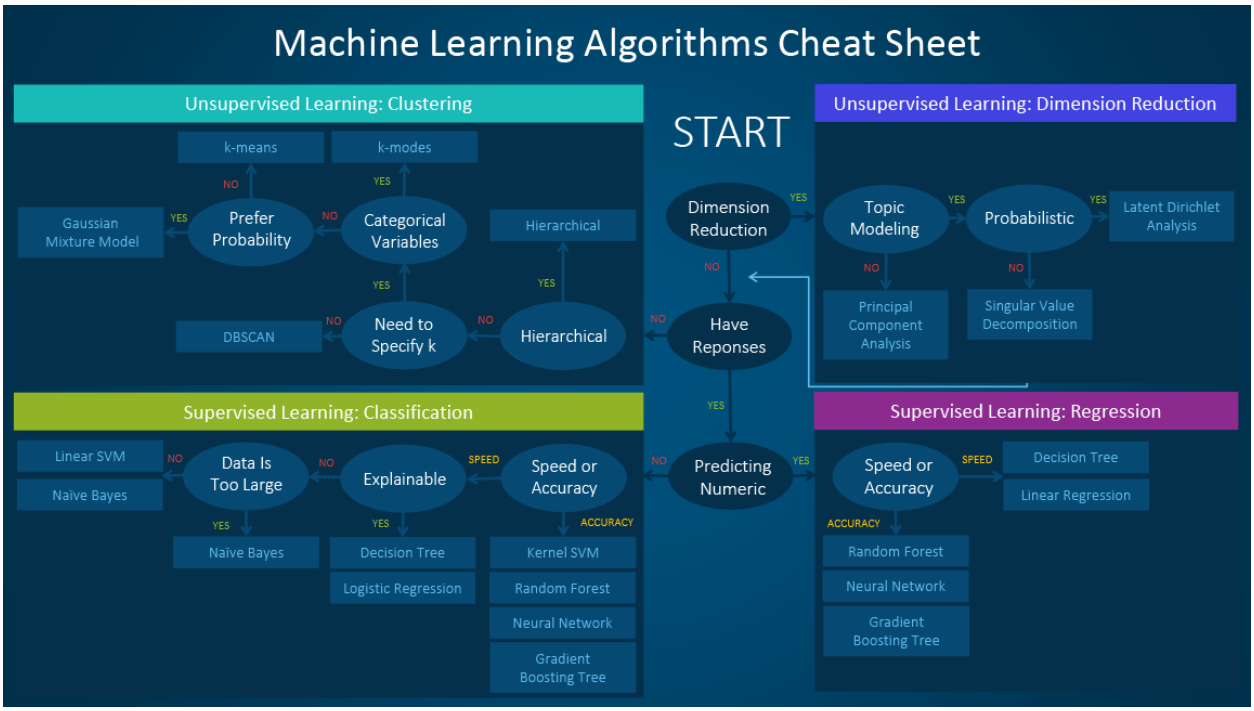

Figure 14 Machine Learning Cheat-sheet. Source: https://blogs.sas.com.

Following the above Taxonomy and Cheat-sheet, we see that our problem is under supervised learning category and then under classification. The methods we tried include some of the most common classifiers, such as K-Nearest-Neighbor (KNN), Logistic regression, Decision Tree, Support Vector Machine (SVM), Neural Network (NN) and RandomForest.

We following typical machine learning development practice by splitting the dataset into training and testing datasets with a ratio of 70:30. Models are trained on the training dataset and their performance are then evaluated and compared using the testing dataset. We are going to illustrate and compare the performance of different algorithms below, and if we do not specify, the performance is always measured on the testing dataset.

Because we have more than 200 variables as predictors, we want to do some rough variable/feature selection. We have used RandomForest method for this purpose. RandomForest method can not only build powerful classifier itself, it can also rank order the predictors by their predictive power or their importance in the forest. In our case, the RandomForest model gives an accuracy ratio of 61%, which is the highest among all the algorithms as we'll see late. In terms of variable selection, usually it works really well in filtering out unimportant variables. The following charts shows the top 12 predictors based on their importance in the underlying forest model. We see that the top predictors are all time related variables, such as day of the week, hour of the day, month of the year and year. This information is very important in business environment, as usually RandomForest is regarded as a blackbox, whereas this information brings out business intuition and gives people clues what the most important features are in the model.

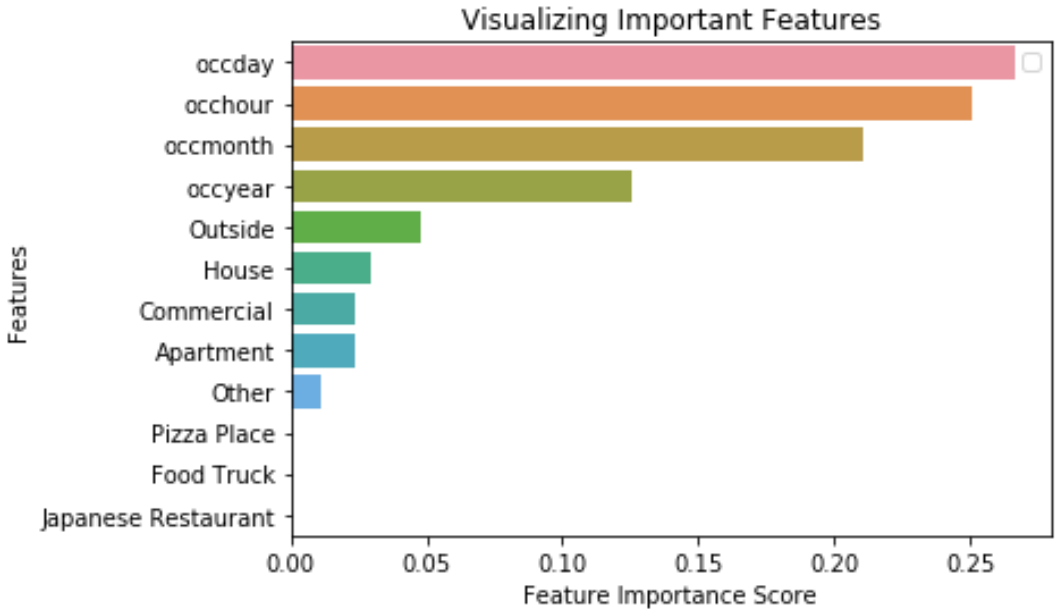

Figure 15 Feature importance from RandomForest method.

To compare the performance of different models, we check the so-called confusion matrix first, as it is rather intuitive. The following chart show the confusion matrix from KNN classifier. The numbers in the diagonal cells are the numbers that are correctly predicted/classified by the algorithm. Certainly, the higher the number in the diagonal the better. Actually, the summation of the numbers in the diagonal cells (655) divided by the total summation of all the numbers (1086) will give the so-called accuracy ratio, which is 55% in this case. It is a rather weak one.

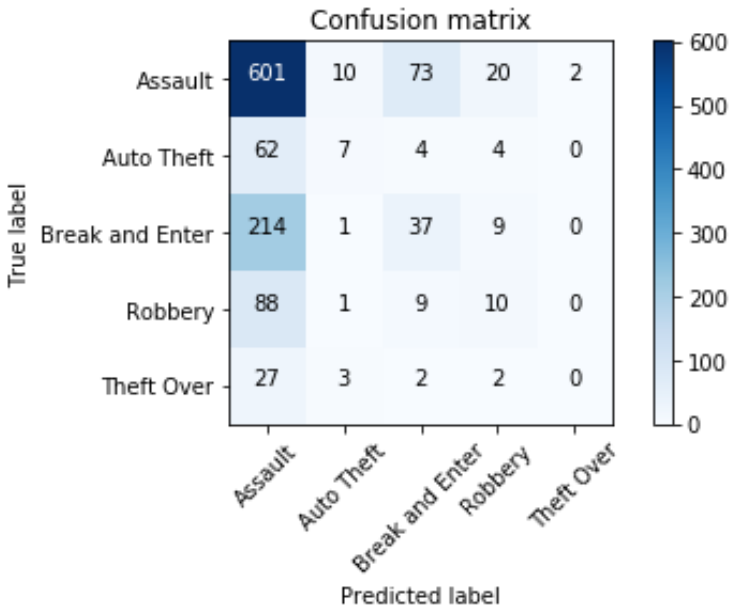

Figure 16 Confusion matrix of KNN method.

We then developed classifiers based on Logistic regression, Decision Tree, SVM and NN. They all give exactly the same prediction without any tuning. They basically predict every incident to be an Assault, which is the highest frequent type of crime. The results itself is certainly not quite useful.

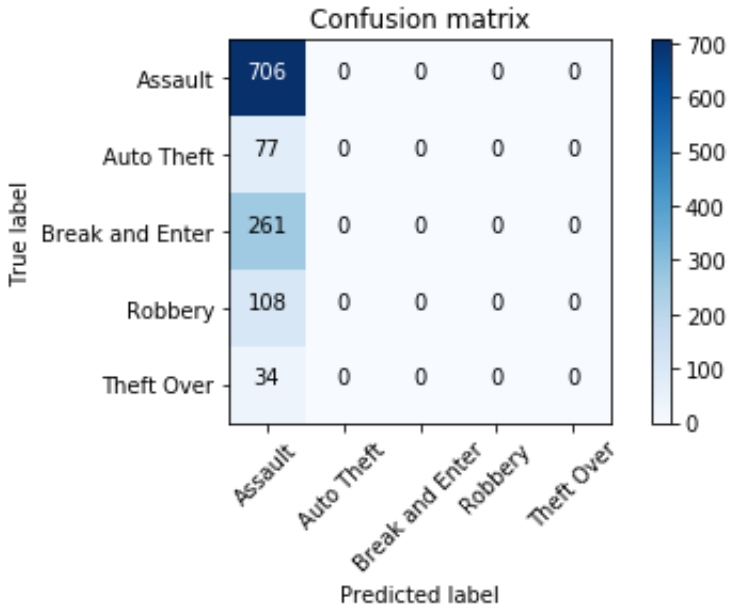

Figure 17 Confusion matrix of Logistic regression, Decision Tree, SVM and NN. Yes, they all share the same confusion matrix.

6 Conclusion

In this project, we have studied the relationship between crime rate/types and commercial activities in Toronto neighborhoods. In particular, we have inspected the relationship/association between certain commercial activities in a neighborhood, represented by the presence as well as density of certain commercial venues in a neighborhood, and the types of crime and overall crime rates in the underlying neighborhood. In order to do so, we have obtained the historical crime statistics from Toronto Open Data Portal, Toronto neighborhood data from Wikipedia, and also commercial venue information from Foursquare. The data were first thoroughly cleaned and then transformed and merged together for further analysis. Different data analytical techniques have been employed to analyze the data, for example, data visualization, descriptive statistics, and more advanced machine learning techniques, such as logistic regression, decision tree, support vector machine (SVM) and neural network (NN). Predictive models have been developed to predict the probability of occurrence of certain crimes with a specific neighborhood, given presence and density of all the commercial venues, and their performances are compared. Overall the model performance is not satisfactory with the best accuracy ratio (achieved by using RandomForest method) being around 62%. However, we did gain extra insight with regard to what might be important factors deciding the types of crimes.