Selecting Toronto Neighborhoods for Pizza Outlets

Abhijit Das

March 16, 2020

1. Introduction

1.1 Background

A pizza chain wants to open outlets in the Toronto neighborhoods, and they are looking for guidance on which neighborhoods might be suitable locations for their restaurants. They have a fixed budget that will allow them to about 10 outlets in the first year, and their management must be able to show traction in the first year as to be able to expand the business in the subsequent years. In their experience in other markets and cities, they have learned that certain locations tend to perform better than others, and locations performing well had some things in common, like they had other comfort food location making it a destination for young crowd. The initial business plan is to follow similar footsteps as in other markets, hence management wants to conduct a study to look for such similarities in the Toronto neighborhoods that could predict their success for new pizza outlets.

1.2 Problem

The management of a pizza chain wants to decide on the locations for new pizza outlets in the Toronto neighborhoods. They want to decide on the new location based on analysis of data on Toronto neighborhoods, the data on existing restaurants that are popular venues of the neighborhoods. After initial review of the Toronto neighborhoods, it was decided that they would segregate the market by the postal codes for the study and target to open one restaurant in one postal code at the maximum. The management needs information on which postal codes or neighborhood would likely be successful locations for new pizza outlets. The management needs the popular restaurants in these postal codes and other parameters of the postal codes to be analyzed to determine the locations that could be successful targets for new pizza locations.

1.3 Interest

The management team of the pizza eatery interested to expand in the Toronto neighborhoods is the primary consumer of this data. The initial interest is to identify the first 10-15 neighborhoods or postal codes that would serve as the first locations to be opened in Toronto. If the initial venture is successful, then further expansion may happen following a similar logic.

The management team needs a list of locations or neighborhoods that would be recommended for opening this pizza outlets based other popular restaurants in the neighborhoods. The management would then select some or all of these location based on the available budget, to launch new pizza outlets.

2. Data Acquisition and Cleaning

2.1 Data Sources

The following datasets to be used for this analysis.

1. The Toronto neighborhood data available publicly on a Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
2. Toronto Geo-Spatial data from (http://cocl.us/Geospatial_data)
3. Location data for Toronto from Foursquare

2.2 Data Cleaning

The Toronto neighborhood data is scrapped from a Wikipedia page. The webpage data loads into a dataframe, the webpage loads as a nested dataframe i.e. a dataframe containing component dataframes. The first dataframe from the webpage contains the Toronto postal code data, this dataframe is extracted into a separate dataframe. The total number of neighborhoods loaded is 287, each row has the columns “PostCode”, “Neighborhood” and “Borough”. As part of the cleaning of the data

- Any null rows are dropped
- Any row having “Borough” as “Not Assigned” is dropped
- Any row having “Neighborhood” as “Not Assigned” is updated such that Neighborhood is assigned the Borough name
- The dataframe is grouped by “PostCode” and “Borough”, if there are multiple neighborhoods they are concatenated into the “Neighborhood” column, using a comma separator
- Two new columns “Latitude” and “Longitude” is added to the dataframe, they are assigned zero value initially

The geo-spatial data for Toronto is loaded into a new dataframe, this provides the postal code and the corresponding latitude and longitude. The original dataframe is populated with the latitude, longitude data by joining with the geo-spatial dataframe.

The “Location Data” from “Foursquare” is extracted using the latitude and longitude tied to the postal codes or neighborhood. The location data for “Food” related venues within a 500 unit radius of the latitude and longitude of the postal code is extracted from “Foursquare”. This location data contains the popular restaurants in those neighborhoods and their categories, this data is transformed into a dataframe. This location data extracted from Foursquare contains the following

- Venue Name
- Venue Latitude
- Venue Longitude
- Venue Category

The venue data contains one row for each venue, hence multiple rows for each neighborhood or postal code. This data has to be cleaned and summarized for processing. After doing some initial analysis it’s found that there are 643 venues in total in all the neighborhoods, and there are 71 unique categories. This data set is transformed for modeling in the following manner

- The 70 unique restaurant venue categories are converted to columns using one-hot encoding
- The neighborhood information is added back to the venues dataframe
- The venues dataframe still contains one row for every venue, hence 604 rows, its further processed as follows
    - The venues dataframe is then grouped by the neighborhood to reduce the dataframe to one row per neighborhood
    - The mean of the occurrence of the venue category is taken as the weight of that category in that neighborhood. Hence the more number of times a venue category occurs in a neighborhood, the higher the weight of that category in that neighborhood
    - The grouping provides one row for each neighborhood with the weight of each venue category in the columns

The 70 venue categories represent a lot of spread, hence I went on to select the following 8 groups of restaurant categories to study the trend. I grouped the 70 categories into one of these groups.

- Pizza Outlet / Italian Cuisine
- North American Cuisine
- Asian Cuisine
- South American Cuisine
- Middle Eastern Cuisine
- European Cuisine
- Healthy Choices

- Comfort Food

The weights of the individual 70 venue categories are added to get the relative weight of these restaurant categories. These 8 restaurant categories weights will be used to build a classification model and then predict the locations that would be good choices for pizza locations.

2.3 Feature Selection

The Toronto neighborhood data after combination with the "Foursquare" location data, contains the following features after the data cleaning exercise.

- Postal Code
- Borough Name
- Neighborhood Name (concatenated and separated by a comma when there are multiple neighborhoods in a Borough)
- Latitude
- Longitude
- Weightage of popular restaurant categories (8 restaurant categories)

The dataframe is clean and contains the required features, this data containing the popular venues will be used to build a classification model. The model is then used to predict the recommended locations for the pizza outlets.

3. Exploratory Data Analysis

3.1 Calculating of target variable

The target variable in the classification of the Toronto neighborhoods by their applicability for new pizza outlets. The criteria and process to classify the postal codes as follows.

The high level criteria for selecting the neighborhoods will be by classification. The following steps will be followed

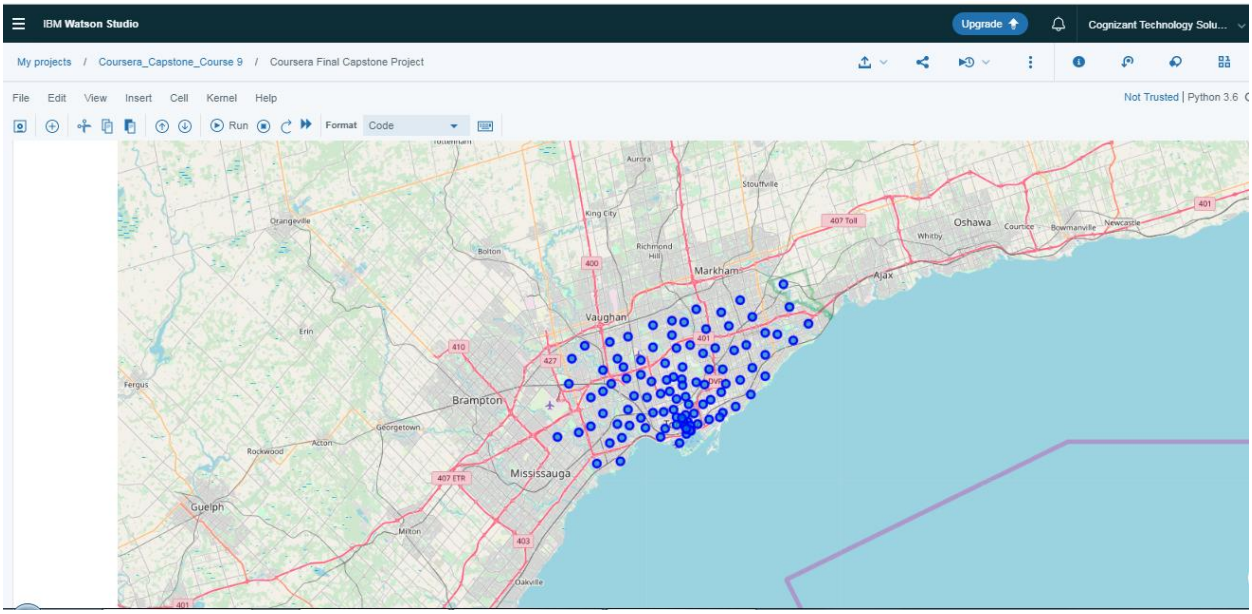
- The neighborhoods or postal codes which have higher rating for pizza outlet are marked as "Recommended" and the ones that have a lower rating are marked as "Not Recommended"
- A subset of this data is taken as the training data
- The training data contains the ratings for all the other venues in that neighborhood
- A classification feature is defined, which will contain "1" to indicate "Recommended" and "0" to indicate "Not Recommended"
- The training data would be used to build a classification model.

- This model is then used to classify the rest of the neighborhoods into either “Recommended” or “Not Recommended’ for pizza outlet.
- The recommended locations are plotted on a map to see the geographical spread

The idea being implemented is that, a neighborhood that currently has “Pizza Outlet” as one of the top venues, has some unique tastes which links them to pizza outlets. The other venues in these neighborhoods influence their affinity for pizza outlets. That is, a venue has “Bowling”, “Pubs” and “Pizza” as popular venues, signifies “Pubs”, “Bowling” and “Pizza” are related, hence classifying based on this would capture this trend. Similarly, a neighborhood that has “Pizza Outlet” as their bottom 10 venues, has some unique tastes which defines the un-popularity of “Pizza Outlets”, the other venues of these neighborhood may influence their un-popularity of “Pizza Outlets”.

3.2 Geographical spread of the Toronto Neighborhoods

To get a perspective of the geographical spread of the Toronto neighborhoods, the neighborhoods are plotted on a map as shown below. This shows that the neighborhoods in the downtown areas are located close to each other i.e. more dense, whereas the ones on the outskirts of Toronto are further apart. This geographic distribution could also reflect in classification provided by the algorithm, the choice of cuisine or type of food may follow a different pattern in the downtown areas versus the outskirts.



3.3 Number of venues in each neighborhood

The number of venues vary over a wide range with the neighborhoods, the following table provides a sample of this variances. The number of venues could potentially be a factor in the classification of neighborhoods as being “recommended” or “not recommended” for pizza outlets. The classification done on the basis of the venues would capture this trend.

Out[30]:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Adelaide,King,Richmond	50	50	50	50	50	50
Berczy Park	50	50	50	50	50	50
Brockton,Exhibition Place,Parkdale Village	24	24	24	24	24	24
Business Reply Mail Processing Centre 969 Eastern	17	17	17	17	17	17
CN Tower,Bathurst Quay,Island airport,Harbourfront West,King and Spadina,Railway Lands,South Niagara	18	18	18	18	18	18
Cabbagetown,St. James Town	47	47	47	47	47	47
Central Bay Street	50	50	50	50	50	50
Chinatown,Grange Park,Kensington Market	50	50	50	50	50	50
Christie	17	17	17	17	17	17
Church and Wellesley	50	50	50	50	50	50
Commerce Court,Victoria Hotel	50	50	50	50	50	50
Davisville	33	33	33	33	33	33
Davisville North	8	8	8	8	8	8
Deer Park,Forest Hill SE,Rathnelly,South Hill,Summerhill West	14	14	14	14	14	14
Design Exchange,Toronto Dominion Centre	50	50	50	50	50	50
Dovercourt Village,Dufferin	14	14	14	14	14	14
First Canadian Place,Underground city	50	50	50	50	50	50
Forest Hill North,Forest Hill West	5	5	5	5	5	5
Harbord,University of Toronto	36	36	36	36	36	36
Harbourfront	47	47	47	47	47	47
Harbourfront East,Toronto Islands,Union Station	50	50	50	50	50	50
High Park,The Junction South	23	23	23	23	23	23
Lawrence Park	3	3	3	3	3	3
Little Portugal,Trinity	50	50	50	50	50	50
Moore Park,Summerhill East	1	1	1	1	1	1
North Toronto West	18	18	18	18	18	18
Parkdale,Roncesvalles	13	13	13	13	13	13
Queen's Park	38	38	38	38	38	38

3.4 Relationship between Pizza outlets and other venues in the neighborhood

After the initial data cleanup and processing, we have a list of postal codes and the weightage of popular restaurant venues in those postal codes. These postal codes are ranked by their weightage of “Pizza” restaurant categories. Some of the neighborhoods that have “Pizza Outlets” as their top venues and some of the neighborhoods that have “Pizza Outlets” as their least frequented venues are taken for training a model, the hypothesis is that the other popular venues for a neighborhood defines their taste as far as popularity of “Pizza Outlets” are concerned. All the venues of a location would be used to train the classification model.

The relationship of popular venues to pizza locations is used to build the classification model. This model then predicts whether other test locations would be popular pizza locations or not i.e. whether the management is recommended to open pizza locations in those postal codes or not.

4. Predictive modelling

This type of problem is suitable for a classification modeling approach as the need is to classify or label the Toronto neighborhoods as either “Recommended” for “Pizza Outlets” or “Not Recommended” for “Pizza Outlets”. Classification is an approach to predict discrete class labels. The second type of predictive modelling is regression, however that is more suitable for continues values.

Classification is a supervised learning approach where we teach the model to categorize some unknown items or data set into discrete classes or categories based on a set of independent variables. The model is built by using a labelled dataset; that is a set of data that has been classified into one of the available categorical values. In this case the labelled data set would be classified as “Recommended” / “Not-Recommended”. The model learns the relationship between a set of independent variables and a dependent target variable i.e. the classification or category. In this case the independent variables are the top venue of each neighborhood and the dependent categorical variable is the “Recommendation” (a “yes” or a “no”).

There are several classification algorithms available, the “K-nearest neighbor” is used to solve the above problem. I think this is the most suitable algorithm to apply to this problem, as the hypothesis is built around the fact that certain venues tend to go along with each other, hence neighborhoods would tend to be related based on the popularity if venues.

4.1 Classification Model

To solve this problem I have applied a “K-Nearest Neighbor” classification algorithm or KNN, as the underlining hypothesis is that neighborhoods that have “Pizza Outlets” as one of their popular venues, also have other common popular venues among them. That is, neighborhoods having “Pizza Outlets” as their popular venues also have “Bowling”, “Italian” as their other popular venues. Hence there is an association between the popular venues of a particular neighborhood, and this will help determine which neighborhoods would be suitable for popular pizza outlets.

To apply the classification model, a new feature “Recommended” is defined, which will contain “1” to indicate “Recommended” and “0” to indicate “Not Recommended”. The independent variables are the popular venues of a neighborhood and the target variable is the classification as “recommended” or “not recommended”.

The training data includes the following.

- Some neighborhoods that have “pizza outlets” in their top venues, which are classified as “recommended” and used to train the model
- Similarly the neighborhoods that have “pizza outlets” in their least frequented venues, which are classified as “not recommended” are used to train the model

The model is finally applied on the entire dataset, to determine which neighborhoods are recommended based on the model.

5. Conclusions

The data analysis process was able to provide a list of neighborhoods or postal codes that are recommended for new pizza outlet locations. This is based on the other “restaurant venues” in the neighborhoods which reflect the taste of the people in that neighborhood or the people who frequent that neighborhood. This offers a data based approach to selecting locations for new pizza businesses, and increases the chances of success for the new restaurants. The following map shows the locations that was selected by the model, the geographical spread seems to suggest that the prediction is balanced and not concentrating on any one area. The spread also ensures that the new outlets are not competing for the same customers.