The Battle of the Neighborhoods

Capstone Project

by
Suresh Jacob

Submitted on 19 May, 2020

Determine location for a new upmarket Sushi Restaurant in the City of Toronto Project

1. INTRODUCTION

A client, who is a wealthy investor, wants to open a fine dining Sushi restaurant in the city of Toronto. As per widely available reports, the popularity of Sushi is growing in Canada, especially in the big cities. The increasing demand for natural food cooked hygienically appeals to the growing health consciousness of consumers. Moreover, there is an accelerating trend among the majority of the population in seeking ethnic cuisines and Sushi is one of the most popular.

The client believes that the Sushi restaurant market is under-represented in Canada in relation to the demand and sees a big potential in getting big returns by investing in this area.

Like all businesses, the returns reflect the initial investment. A fine dining establishment will cost a significant amount of money to set up and therefore the client requires to have the lowest possible pay-back period and maximum ROI. The client is willing to set up the business in any suitable neighborhood in the city of Toronto.

2. BUSINESS PROBLEM

The oft-quoted phrase when it comes to real-estate investment is location,location,location. This applies to customer-facing businesses too. The logic is that higher the foot-traffic in a location, greater the business opportunities. Similarly, for a restaurant to be successful it must be located in the most favorable location.

Traditionally, choosing the best location for a restaurant would involve doing market research by conducting surveys through the phone or internet of a large enough number of potential customers. This requires a lot of time and effort and the data collected is mostly subjective. A truer , quicker and cheaper solution of identifying the best location must be explored.

Choosing the right location will mean quicker pay-back/ROI by ensuring the most business in terms of customers visits and few competition in the same business space (few or no restaurants of the same type).

Author: Suresh Jacob
Date: 19 May 2020

Determine location for a new upmarket Sushi Restaurant in the City of Toronto Project

3. DATASET

The dataset used for analysis for this project was collected from publicly available data.

How the dataset used for analysis was created

The dataset was created by merging the geospatial data containing latitude and longitude data for postal codes (http://cocl.us/Geospatial_data) with scraped data (using the BeautifulSoup4 library) from the wikipedia table for postal codes and neighborhoods ('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:M'); the common key being the postal code. After cleaning, deleting 'non-assigned' postal codes and dropping the postal code column which is not needed for analysis, the final dataset was written out to the file csv.

This file was created for the assignment portion (Segmenting and Clustering Neighborhoods in Toronto) of this Capstone and was reused in the interest of saving a bit of time.

3.1 Original datasets

a) - First source: http://cocl.us/Geospatial_data
The file obtained from the source URL provides a list of postal codes for the City of Toronto (first three letters of the six letter Canadian postal code) and the associated GPS coordinates (latitude and longitude).

b) - Second source: 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:M'
The file obtained from the source URL provides a list of postal codes for the City of Toronto (first letter of the prefix – M) and the associated names for boroughs and neighborhoods.

3.2. Method Overview

a) Download the csv file from the source URL mentioned in section I above containing the postal code prefixes and GPS coordinates.
b) Scrape the table containing the postal code prefixes and associated neighborhood information from the source URL mentioned in section I above.
c) Load the csv data into pandas dataframe.
d) Merge the two dataframes (inner join) with postal code as key.
e) Drop those rows that have ‘Not assigned’ in the Borough column and no data in the Neighborhood column.
f) Reset the index for the dataframe.
g) Ensure that the dataframe is clean.
h) The postal code column can be dropped as it will not be used for analysis.

Author: Suresh Jacob
Date: 19 May 2020

i) Store the cleaned dataframe to csv which will be used to read back into a dataframe as required.

4. Methodology (Description of data analysis)

The methodology section describes the method of

4.1. Exploratory Data Analysis

The data was imported into dataframes as described in section 3.2.

a) The dataframe contained the required attributes of neighborhood and Longitude and Longitude to get the needed venue neighborhood information. Using the information in the dataset, the Foursquare API returned a json file containing the venue information such as name of venue and latitude and longitude of each venue for each neighborhood.
b) The dataset was run through the Folium API to get a geographical representation, ie a map. The map showed that the GPS coordinates were indeed situated in the province of Toronto, centered in the core of the city.
c) The Foursquare API was used again to retrieve nearby venues (limited by 100 to abide by Foursquare requirements).
d) The dataset was narrowed down to those neighborhoods which have the highest number of venues The idea behind choosing those neighborhoods with the highest number of venues is that higher venue density equates directly to higher flow of people traffic.
e) To explore this, bar charts were plotted to see that five neighborhoods had the highest number of venues. They are

I) Commerce Court, Victoria Hotel
II) First Canadian Place, Underground City
III) Garden District, Ryerson
IV) Harbourfront East, Union Station, Toronto Islands
V) Toronto Dominion Centre, Design Exchange

f) The dataset was trimmed down to just the five neighborhoods along with the associated GPS coordinates.
g) Each of the five neighborhoods were explored to see which venues had the highest number including eateries and non-eateries. This will give an idea of the location as to whether people consider it to be a go to area for food.
h) The dataset was then divided into five dataframes, one for each neighborhood to do more exploration.
i) Since the specific type of neighborhood is a Sushi Restaurant, the five neighborhoods were explored for prevalence of that type of restaurant.

Author: Suresh Jacob
Date: 19 May 2020

j) The five datasets were also explored to see which of the five neighborhoods had the most number of eateries. The neighborhood which has the most eateries is clearly one that is the go-to area for people wanting to eat.

4.2 K-means clustering used as a Machine Learning technique

There are many ML techniques but they have to be selectively used depending on the type of dataset that is being analyzed.

The dataset that was obtained from Foursquare contained the features of a Neighborhood (string), the latitude and longitude values for each Neighborhood (float) and Venues (string). There are no other numerical attribute values for the key features (venues)

The dataset therefore does not lend itself to be easily analyzed by training and testing. This leaves clustering techniques such as Density based clustering and K-means clustering. Of the two, K-means is easier to implement and is generally used for Customer segmentation purposes. Since segmentation is what can be done on the dataset, K-means clustering was chosen as the clustering technique for analysis.

In choosing the cluster size the concern was to avoid over-fitting as that would skew the clustering and the end result. As executed, the cluster size of three on five unique neighborhoods produced clear and logical clustering helping to determine the solution.

5. RESULTS

Starting with the ML technique of K-means clustering which was used to do segmentation on the dataset, provided clear direction towards the expected solution.

The algorithm was given the five neighborhoods chosen for the most density of venues to reveal which locations have what types of venues as the most common indicating favored businesses in that area.

From the clustered data, the following can be understood:
The number of clusters are three, 0, 1,2.

Cluster 0 has the following neighborhoods

- Toronto Dominion Centre, Design Centre
- Commerce Court, Victoria Hotel
- First Canadian Place, Underground city

Author: Suresh Jacob
Date: 19 May 2020

Determine location for a new upmarket Sushi Restaurant in the City of Toronto Project

Cluster 1 has the following neighborhoods
- Garden District, Ryerson

Cluster 2 has the following neighborhoods
- Harbourfront east, Union Station, Toronto Islands

5.1 ML Clustering analysis

The cluster data shows:

There are three neighborhoods in Cluster 0 and the top ten venues include;
At least eight (8) food related venues.
The first most common venue is a Coffee shop.

There is one neighborhood in Cluster 1 and the top five venues include;
At least six (6) food related venues.
The first venue is a Clothing store.

There is one neighborhood in Cluster 2 and the top five venues include;
At least five (5) food related venues.
The first venue is a Coffee shop.

Clearly, the clustering process has shown that Cluster 0 is where one must concentrate to open an eatery!

The K-means clustering technique has clearly shown that locations that have eateries in the top five venues are correctly clustered together and the one neighborhood out of the five which has the top most favored/common venue as a 'Clothing Store' is rightly in a separate cluster.

The ML technique has indicated that investing in an eatery of any kind in Cluster 0 neighborhoods will be most optimum. The neighborhoods in Cluster 0 are Toronto Dominion Centre, Design Centre, Commerce Court, Victoria Hotel, First Canadian Place, Underground city.

Author: Suresh Jacob
Date: 19 May 2020

5.2 Further analysis

Exploratory analysis was used to arrive at a singular point, ie the neighborhood that is best suited as a location to open an upmarket Sushi restaurant that will also ensure the best returns on investment.

Tables and associated Bar graphs were used to reveal:

a) Which neighborhoods had the highest number of venues?

The neighborhood of Commerce Court has one of the highest number of venues overall among all the neighborhoods in the City of Toronto. This is an indicator of the density of people traffic which is a prime requirement when looking for high volume of business.

b) Which neighborhoods had the highest number of eateries?

Commerce Court is a venue with the second highest number of eateries and this an indicator that this neighborhood is a preferred choice among people when they are looking for a place to eat.

c) Which neighborhoods had no or few Sushi restaurants?

The data shows that Commerce Court neighborhood has no Sushi Restaurants.

d) Where is the neighborhood located?

Commerce Court neighborhood is located in the financial district of the City of Toronto and we can expect the highest number of well-heeled clientele who have a penchant for fine dining. A Sushi restaurant, first of its type in the neighborhood, marketed as an upscale eatery will be very profitable in this location. And if it matters at all, Commerce Court is only about one km from the lake (Lake Ontario) as shown on the map, which makes for better atmosphere that will be appreciated by the clientele.

6. DISCUSSION

The dataset lends itself fairly well to be analyzed. It is remarkable that with publicly available data like GPS coordinates and postal codes one is able to determine an optimum location to start a business, purely by data exploration and analysis using software.

However, the limited number of attributes of the features did not allow using other ML techniques easily. A dataset that had more numerical data can be more easily manipulated and worked on by the ML algorithms. As a result, the results had to rely

more on knowledge gleaned from exploratory analysis, although the K-means clustering technique did work to point in the right direction.

7. CONCLUSION

The results of Data analysis clearly gives maximum confidence to declare to the client that Commerce Court is the best possible choice of location to start their dream Sushi Restaurant.

The Commerce Court Neighborhood is the winner of the battle of the Neighborhoods!

Author: Suresh Jacob
Date: 19 May 2020