Opening a New Indian Restaurant in Toronto, Canada

By-Jayesh Sharma

October'2019

Introduction

Toronto is usually crowned as most diverse city in the world. Cultural groups from various Countries are migrating to Canada, especially, Indians who are migrating roughly 25,000–30,000 each year (which is now the second-most populous cultural group immigrating to **Canada** each year, behind Chinese immigrants who are the largest group).

Generally, Indians struggle to find their local foods in foreign countries and end-up to cook their foods on their own. So, this is a huge opportunity for the potential investors/businessmen to fulfil the needs of the Immigrants and make their profit out of it. But location of restaurant plays very important role, in whether the restaurant will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Toronto, Canada to open a new Indian Restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Toronto, Canada, if an Investor/businessmen is looking to open a new Indian Restaurant, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new Indian Restaurants in the most Diversed city of world i.e. Toronto.

Data

To solve the problem, we will need the following data:

• Demographics of neighbourhoods in Toronto -This defines the scope of this project which is confined to the city of Toronto, Canada.

• Latitude and longitude coordinates of those neighbourhoods-This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to Restaurants-We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods) contains a list of neighbourhoods in Toronto, with a total of 174 neighbourhoods and list of second most common language in respective neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Restaurant typically Indian, Asian, Fast-Food category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Toronto. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data and the second most common language. Then, using names of 22 official languages that are used in India, we will select the neighbourhoods having majority in them.

Now,We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform

a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Toronto. (as shown below)

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 1500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Indian Restaurant” data, we will filter the “Indian/Asian/Fast-Food Restaurant” as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the

simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Restaurants”. The results will allow us to identify which neighbourhoods have higher concentration of Indian Restaurants while which neighbourhoods have fewer number of Restro’s. Based on the occurrence of Restro’s in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Restro’s.

Results

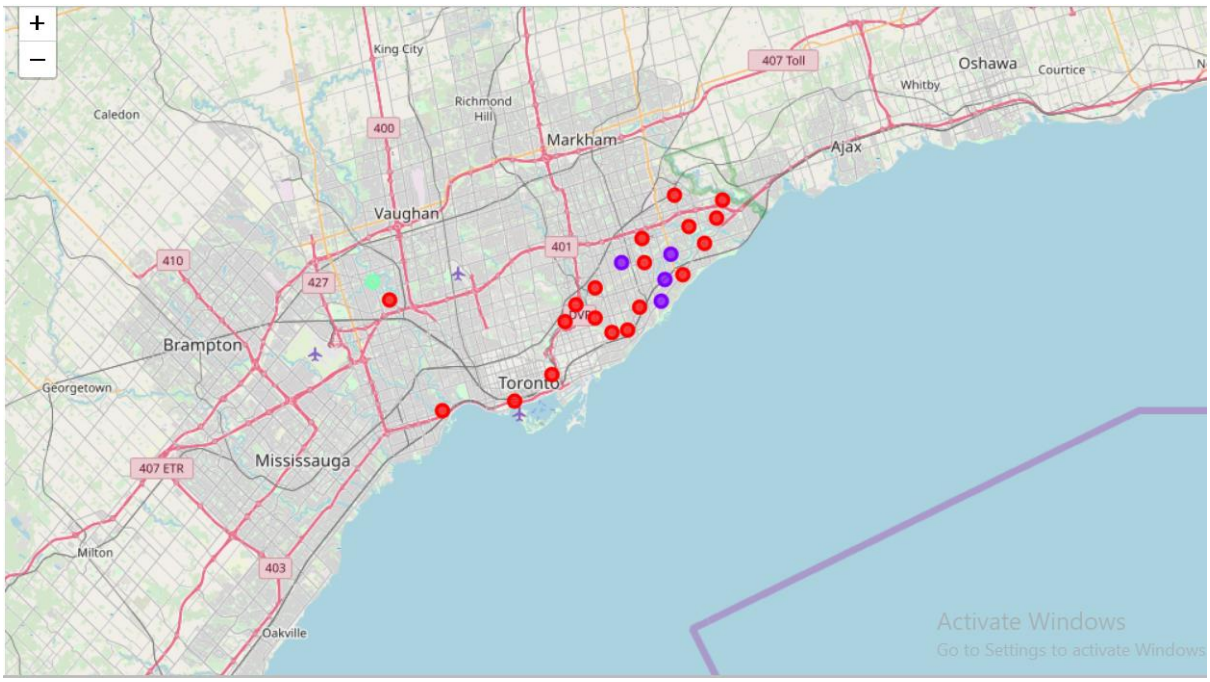
The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Restaurants”:

• Cluster 0: Neighbourhoods with low number to no existence of Indian/Asian Restaurant

• Cluster 1: Neighbourhoods with moderate number of Restaurant

• Cluster 2: Neighbourhoods with high concentration of Indian Restaurants.

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

Most of the Restaurants have highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to totally no Indian/Asian Restaurant in the neighbourhoods. This represents a great opportunity and high potential areas to open new Indian Restro as there is very little to no competition from existing Restro's. Meanwhile, Indian Restro's in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Restaurants. Therefore, this project recommends Investors to capitalize on these findings to open new Indian Restaurants in neighbourhoods in cluster 0 which have high percentages of **Indian Languages** like in Crescent Town, Humberwood etc. Property developers with unique selling propositions to stand out from the competition can also open new Restaurants in neighbourhoods in cluster 1 with moderate competition. Lastly, Investors are advised to avoid neighbourhoods in cluster 2 which already have high concentration of Indian Restaurant and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider the factor of frequency of occurrence of Indian Restaurants and Second Most common language, there are other factors such as population, diversity and income of residents that could influence the location decision of a new Restaurant. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new Indian Restaurant. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new Indian Restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 with the highest common Indian languages are the most preferred locations to open a new Indian Restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Indian Restaurant.

APPENDIX-

Cluster 0-

Bendale
Crescent Town
Flemingdon Park
Fort York/Liberty Village
Highland Creek
Humberwood
Malvern
Morningside
O'Connor-Parkview
Oakridge
Regent Park/Trefann Court
Rouge
Scarborough City Centre
Scarborough Junction
Scarborough Village
Smithfield
Thorncliffe Park
Victoria Village
West Hill

Cluster 1-

Cliffcrest
Dorset Park
Eglinton East
Woburn

Cluster 2-

Thistletown