Trying to find a good location to open a Indian restaurant using Data Science and ML Techniques

Introduction

According to a study of cuisine trade, Indian food is the fourth most popular cuisine the world. Its simplicity resulted in its popularity and it is a very diverse cuisine. Most of these restaurants abroad have been opened by the native Indians who have settled abroad and do not want to loose touch with their food and culture, but surprisingly these restaurants are visited more by the international customers than the native Indians. However there are not many Indian restaurants out there in certain places and these places tend to miss an opportunity to explore a beautiful, colorful and rich cuisine. In this project our goal is to find a great location to open an Indian restaurant where it would gain popularity amongst the locals but at the same time not run on loss as the cuisine is not really popular or likeable there.

Business Problem

The objective of this project is to find the perfect place to open an Indian restaurant in Toronto, Canada. Though Toronto has a lot of Indians, there are not many Indian restaurants, I am put in a situation where I want to find suitable locations to open my Indian restaurant. My intention is to find the answer to my business problem “Where should I open my indian restaurant?”. In this capstone project, with the help to data science techniques and machine learning methods (like clustering), we intend to find a solution to our business problem.

Target Audience

The audience that we are targeting are all the Indian households or entrepreneurs who want to open or expand their Indian restaurant in great places as it is a great opportunity for not just entrepreneurs but stay-at-home parents or partners who are not working but are looking for means to earn.

Data Description

To solve this problem, I would be requiring the following data:

1. List of neighborhoods in Toronto. This defines the scope of this project which is confined to the city of Toronto, Canada.
2. Coordinates of these neighborhoods. This is required in order to plot the map and also to get the venue data.
3. Venue data related to our business. We will use this data to perform clustering on the neighbourhoods.

Sources of Data

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of neighbourhoods in Toronto Canada with a total of 103 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page. Then we will get the geographical coordinates of the neighbourhoods to get the latitudes ad longitudes of the neighbourhood.

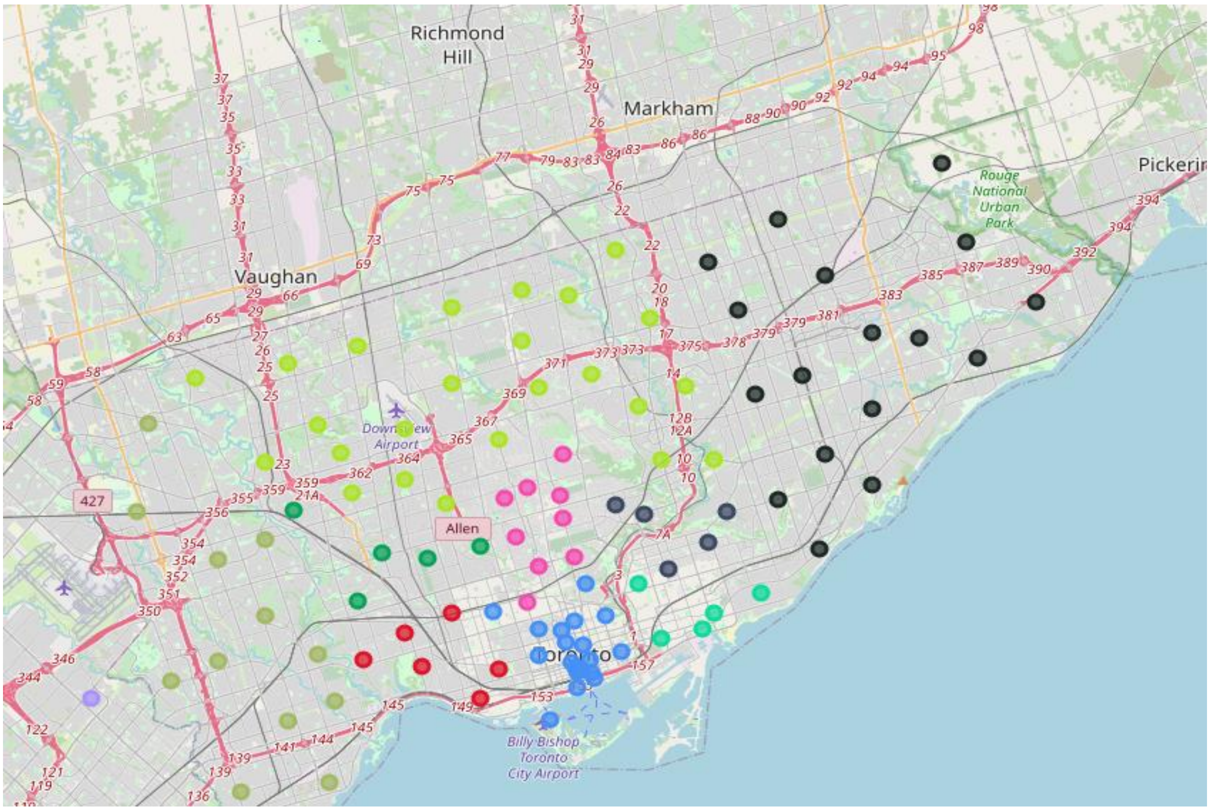
This project will also use Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

This project makes 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).

Methodology

First, I need to get the list of neighbourhoods in Toronto, Canada. This is possible by extracting the list of neighbourhoods from Wikipedia page that is mentioned in the source of data section of this report. I performed the web scraping by utilizing pandas html table scraping technique as it is easier and more convenient to pull tabular data directly from a web page into dataframe.

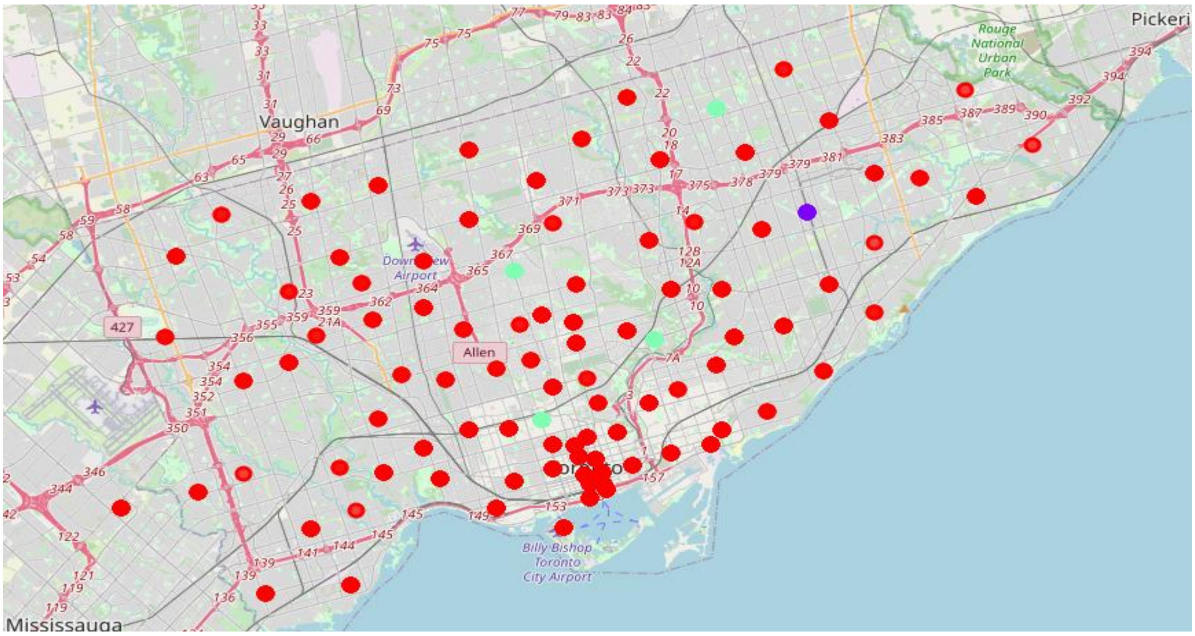
However, it is only a list of neighbourhood names and postal codes. I will need to get their coordinates to utilize Foursquare to pull the list of venues near these neighbourhoods. To get the coordinates, I used the csv file provided by IBM team to match the coordinates of Toronto neighbourhoods. After gathering all these coordinates, I visualized the map of Toronto using Folium package to verify whether these are correct coordinates.



After that, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analysed each neighbourhood by grouping the rows by neighbourhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for “Indian restaurants”. Lastly, I performed the clustering method 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 it is highly suited for this project as well. I have clustered the neighbourhoods in Toronto into 3 clusters based on their frequency of occurrence for “Indian food”.

With the help of the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.



Results

The results from k-means clustering show that we can categorize Toronto neighbourhoods into 3 clusters based on how many Indian restaurants are in each neighbourhood:

- Cluster 0: Neighbourhoods with high number of Indian restaurants
- Cluster 1: Neighbourhoods with little to no Indian restaurants
- Cluster 2: Neighbourhoods with moderate number of Indian restaurants

The results are visualized in the above map with Cluster 0 in red colour, Cluster 1 in purple colour and Cluster 2 in mintgreen colour.

Recommendations

Most of Indian restaurants are in Cluster 0 which is around Adelaide, King, Richmond areas and lowest (close to zero) in Cluster 1 areas which are Dorset Park, Wexford Heights. Also, there seems to be good opportunities to open an Indian near Bedford Park, Lawrence Manor East or The Annex, North Midtown, Yorkville as the competition seems to be low. Upon taking a look at the venues, it seems that Cluster 1 might be a good location as there are not a lot of Indian restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Indian restaurant in these locations with little to no competition. Nonetheless, if the food is authentic, affordable and good taste the restaurant would gain fame anywhere.

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. entrepreneurs or households planning to open a restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new Indian restaurant as there is some healthy competition there. The findings of this project will help the entrepreneurs to make use of the opportunity to high potential locations while avoiding overcrowded areas in their decisions to open a Indian restaurant.