Using Machine Learning to find locations to open an Indian Restaurant

A.B.M. Fahim Shahriar
abmfahimshahriar@iut-dhaka.edu

June 2, 2020

1 Introduction

1.1 Background

For this Capstone project, I am creating a hypothetical scenario for a concept Indian guy who wants to open an authentic Indian restaurant in Toronto area. The idea behind this project is that there may not be enough Indian restaurants in Toronto and it might present a great opportunity for this entrepreneur who is based in Canada. As Indian food is very popular around the world and is similar to other Asian cuisines, this entrepreneur is thinking of opening this restaurant in locations where Asian food is popular.So,finding the location to open such a restaurant is one of the most important decisions for this entrepreneur.So, I am designing this project to help him find out the most suitable location where he can make good profit out of his business.

1.2 Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Indian restaurant in Toronto, Canada. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open an Indian restaurant, where should they consider opening it?

1.3 Target Audience

The entrepreneur who wants to find the location to open authentic Indian restaurant

2 Data

To solve this problem, I will need below data :

- List of neighborhoods in Toronto, Canada.
- Latitude and Longitude of these neighborhoods.
- Venue data related to Asian restaurants. This will help us find the neighborhoods that are most suitable to open an Indian restaurant.

3 Extracting Data

- Scrapping of Toronto neighborhoods via Wikipedia
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

4 Methodology

Firstly, I got the neighborhood data of Toronto, Canada from scraping the following webpage. (`https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M`) I did the web scraping using pandas html table scraping method since it is easier and more convenient for this case. I only got the names of neighborhood of Toronto. So to work with the data, I used the CSV provided by IBM to get the co-ordinates of the neighborhoods and then merged the two dataframes. After that I visualized the data using Folium package and generated map of neighborhoods of Toronto and verified the co-ordinates.

Next, Using Foursquare API call to get the list of top 100 venues within 500 meters radius. I opened a developer account in Foursquare API and using my credentials, I submitted the API call. From Foursquare, I was 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 analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

In the data, from venue category I looked for 'Thai restaurants' as Indian and Thai food both are from same continent that why I think they share some common property. So, if I look for the neighborhoods with Thai restaurants I think I can predict the preferred neighborhood for the Indian restaurant. Next, I applied K-Means clustering algorithm on the data with 3 cluster numbers. K-Means is a very popular clustering algorithm for scenarios like this one. Based on the result I will be able to recommend neighborhood where the Indian restaurant can flourish well. The mean frequency in each cluster refers to how much 'Thai restaurants' are there around the neighborhood. So, if a cluster has less 'thai restaurant' frequency then there is less competition and an Indian restaurant can do quite a business there.

5 Result analysis

This picture shows the neighborhood before clustering:

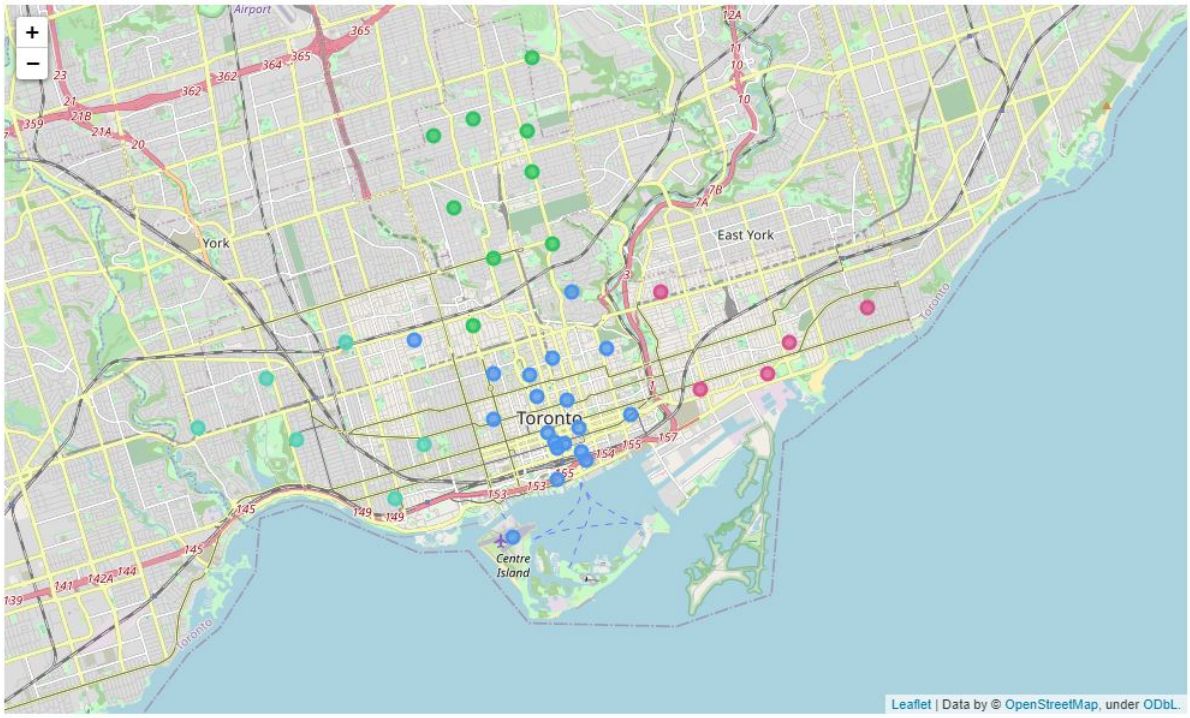

Figure 1: Neighbors before clustering

This picture show neighborhoods after clustering:

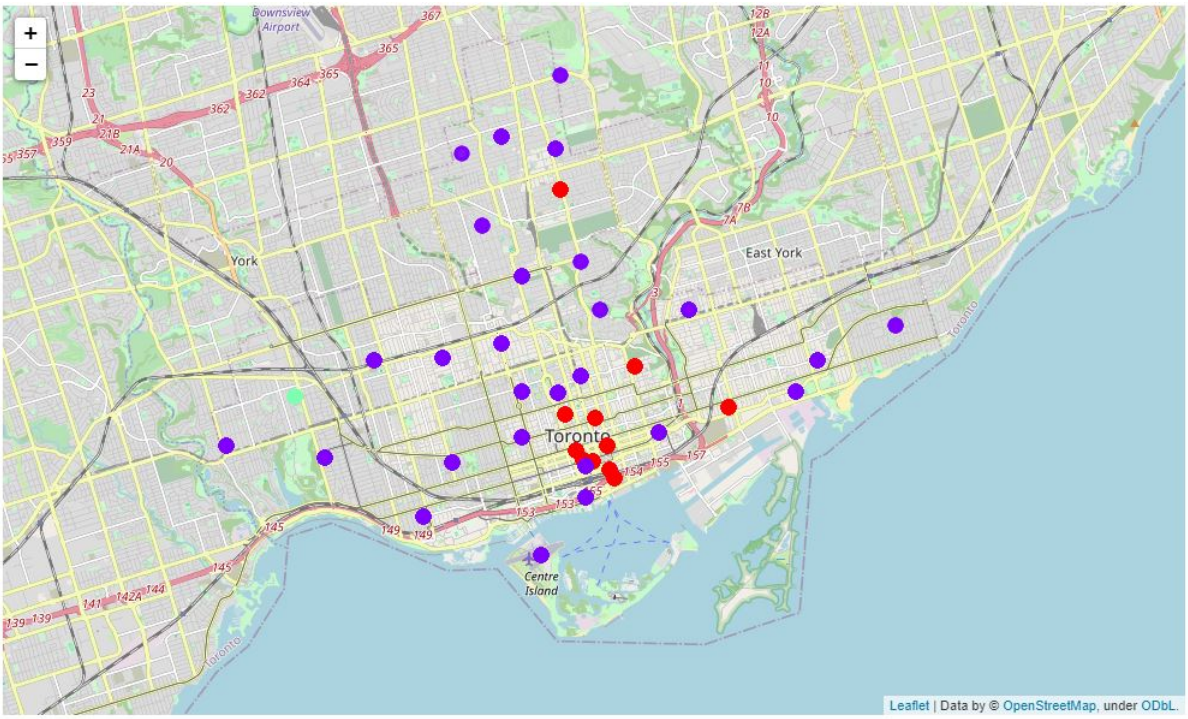

Figure 2: Neighbors after clustering

Now lets see the cluster dataframes to understand better.

	Neighborhood	Thai Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Berczy Park	0.017241	0	43.644771	-79.373306	LCBO	43.642944	-79.372440	Liquor Store
32	Studio District	0.024390	0	43.659526	-79.340923	Avling Kitchen & Brewery	43.661515	-79.338117	Brewery
32	Studio District	0.024390	0	43.659526	-79.340923	EAT BKK Thai Kitchen	43.660450	-79.343113	Thai Restaurant
32	Studio District	0.024390	0	43.659526	-79.340923	Chez Nous	43.659640	-79.346900	Wine Bar
32	Studio District	0.024390	0	43.659526	-79.340923	Boxcar Social	43.659723	-79.346871	Café
32	Studio District	0.024390	0	43.659526	-79.340923	Starbucks	43.660742	-79.342362	Coffee Shop
32	Studio District	0.024390	0	43.659526	-79.340923	Wonder Pens	43.662809	-79.341452	Stationery Store
32	Studio District	0.024390	0	43.659526	-79.340923	eastside social	43.661289	-79.339155	Comfort Food Restaurant
32	Studio District	0.024390	0	43.659526	-79.340923	Tabule	43.659731	-79.346341	Middle Eastern Restaurant
32	Studio District	0.024390	0	43.659526	-79.340923	Purple Penguin Cafe	43.660501	-79.342565	Café

Figure 3: Cluster 0

	Neighborhood	Thai Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
37	Toronto Dominion Centre, Design Exchange	0.0	1	43.647177	-79.381576	Starbucks	43.650159	-79.377793	Coffee Shop
37	Toronto Dominion Centre, Design Exchange	0.0	1	43.647177	-79.381576	Toronto Maple Leafs Hockey Club	43.645786	-79.379803	General Entertainment
37	Toronto Dominion Centre, Design Exchange	0.0	1	43.647177	-79.381576	Druxy's	43.648015	-79.379907	Deli / Bodega
37	Toronto Dominion Centre, Design Exchange	0.0	1	43.647177	-79.381576	Over Easy	43.648077	-79.377599	Breakfast Spot
37	Toronto Dominion Centre, Design Exchange	0.0	1	43.647177	-79.381576	Fast Fresh Foods	43.647708	-79.379549	Salad Place
37	Toronto Dominion Centre, Design Exchange	0.0	1	43.647177	-79.381576	Szechuan Express	43.646973	-79.379549	Chinese Restaurant
37	Toronto Dominion Centre, Design Exchange	0.0	1	43.647177	-79.381576	Spirit of Hockey	43.647047	-79.377537	Sporting Goods Shop

Figure 4: Cluster 1

	Neighborhood	Thai Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	famous last words	43.665181	-79.468471	Speakeasy
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	Lithuania Park	43.658667	-79.463038	Park
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	nodo	43.665303	-79.465621	Italian Restaurant
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	Indie Alehouse	43.665475	-79.465290	Gastropub
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	Hole in the Wall	43.665296	-79.465118	Bar
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	Junction Flea	43.665258	-79.462868	Flea Market
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	ARTiculations	43.665550	-79.467194	Arts & Crafts Store
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	Mjölk	43.665432	-79.467962	Furniture / Home Store
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	Isaan Der	43.665311	-79.468078	Thai Restaurant
15	High Park, The Junction South	0.083333	2	43.661608	-79.464763	Junction City Music Hall	43.665334	-79.466253	Music Venue

Figure 5: Cluster 2

From the. cluster table we can conclude that,

- ”Cluster 0” has few thai restaurants in the corresponding neighborhoods.
- ”Cluster 1” has no thai restaurants in the corresponding neighborhoods.
- ”Cluster 2” has most thai restaurants in Toronto.

6 Result Discussion

Most of Thai restaurants are in Cluster 2 which is around High Park, the Junction South and none in Cluster 0 areas which are South district, Berczy Park. areas. There are few restaurants in cluster 1 near Toronto dominion center, Design exchange.

So, to open an Indian restaurant, I think the neighbors of cluster 1 would be much better as there is less competition. Moreover, the neighborhoods of cluster 0 has no thai restaurant and one might think it would be a good opportunity to open Indian restaurant there as there is no competition. But this might be due to the fact that Asian food is not popular in these part of Toronto. So, I think it would not be wise to open the restaurant in cluster 0 neighborhoods. Again, since in cluster 2 number of thai restaurants is higher, so the competition is also high there. So, it would not be a profitable idea to open a new restaurant there as people are less likely to try new restaurants over old ones.

So, I would recommend to open the Indian restaurant in the neighborhoods of cluster 1.

7 Future work

In this project, I used only thai restaurant category to cluster the neighbors and there might be other Asian restaurants in the neighborhoods of Toronto. In future work, I might try using the other venue categories as well and analyze how it affects the cluster and so on. I might also take into consideration about the density of Asian population in the neighborhoods of Toronto.

8 Conclusion

In this project, firstly we identified the business problem and specified the data required to solve the problem. After that, we pre-processed the data and then applied machine learning algorithm(K-Means clustering) and thus recommending the stakeholder where to open his Indian restaurant.

9 References

List of neighborhoods in Toronto:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M Foursquare Developer Documentation:
https://developer.foursquare.com/docs