Coursera IBM Data Science Capstone Project : Opening a new Indian Restaurant in Toronto

Report Prepared by :
Abhinav Chandan
(www.github.com/abhinavchandan)

TABLE OF CONTENTS

Introduction :

For the Capstone project, I am creating a hypothetical scenario for a concept Indian restaurateur who wants to explore opening an authentic purely vegetarian Indian restaurant in Toronto area. The idea behind this project is that there may not be enough vegetarian Indian restaurants in Toronto and it might present a great opportunity for this entrepreneur who is based in Canada. A high number of Indians are vegetarians and prefer not to eat in restaurants where non-vegetarian food is served. With the purpose in mind, finding the location to open such a restaurant is one of the most important decisions for this entrepreneur and I am designing this project to help him find the most suitable location.

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 ansIr the business question: In Toronto, if an entrepreneur wants to open a vegetarian Indian restaurant, where should they consider opening it?

Data Requirement :

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 Indian restaurants. This will help us find the neighborhoods that are most suitable to open a vegetarian Indian restaurant.

Data Collection:

- 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 neighbourhoods.

Methodology :

First, the list of neighborhoods in Toronto is obtained. This is done by extracting the list of neighborhoods from Wikipedia page:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

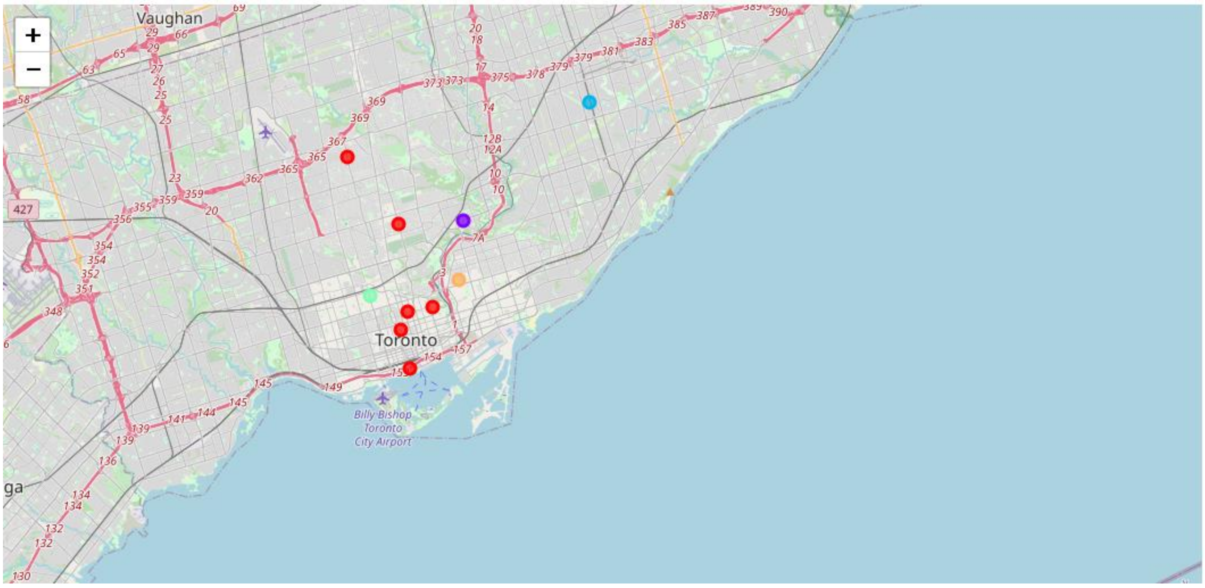
The methodology used for web scraping was utilizing pandas html table scraping method as it is more convenient to pull tabular data directly from a web page into dataframe. However, it is only contains a list of neighbourhood names and postal codes. Next, we need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, we using Geocoder package but it was not functional so the csv file provided by IBM team to match the coordinates of Toronto neighborhoods was used. After gathering all these coordinates, the map of Toronto was visualized using Folium package to verify whether these are correct coordinates.

Next, we use Foursquare API to pull the list of top 100 venues within 500 meters radius. We have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, we can obtain the names, categories, latitude and longitude of the venues. With this data, we can also check how many unique categories that we can get from these venues. Then, we 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

Here, we look specifically for the restaurants. Once we have a dataset containing only restaurants, we can identify clusters where many restaurants are located. Previously, we tried running the model, I was looking only for "Indian restaurants" but there are very few results (maybe due to Foursquare categorization) so we instead try to find restaurant hotspots where Indian Restaurants are popular. Finding such instances will help us in determining locations where we can open a Veg Indian Restaurant. Lastly, we feed the data into the k-means clustering. K-means clustering algorithm identifies k number of centeriods, 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. We have clustered the neighborhoods in Toronto into 5 clusters based on their frequency of occurrence for restaurants in neighborhoods were Indian Restaurants were more in number. This will help us in maximizing customers just on the basis of restaurant location. Based on the results (the concentration of clusters), We will be able to recommend the ideal location to open the restaurant.

Results :

Clusters



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

- Cluster 0: Neighborhoods with low popularity Indian Restaurants
- Cluster 1: Neighborhoods with highly popular Indian Restaurants and Asian Cuisine Restaurants (Asian, Chinese & Thai)
- Cluster 2: Neighborhoods with highly popular Indian Restaurants
- Cluster 3: Neighborhoods with moderately - highly popular Indian Restaurants and highly popular Vegetarian Restaurants
- Cluster 4: Neighborhoods with moderately popular Indian Restaurants

The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple color, Cluster 2 in light blue color, Cluster 3 in green color and Cluster 4 in orange color

Recommendations :

. There are good opportunities to open in Cluster 1(Thorncliffe Park) and Cluster 2 (Dorset Park, Wexford Heights, Scarborough Town Centre). Looking at nearby venues, it seems Cluster 1 might be a good location as there are not a lot of Asian restaurants in these areas. A great location to open Veg Indian restaurants will be in Cluster 3 (The Annex, North Midtown, Yorkville). Therefore, this project recommends the entrepreneur to open an authentic Indian restaurant in these locations with visitors interested in both Vegetarian Food and Indian Food. Nonetheless, if the food is authentic, affordable and nutritious.

Limitations and Suggestions for Future Research :

In this project, we only take into consideration of one factor: the number of Indian restaurants in each neighborhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a new restaurant. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration of these factors. In addition, I am relying on the existence of Thai restaurants only for this project but future research can take into consideration of other variables such as the population level of Indians in each neighborhood etc.

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 the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.

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