IBM PROJECT FINAL

IBM Final Report

ANKUR KHANDELWAL

Contents

Introduction:

For this Capstone project, I am creating a hypothetical scenario for an entrepreneur who wish to open an authentic Nepalese restaurant in Toronto area. He wishes to consider this opportunity based on data science. The idea behind this project is that there may not be enough Nepalese restaurants in Toronto and it might present a great opportunity for this entrepreneur who is based in Canada. As Nepalese food is very similar to other Asian cuisines, this entrepreneur is thinking of opening this restaurant in locations where Asian food is popular (that is, the area with huge number of Asian restaurants in density). 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 such a person find the most suitable location based on data science.

Business Problem:

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Nepalese restaurant in Toronto, Canada. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to analyse the business question:

In Toronto, if an entrepreneur wants to open a Nepalese restaurant, where should they consider opening it?

Target Audience:

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

Data:

To solve this problem, I will need below data:

1) *List of neighbourhoods in Toronto, Canada.*
2) *Latitude and Longitude of these neighbourhoods.*
3) *Venue data related to Asian restaurants.*

This will help us find the neighbourhoods that are most suitable to open a Nepalese restaurant.

Extracting the Data:

1) *Scrapping of Toronto neighbourhoods via Wikipedia.*
2) *Getting Latitude and Longitude data of these neighbourhoods via Geocoder package.*
3) *Using Foursquare API to get venue data related to these neighbourhoods*

Methodology:

First, I need to get the list of neighbourhoods in Toronto, Canada. The possible way to extract the list of neighbourhoods from Wikipedia page (" https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M "). I did the web scraping by utilizing pandas html table scraping method as it is easier and more convenient to pull tabular data directly from a web page into data frame. 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 tried using Geocoder package but it was not working so 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.

Next, 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 analyse 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 "Thai restaurants". Previously, when I ran the model, I was looking for "Asian restaurants" but there are very few results (maybe due to Foursquare categorization) so I looked for the restaurants closest to Nepalese cuisine taste (side note: Nepalese food and Thai food are very similar in taste, so my justification is that if there are people who enjoyed Thai food, they likely are going to enjoy Nepalese food too!)

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 "Thai food". Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurants.

Obtained Results:

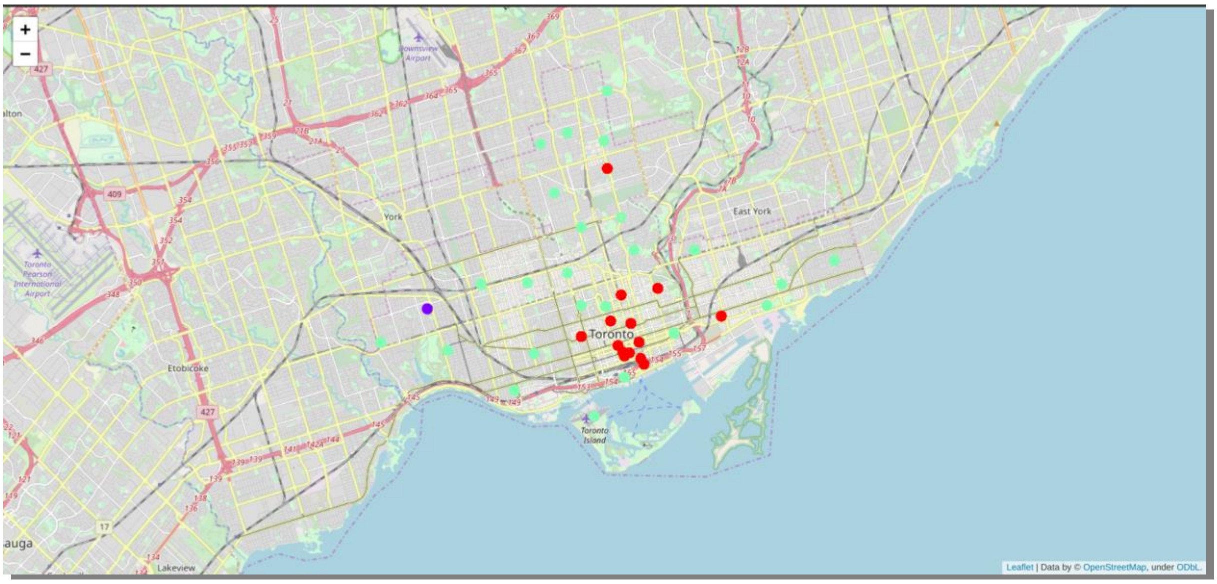
+
−
York
East York
Toronto
Etobicoke
Toronto Pearson International Airport
Downsview Airport
Toronto Island
Lakeview
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig. 1 : Displaying Clusters.

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

1. *Cluster 0: Neighbourhoods with little or no Thai restaurants.*
2. *Cluster 1: Neighbourhoods with no Thai restaurants.*
3. *Cluster 2: Neighbourhoods with high number of Thai restaurants*

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

Recommendations:

Limitations and Suggestions for Future Research:

In this project, I only take into consideration of one factor: the occurrence / existence of Thai restaurants in each neighbourhood. 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 existence of Asian restaurants, Asian population level in each neighbourhood 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:

1. *List of neighbourhoods in Toronto:* https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2. *Foursquare Developer Documentation:* https://developer.foursquare.com/docs