Using Machine Learning to open an Ice-Cream shop in the areas of Toronto, Canada

Introduction:

For the Capstone project, I'm considering a hypothetical scenario wherein there is a start-up Ice-Cream business which has been recently setup in Canada. They are placed in Toronto and want to experiment with the idea of setting up an Ice-Cream parlour in the area of Toronto, Canada. This new Ice-Cream venture start-up is unique as it uses fresh vegetables, pulses and grains combined with fruits to make a premium and nutritious ice-cream. As Canadian's are health conscious and the company aims at making delicious yet nutritious ice-cream products, they want to open a parlour in Toronto, Canada since they are residing near the area. Also, the start-up does not have sufficient resources and funds to compete with big ice-cream retailers and don't want to partner with any major ice-cream outlet yet. So, an ideal location for the ice-cream parlour is of extreme importance for the future of the young Ice-Cream start-up.

Aim of the Ice Cream start-up:

To open its first ice cream outlet in an **ideal location in Toronto, Canada** to test their product's popularity among masses and decide the future of the product.

Business Problem:

Finding the ideal location for the Ice-Cream start-up to open a new parlor in Toronto, Canada. By using data science techniques and machine learning such as clustering, this project aims to provide solutions to business problem at hand: If an Ice-Cream start-up wants to open an Ice Cream parlor , where areas would be most suitable in Toronto, Canada?

Target Audience:

The ice-cream start-up wants to mainly target middle-aged to older adults with more nutritious options of ice-cream which are tasty yet are healthy. In general it caters to all the sections of the society.

Data:

To solve this problem, I will need below data:

- Neighbourhoods in Toronto, Canada.
- Latitude and Longitude of these neighbourhoods.
- Venue data related to Ice-Cream.

This will help us find the neighbourhoods that are most suitable to open an ice-cream parlour.

Extracting the data:

- Web Scrapping of Toronto neighbourhoods, Canada via Wikipedia
- Getting the Latitude and Longitude data of these neighbourhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighbourhoods

Methodology:

Firstly, get the list of neighbourhoods in Toronto, Canada. This is possible by extracting the list of neighbourhoods from Wikipedia page:
("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M")

Perform web scraping on the Wikipedia page by utilizing BeautifulSoup package and then converting the table in the Wikipedia page to a pandas dataframe. However, it is only a list of neighbourhood names and postal codes. Next, I got the coordinates of the places in Toronto, Canada by using the Foursquare API to pull the list of venues near these neighbourhoods. To get the coordinates, initially I tried Geocoder package it failed to work on my machine. So, I used the coordinates from the csv file provided by IBM to match the coordinates of Toronto neighbourhoods. After gathering all the 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.

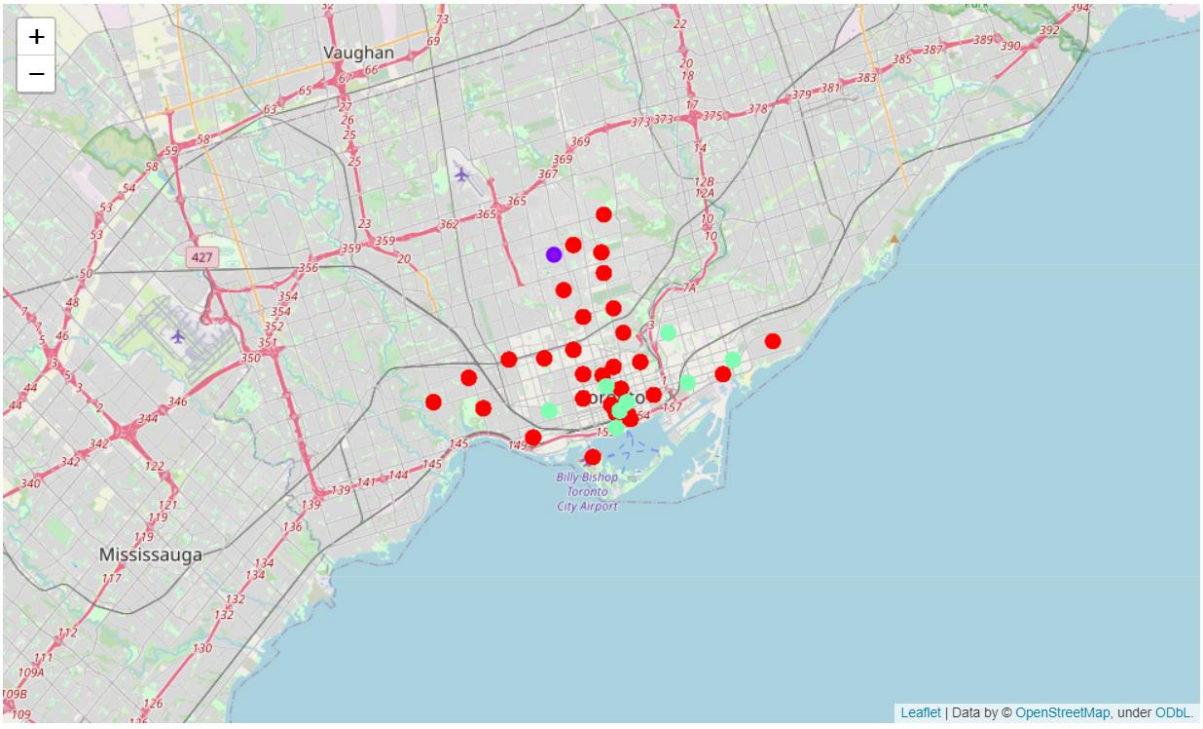
In order to use the Folium API I had to create a Folium Developers Sandbox account which is free. I then used my account ID and API key to pull the data. From Foursquare API, I was able to pull the names, categories, latitude and longitude of the venues. With this data, I checked how many unique categories 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. Next, I specifically looked for "Ice-Cream".

Lastly, I performed the clustering by using k-means. 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 a popular unsupervised machine learning algorithm and ideal for this project.

I have clustered the neighbourhoods in Toronto into 3 clusters based on their frequency of occurrence for “Ice-Cream”. Based on the results (the concentration of clusters), I can recommend the ideal location to open the parlor.

Results:



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

- Cluster 0: Large Neighbourhoods with little or no Ice-Cream parlours
- Cluster 1: Small Neighbourhoods with ample Ice-Cream parlours
- Cluster 2: Large Neighbourhoods with at most 1 ice-cream parlour in each neighbourhood

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:

My take is, that the ice-cream start-up should open its ice-cream parlour in the neighbourhoods of Cluster 0. The advantage is that as there isn't any major ice-cream retailer there will be a great demand for the ice-creams. An added advantage is that there seems to be many cafe, restaurant and entertainment hubs which will give the provide the ice-cream business enough market popularity and exposure and hopefully if the ice-cream is a hit with consumers it will be a major win for the young ice-cream start-up.

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 utilizing k-means clustering and providing a recommendation to the stakeholder.

References:

- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M Foursquare Developer
- Documentation: https://developer.foursquare.com/docs