Machine Learning location evaluation to start an Italian Restaurant in Toronto

1.Introduction

1.1 Background

For this Capstone project, the hypothetical scenario is represented by an Italian who wants to explore opening opportunity of an authentic Italian restaurant in Toronto area.

The idea behind this project is that there may not be enough Italian restaurants in Toronto. Let's say Italian food is similar to other European cuisines: this Italian person might think to open his restaurant in locations where European food is popular (in other words, having many European restaurants in the neighborhood).

With this aim, find the right location to open such a restaurant is one of the most important decisions for this Italian person, and with this project I would like to help him find the most suitable location.

1.2 Business Problem

The objective of this capstone project is to find the most suitable location for the Italian person to open a new Italian restaurant in Toronto, Canada. The question data science methods and machine learning methods such as clustering can help us to have a solution is: in Toronto, if anyone wants to open an Italian restaurant, where should he consider opening it?

1.3 Target Audience

The Italian person who wants to find the best location to open authentic Italian restaurant.

2.Data

To solve this problem, we need the below list of data:

- List of neighborhoods in Toronto, Canada.
- Latitude and Longitude of these neighborhoods.
- Venues data related to European restaurants: this will help us find the neighborhoods that are most suitable to open an Italian restaurant.

3. Extracting Data

- Scraping 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

First of all, I imported all the necessary libraries.

Then, I got the list of neighborhoods in Toronto, Canada by extracting them from Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

The web page was scraped with pandas html table scraping method.

At the end I got a list of neighborhood names and postal codes. I will need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I used Geocoder package: and after gathering all these coordinates, I visualized the map of Toronto using Folium package to verify whether these are correct coordinates.

Next step, I used Foursquare API to pull the list of top 100 venues within 500 meters radius. Earlier I created a Foursquare developer account in order to obtain account ID and API key to pull the data.

From Foursquare, I was able to get the names, categories, latitude and longitude of the venues.

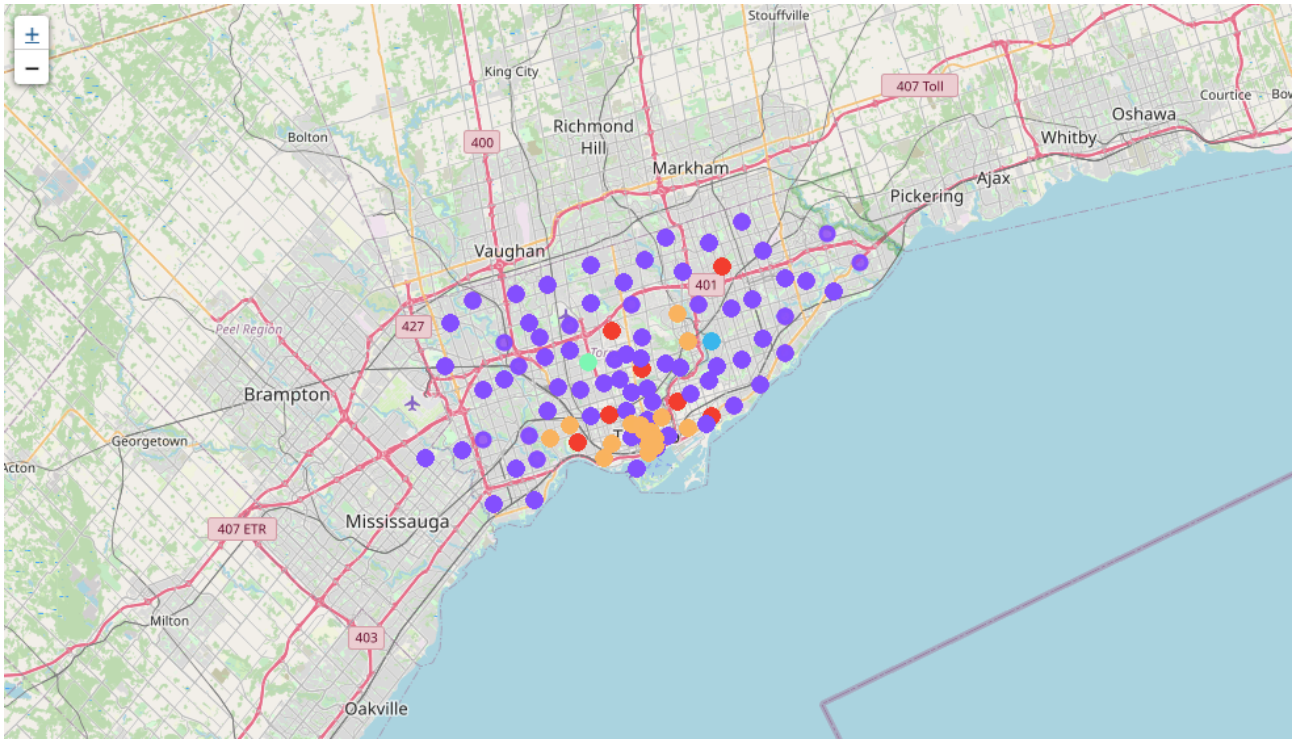
With this data, first I checked how many unique categories that I can get from these venues, then, I analyzed each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Later, I did some queries to find out typical European Restaurants (I limited queries to Italian, Spanish, Portuguese and English), closest to Italian cuisine.

Lastly, I performed the clustering method by using 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. I have clustered the neighborhoods in Toronto into 5 clusters. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

5. Results

Here below I reported the clusters:



6. Recommendations

From analysis, it came out that better clusters to start an Italian Restaurants are 0, 2, 3.

7. Limitations and suggestions

It could be useful to extend analysis to other European Restaurants.

8. 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.

Angelo Antonio Biancamano
(https://github.com/olegna1984/Coursera_Capstone)

(https://github.com/olegna1984/Coursera_Capstone/blob/master/Battle%20of%20neighborhoods%20week%201.ipynb)