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

Report By:

Syed Muhammad Areeb Imran(www.github.com/syedareeb)

1.Introduction

1.1 Background

For this Capstone Final Project, I am creating a hypothetical scenario in which an entrepreneur wants to open an Italian restaurant in Toronto. The main idea behind this approach is that this entrepreneur wants to open his restaurant in an area where there are Italian cuisine is popular so it will be easy to attract customers in that particular area and the restaurant will have a higher probability of success. With this goal in the mind this project has been designed to satisfy the need of the entrepreneur to upmost extent by finding the most suitable location for him to open his restaurant.

1.2 Business Problem

The main objective of this capstone project is find the most suitable location by using the techniques and skill acquired from the course such as machine learning, data visualization and data analyzing, fulfilling the aim of this project to provide an answer to the business question: In Toronto , if an entrepreneur wants to open an Italian restaurant, what would be the ideal location?

1.3 target Audience

The entrepreneur who want to find the suitable location to open an Italian Restaurant.

2.Data

To solve this problem, following data will be required:

- List of neighborhoods in Toronto, Canada.
- Latitude and Longitude of these Neighborhoods.
- Venue Data related to Italian restaurants. This will help us in finding the suitable neighborhood to open the restaurant.

3.Extracting Data

- Scrapping of Toronto neighborhoods via Wikipedia.
- Getting the latitude and longitude data of these neighborhoods using geocoder package.
- Using Foursquare API to get venue data of the neighborhoods.

4.Methodology

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia page (**https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M).** I did the web scrapping by utilizing pandas html table scrapping method as it is easier and more convenient to pull tabular data directly from a web page into dataframe. However, it is only 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 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 neighborhoods. 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 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, I made a justification to look for "Italian Restaurants". 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 neighborhoods in Toronto into 4 clusters based on their frequency of occurrence for "Italian Food". Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

5.Results

Clusters.

The results from k-means clustering show that we can categorize Toronto neighborhoods into 4 clusters based on how many Italian restaurants are in each neighborhood:

- Cluster 1: neighborhoods with very little number of Italian Restaurants.
- Cluster 2: Neighborhoods with little to no Italian Restaurants.
- Cluster 3: Neighborhoods with a large number of Italian Restaurants.
- Cluster 4: Neighborhoods with also large number of Italian Restaurants.

6.Recommendations

Most of Italian restaurants are in Cluster 1, 3 and 4 which is around India bazaar, Riverdale, Central Bay Street and University of Toronto. There are little to no restaurants in Cluster 2. So cluster 2 holds the greatest potential for an Italian restaurant to succeed.so the area around North Toronto West, Richmond, North Midtown and Regent Park are good places for the entrepreneur to open his Italian Restaurant and have a higher chance of success due to little to zero competition. Nonetheless, if the food is authentic, affordable and good taste, I am confident that it will have great following everywhere.

7.Limitations and Suggestions for Future Research

In this project, I only take into consideration of one factor: the occurrence / existence of Italian 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.

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.

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