DATA SCIENCE CAPSTONE PROJECT Report

For IBM Professional Certificate

Made by

Kateryna Zavidniuk

02/27/2021

This is the capstone project for IBM Data Science Professional Certificate. In this project, I was proposing an idea for a concept that there may not be enough Small Family Restaurants in some Toronto Area where they should be very popular. As it might be a big opportunity for a small businesses that are based in Toronto (Canada). The Small Family Restaurants should be very popular among the family-oriented community, so this businessman might think of opening his/her business in areas where residential areas located. With this idea, finding the location to open such a restaurant is one of the most important decisions for this businessman and I am creating this project to help him/her find the most advantageous location.

BUSINESS PROBLEM

The purpose of this capstone project is to look for and find the most profitable location for the businessman to open a new Small Family Restaurant in Toronto (Canada). Using data science methods and tools together with machine learning algorithms such as clustering, this project aims to provide recomendations to answer the business question: Where should be the new Small Family Restaurant in Toronto?

PURPOSE AUDIENCE

The businessman who wants to find the best place to open the Small Family restaurant.

DATA

To solve this task, we will need to use the following data:

- Latitude and Longitude of these neighbourhoods;
- List of neighbourhoods in Toronto(Canada);
- Venue data related to small family restaurants' location.

This will help us find the neighbourhoods that are more advantageous to open a Small Restaurant.

EXTRACTING THE DATA

- Getting Latitude and Longitude data of these neighbourhoods via Geocoder package;
- The scrapping of Toronto neighborhoods via Wikipedia;
- Using Foursquare API to get venue data related to these neighborhoods.

METHODOLOGY

To begin with, I got the list of neighborhoods in Toronto (Canada). I did this by extracting the list of neighborhoods from Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

I have done the web scraping by utilizing pandas HTML table scraping method, as it was easier and more easy to pull tabular data directly from a web page into the data frame. Although, it was only a list of neighborhood names and postal codes. I needed to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I used 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 these data coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Then, 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 was able to pull the names, categories, latitude, and longitude of the venues. With this data, I could also check the quantity of the unique categories that I could 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. Here, I made a justification to specifically look for “Family restaurants”. At the end, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocated 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 small family restaurant. Based on the results (the concentration of clusters), I was able to recommend the ideal place to open the restaurant.

CLUSTERS

The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Small Family Restaurants are in each neighborhood.

- Cluster 0: Neighbourhoods with the less number of Small Family Restaurants.
- Cluster 1: Neighbourhoods with no Family restaurants.
- Cluster 2: Neighbourhoods with the most number of Small restaurants

CONCLUSION

RECOMANDATIONS

The most number of the Small Family restaurants are in cluster 2 which is around Central Bay Street, Church and Wellesley, Berczy Park, Union Station, Richmond, the lowest in Cluster 1 areas which are in North Toronto West and Parkade areas. So this is the best place to open one of such kind restaurants.