Using Machine Learning to find the location to open a Filipino restaurant in Toronto

1. Introduction

1.1. Background

For this project, I am using a hypothetical situation where a Filipino restaurant owner wants to open a new Filipino restaurant in the Toronto area. The idea behind this scenario is that there may not be enough Filipino restaurants in Toronto and presents an excellent opportunity for an entrepreneur based in Canada. Filipino food is similar to other Asian cuisines the entrepreneur in thinking of opening restaurant in locations where other Asian food is popular. Therefore, finding the location is one of the most important decisions for the client and data science methodology leveraged with geographical data is one of the most effective methods to make such a decision.

1.2. Business Problem

The objective of this project is to find the most suitable location for an entrepenuer to open a new Filipino restaurant in Toronto, Canada. By using data science and machine learning methods such as clustering combined with geographical location data from Foursquare this project aims to provide solutions to the business problem; where is the best potential location for the opening of a Filipino restaurant in the city of Toronto, Canada?

1.3. Target Audience

The entrepreneur who wants to find the location to open a new Filipino restaurant

2. Data

To solve the business problem, the following data will be required:

- A list of neighbourhoods in Toronto, Canada
- The geographical location in latitude and longitude of these neighbourhoods
- Venue data related to Asian cuisine restaurants in the city of Toronto, Canada. This will allow us to find the neighbourhoods most suitable to open a new Filipino Restaurants.

To fulfil the requirements of the data that will be required to implement the solution:

- A list of Toronto neighbourhoods scraped from Wikipedia
- Getting Latitude and Longitude information for the neighbourhoods through the geocoder package
- Using the Foursquare API to get the venue data related to all the neighbourhoods scraped from the Wikipedia list.

3. Methodology

A list of neighbourhoods in Toronto, Canada was extracted from the list of neighbourhoods Wikipedia page. Web scraping of the Wikipedia page was carried out by utilising the pandas html table scraping method to pull the data directly from the webpage into a dataframe. The generated dataframe only contains a list of neighbourhood names and postal codes. The dataframe was joined to data provided by the IBM team to match latitude longitude coordinates

to the names of the Toronto neighbourhoods. The neighbourhoods were visualised using the Folium package to verify the geographical positions of the Neighbourhoods.

The Foursquare API was used to request a list of top 100 venues with a 500-meter radius. To use the API a Foursquare developer account was created and the credentials used to verify the request. Using the Foursquare API, the names, categories, latitude and longitude of venues were provided, furthermore using the data can be examined for unique categories that the venues belong to. The neighbourhoods were analysed by grouping the instances by neighbourhood and taking the mean frequency occurrence of each venue category. This prepares the data for later processing and data mining steps.

A decision was made during the data analysis stage to search the location of 'Thai Restaurants' through the FourSquare API. This was due to a lack of venues under the 'Asian Restaurant' or associated venues such as 'Chinese Restaurant' under the FourSquare API. 'Thai Restaurant' appeared most frequently under the umbrella of 'Asian Restaurant', however, Filipino and Thai food have very similar tastes and ingredients therefore a person that would enjoy one type of cuisine would most likely enjoy the other.

Finally clustering of venues was performed by using the K-means algorithm. K-means clustering identifies k number of centroids and then allocates every data point to the nearest cluster while keeping the number of centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is suited for this business problem. The neighbourhoods of Toronto were separated into different clusters based on the frequency of the occurrence of 'Thai Restaurant' venues within each neighbourhood. Based on the results of the clustering an ideal location for the restaurant was recommended.

4. Results

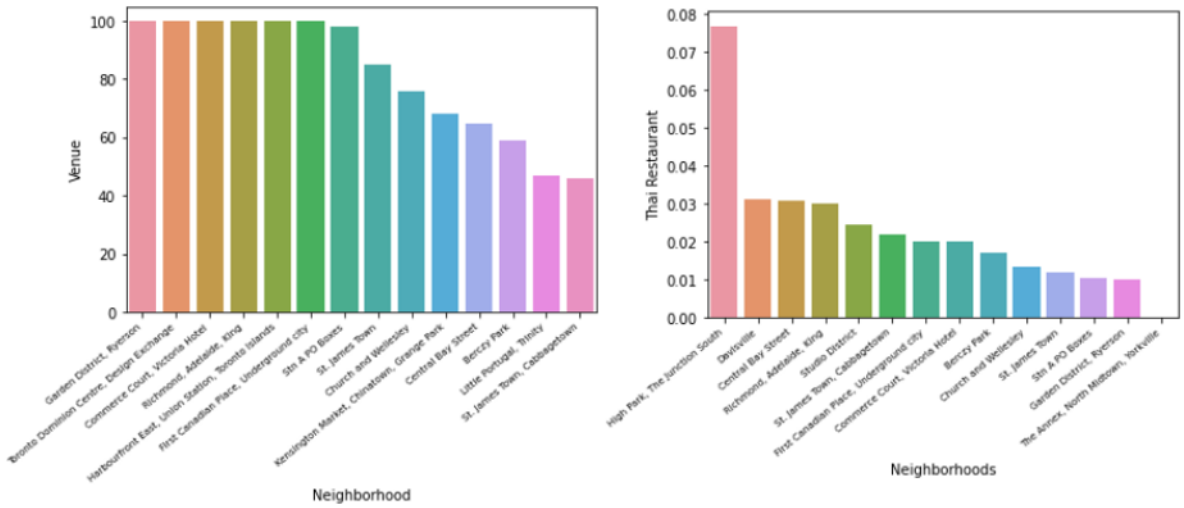

Figure 1 shows the top eleven neighbourhoods with the most number of venues in the city of Toronto (left) and the neighbourhoods with the highest percentage of Thai restaurants in central Toronto (right). It can be seen that there are 6 neighbourhoods with at least 100 venues, only 2 of which contain a Thai restaurant, only a single neighbourhood contains more than 1 Thai restaurant. There are only 13 Thai restaurants in the central Toronto area.

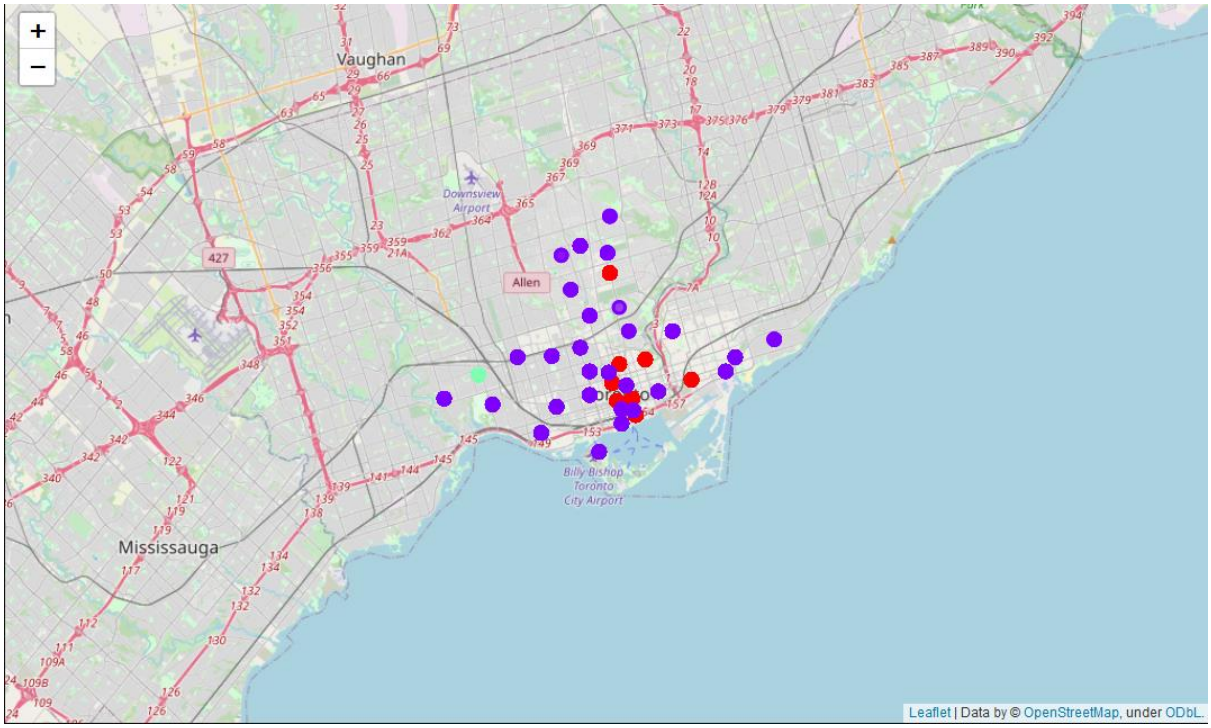

Figure 2: A map of Toronto showing the central neighbourhoods that are have been assigned to one of 3 clusters based on how many restaurants there are in the neighbourhood, neighbourhoods with no Thai restaurants (purple), neighbourhoods with one Thai restaurant (red) and neighbourhoods with a high number of Thai restaurants (green)

Figure 2 shows the results from k-means clustering where Toronto neighbourhoods have been assigned to a cluster based on how many Thai restaurants in each neighbourhood. There are

three clusters, neighbourhoods with no Thai restaurants, neighbourhoods with one Thai restaurant and neighbourhoods with more than one Thai restaurant. The neighbourhood that contains more than one restaurant is the high park and the Junction south, which is a fair distance from downtown Toronto. The other neighbourhoods that contain a single Thai restaurant in the downtown area but there are a large number of neighbourhoods in Toronto that do not have a single Thai restaurant.

5. Discussion and Recommendations

5.1. Discussion

Toronto is the largest city in Canada and the fourth largest in North America, it is a cosmopolitan city with a bustling food scene. There are 39 neighbourhoods in all Toronto however, there are only 13 neighbourhoods that have Thai restaurants. High Park and the junction south have the highest concentration of Thai restaurants, this neighbourhood is far away from downtown, which would lead to low foot traffic. The majority of restaurants are centrally located in the downtown area, around the Adelaide, King and Richmond areas. There are six central neighbourhoods that have over 100 venues within the zip code and will have a great deal of competition trying to open a new restaurant in the area.

During the course of the data analysis for this project, only one factor has been taken into account; the existence of Thai Restaurants in each neighborhoods. There are many more factors that need to be taken into account such as population density, the income of residents in the neighbourhood. However, this requires looking for further data sources for this project within the short timeframe for this project. Future research can take into consideration these factors. Furthermore, Thai restaurants was used for this project due to a shortage of Asian restaurants but also does not take into account of the local population demographics.

5.2. Recommendation

There is a good opportunity to open near Kensington market, Chinatown and Grange park, this is mainly due to the downtown foot traffic but only has 60 venues within the local neighbourhood. This is a compromise between the amount of foot traffic and only a limited number of venues to limit the competition. This project recommends any of the neighbourhoods near the downtown area that is a member of the purple cluster from figure 2 would provide good locations for the new Filipino restaurant. This would provide the entrepreneur with little to no competition but relatively high foot traffic.

6. Conclusion

Over the course of 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 utilising k-means clustering and providing a final recommendation to the stakeholder.

7. References

List of Neighbourhoods in Toronto:

'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare Developer documentation: https://developer.foursquare.com/docs