The Battle of Neighborhoods - Coursera IBM Capstone Project

1. Introduction: Business Problem

Toronto is the capital of the Canadian province of Ontario. With a registered population of 2,731,571 in 2016, [14] is the most populous city in Canada and the fourth most populous city in North America. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (in 2016) around the western end of Lake Ontario, [15] while the Greater Toronto (GTA) area itself had a population of 6,417. 516 in 2016. Toronto is an international center for business, finance, arts and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

Therefore, it became very challenging for the stakeholder or for the new business to decide in which area they should start their business in order to obtain the highest revenue with the least possible competition.

Target Audience: New businesses that want to open a restaurant in Toronto.

This project tries to solve the problem above by suggesting to the Target Audience which is the best place to open new restaurants and obtain the maximum profits in Toronto.

2. Data

Based on the definition of our problem, the factors that will influence our decision are:

All existing restaurants in the neighborhood (any type of restaurant) Age group of people with their income Distance from the neighborhood to the city center We decided to use a regularly spaced grid of places, centered in the city center, to define our neighborhoods.

The following data sources will be needed to extract / generate the necessary information:

candidate area centers will be generated algorithmically and approximate addresses of centers in these areas will be obtained using https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M number of restaurants and their type and location in each neighborhood will be obtained using the Foursquare API.

3. Methodology

The main goal of this project is to find the best place to open a new restaurant in Toronto, Canada, based on competition in different locations and its population.

So, to do this, I used 2 different data sets available, as mentioned above. These 2 sets of data contain information about the locality of Toronto, different age groups of people, and population.

To solve the problem, I will use the "K-Means clustering algorithm". K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (that is, data without defined categories or groups). The purpose of this algorithm is to find groups in the data, with the number of groups represented by the K variable. The algorithm works iteratively to assign each data point to one of the K groups based on the characteristics that are provided. The data points are grouped based on the similarity characteristic. The results of the K-means clustering algorithm are:

The K cluster clusters, which can be used to label new data labels for training data (each data point is assigned to a single cluster)

In addition, I will be using different maps to give a clear view to the target audience.

Steps we took for analysis:

Required data ** collected: location and type (category) of each restaurant in our latitude and longitude. We also have the type of restaurants in particular in the locality. We explore the 'restaurant density' in different areas of Toronto - we will use K-mean to identify some promising areas close to the center with a low number of restaurants and their type. We explore the most promising areas and within them create clusters of locations that meet some basic requirements established in discussion with stakeholders: we will take into account locations with fewer restaurants within a radius of 500 meters, We will present the map of all these locations, but also we will create clusters (using k-means grouping) of these locations to explore the neighborhood.

4. Analyze

Data identification, capture and cleaning

Search and identify the relevant data source and capture it. We are using wikipedia to obtain data about Toronto, Canada. Then, we remove the entire redundant value (data wipe). Then we combine neighborhoods similar to the Bronx. The data is now clean and ready to use.

Combine different data sources and classify the neighborhood based on longitude and latitude

Now, let's combine the neighborhood data set with the postal address and the data set with Latitude and Longitude and save them in a separate data frame. The resulting data frame contains details on Postal Code, Brough, Neighborhood, Latitude and Longitude. Then, visualize it using the folio map.

Explore Toronto neighborhoods

First, we explored all neighborhoods in the city of Toronto, using the latitude and longitude data, using the Foursquare API to get the restaurant locations available in Toronto.Explore the unique neighborhood categories.Filter the details of the locations for everyone the possible 'restaurants'. Find each neighborhood along with the main most common locations. Identify the top 10 locations for each neighborhood.

Clustering

With an assumption of 5 clusters, use the K-Cluster algorithm to reach 5 different clusters in Toronto with a similar set of locations. Explore each cluster and determine the categories of discriminating locations that distinguish each cluster. Identify clusters and neighborhoods / neighborhoods with maximum number of restaurants and their types.

5. Results and discussion

Our analysis shows that while there are a large number of restaurants in Toronto, there are low density pockets of restaurants very close to the city center. We have 4 neighborhoods and 74 neighborhoods within the geographic coordinate of 43.653963, -79.387207.

Based on our initial assumption that the cluster with the maximum number of restaurants will have the best chance of having a new restaurant due to the need of the area. Based on the resulting clusters, it appears that Cluster 1 and Cluster 5 have a greater number of restaurants than the rest of the clusters.

It is quite possible that there is a very good reason for the small number of restaurants in any of these areas, reasons that would make them unsuitable for a new restaurant, regardless of the lack of competition in the area. The recommended zones should therefore only be considered as a starting point for a more detailed analysis that could eventually result in a location that not only has no competition nearby, but also other factors taken into account and all other relevant conditions met. .

6. Conclusion

The objective of this project was to identify areas in Toronto with a low number of restaurants, in order to help interested parties to narrow the search for the ideal location for a new restaurant. When calculating the density distribution of the restaurant from Foursquare data, we first identify general neighborhoods that warrant further analysis and then generate an extensive collection of locations that satisfy some basic requirements regarding existing restaurants in the vicinity. The grouping of these sites was then carried out in order to create the main zones of interest (containing the largest number of potential sites) and the addresses of these zone centers were created to be used as starting points for the final exploration by the interested parties .

The final decision on the ideal location of the restaurant will be made by the stakeholders based on the specific characteristics of the neighborhoods and locations in each recommended area, taking into account additional factors such as the attractiveness of each location (proximity to the park or water), noise levels / proximity to main roads, availability of properties, prices, social and economic dynamics of each neighborhood, etc.