Wholesale Grocery Recommender

IBM - Capstone Report

1. Introduction

1.1 Background

There is a groceries contractor in one of the boroughs of Toronto (Scarborough). This contractor provides places such as: Different types of Restaurants, Bakery, Breakfast Spot, Brewery and Café with fresh and high-quality groceries. The contractor wants to build a warehouse for the groceries it buys from villagers and farmers inside the borough, so that they will support more customers and also bring better "Quality of Service" to the old customers.¶

1.2 Problem

For example, if the warehouse is close to those old and famous restaurants, then the vegetables and other groceries would be delivered to the restaurant in the right time and there would be no delay so the restaurant cooks can start their job from the morning and the Quality of Service will be high and this contractor will gain more reputation and income.

1.3 Interest

The contractor should build this warehouse where it is closest to its customers in order to minimize the cost of transpotation in addition to the example above. which neighborhood (in that borough) would be a better choice for the contractor to build the warehouse in that neighborhood. Finding the right neighborhood is our mission and our recommender system will provide this contractor with a sorted list of neighborhoods in which the first elemnt of the list will be the best suggested neighborhood.¶

2. Data Sources

We will need geo-locational information about that specific borough and the neighborhoods in that borough. We specifically and technically mean the latitude and longitude numbers of that borough. We assume that it is "Scarborough" in Toronto.

This is easily provided for us by the contractor, because the contractor has already made up his mind about the borough. The Postal Codes that fall into that borough (Scarborough) would also be sufficient for us. I fact we will first find neighborhoods inside Scarborough by their corresponding Postal Codes.

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. By locational information for each venue we mean basic and advanced information about that venue. For example there is a venue in one of the neighborhoods. As basic information, we can obtain its precise latitude and longitude and also its distance from the center of the neighborhood. But we are looking for advanced information such as the category of that venue and whether this venue is a popular one in its category or maybe the average price of the services of this venue. A typical request from Foursquare will provide us with the following information:

[Postal Code] [Neighborhood(s)] [Neighborhood Latitude] [Neighborhood Longitude] [Venue] [Venue Summary] [Venue Category] [Distance (meter)]

[M1L] [Clairlea, Golden Mile, Oakridge] [43.711112] [-79.284577] [Tim Hortons] [This spot is popular] [Coffee Shop] [592]¶

3. Methodology

Part 1: Identifying Neighborhoods inside "Scarborough"

We will use Postal Codes of different regions inside Scarborough to find the list of neighborhoods. We will essentially obtain our information from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and then process the table inside this site. Images from dataframes and also from maps will be provided in the presentation. Here we only present our strategy and how we got the mission accomplished.

Part 2: Connecting to Foursquare and Retrieving Locational Data for Each Venue in Every Neighborhood

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 1000 meter. It means that we have asked Foursquare to find venues that are at most 1000 meter far from the center of the neighborhood. (I think distance is measured by latitude and longitude of venues and neighborhoods, and it is not the walking distance for venues.)

Part 3: Processing the Retrieved Data and Creating a DataFrame for all the venues inside Scarborough

When the data is completely gathered, we will perform processing on that raw data to find our desirable features for each venue. Our main feature is the category of that venue. After this stage, the column "Venue's Category" will be One-hot encoded and different venues will have different feature-columns. After On-hot encoding we will integrate all restaurant columns to one column "Total Restaurants" and all food joint columns to "Total Joints" column. We assumed that different restaurants use the Same raw groceries. This assumption is made for simplicity and due to not having a very detailed dataset about different venues.
Now, the dataset is fully ready to be used for machine learning (and statistical analysis) purposes.

Part 4: Applying one of Machine Learning Techniques (K-Means Clustering)

Here we cluster neighborhoods via K-means clustering method. We think that 5 clusters is enough and can cover the complexity of our problem. After clustering we will update our dataset and create a column representing the group for each neighborhood.

4. Results

Now, we focus on the centers of clusters and compare them for their "Total Restaurants" and their "Total Joints". The group which its center has the highest "Total Sum" will be our best recommendation to the contractor. {Note: Total Sum = Total Restaurants + Total Joints + Other Venues.} This algorithm although is pretty straightforward yet is strongly powerful.

Based on this analysis, the best recommended neighborhood will be:
{'Neighborhood': 'Agincourt',
'Postal Code': 'M1S'
'Neighborhood Latitude': 43.7942003,
'Neighborhood Longitude': -79.26202940000002}

5. Discussion

In this study, I analyzed the relationship grocery vendors and location demand. In my opinion a key take back from this analysis is that regardless of an area's socio-economic status, wholesale contractors tend to prefer an are where there is a large population. This might seem obvious since everyone requires groceries and more people means more business. But, what people have overlooked is the business aspect of it. The more accessible and affordable a warehouse is, the easier it is to transport goods even if it is required to position far away from the market.

6. Conclusion

From the results above, we can safely conclude that effective data analysis provides us with reasonable and rational decision making rules to help us analyze various aspects of a business What might seem obvious may have many caveats to it and it is important to understand all of them and more importantly understand why do the exist.