IBM Data Science Capstone Project
Dheeraj Patta

Finding Optimum Location for opening an Indian Grocery Store in Toronto (using Machine Learning)

Introduction (Background)

For this Capstone project, I'm building a solution for a friend who is an entrepreneur based in Canada. He is willing to explore and open an Indian grocery store in Toronto and expand to Asian Groceries chain. There is healthy population of Asians in Canada and there might be a great demand for Indian cuisines and ingredients, thanks to all the Asian and Indian restaurants. However, there will be heavy competition in the market and there is an absolute need to find the best place and resources to succeed.

Problem Statement –

The objective of this capstone project is to find the most suitable location for my friend to open a new Indian grocery store and later expand to Asian grocery chain in Toronto, Canada.

We need to identify the high demand and low demand neighborhoods as well as population density, income per capita, crime rate and near-by venues/attractions. We will be using data science methods and machine learning methods such as clustering to provide solutions to answer the following business question:

What is the best place to open an Indian grocery store in Toronto?

Data Requirements –

We need to have access to location and geographical data to predict the optimum location for opening an Indian grocery store. The following are some of the data points necessary to perform this analysis and predictions.

- List of neighborhoods in Toronto, Canada.
- Latitude and Longitude of these neighborhoods.
- Venue data related to Asian restaurants.

The Foursquare API location data is a great place to start the analysis to identify the clusters of neighborhood and optimum location for opening the grocery store.

Data Collection –

The following data needs to be collected to perform the necessary analysis – We will make use of Wikipedia information for scraping Toronto neighborhood related data. We'd also use Geocoordinates (Longitude and Latitude) via Geocoder package

Finally, we'd extract Venue and Nearby location information using Foursquare API.

Methodology –

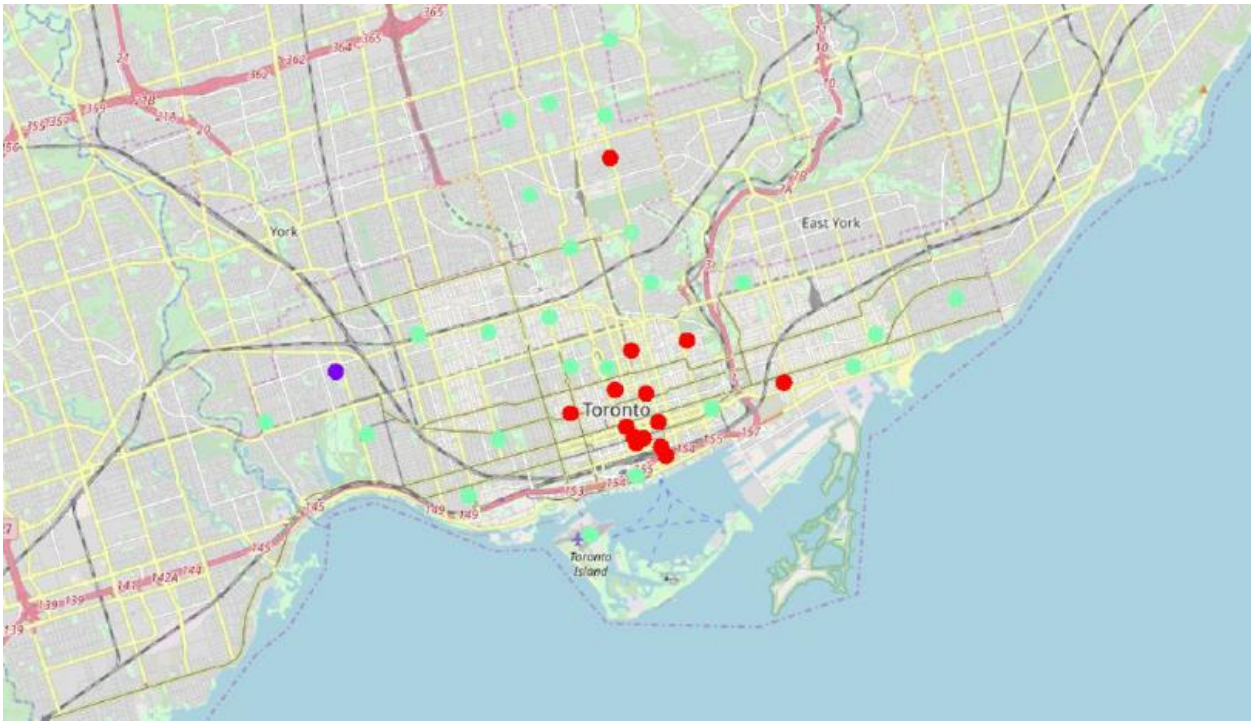
We will extract the location information and geocoordinates of Toronto and its respective neighborhoods. The list of boroughs and neighborhoods are scraped from Wikipedia page using html table scraping method. This gives us the list of neighborhood names and postal codes. Next, we have loaded geocoordinates (longitude and latitude) information using csv load. Finally, we made an API call to Foursquare API for Venue information and Near by places using the Foursquare Developer account. This provides the list of top 100 venues within a radius limit of 500 meters.

Foursquare API provides the detailed information related to Names, Categories, Neighborhoods, latitude and longitude of the Venues. We've performed the necessary Exploratory Data Analysis and Grouping by Neighborhood of each Venue category.

We've picked "Grocery Store" as our category and performed K-means clustering to understand the density of grocery stores. 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 perfectly fits the need of this problem statement.

The data is now clustered into 3 areas (k = 3) based on the frequency of the occurrence of "Grocery Store" and based on the clusters we can find the optimal location/neighborhood for opening up an Indian Grocery store.

Results –



The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Grocery Stores are in each neighborhood:

- Cluster 1: Neighborhoods with high number of Grocery Stores
- Cluster 3: Neighborhoods with good number of Grocery Stores
- Cluster 2: Neighborhoods with less Grocery Stores

Most of the Categories are in Cluster 1 and 2. The neighborhoods in Cluster 1 and 2 are densely occupied by Grocery Stores and Restaurants and are near by various other locations.

There are very few Grocery Stores in Cluster 3 in the neighborhood of Christie. According to the location data and nearby venues, it is less competitive and more probability of Success in the neighborhood of Christie for opening up a new Indian Grocery Store.

Limitations –

There are many factors that can contribute to the success of opening an Indian grocery store in a given location like per capita income, rents, population density, Indian population and operation planning.

However, in this exercise we assumed the existence of any Grocery stores and Nearby venues as the only factor for our prediction of the optimum neighborhood cluster.

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

New York Clustering Lab – Week 2

Segmentation and Clustering Class – Week 2