IBM Applied Data Science Capstone

Predicting the Best Location to Open a New Indian Restaurant in New Delhi, India

By: Dinesh Ogirala

July 2020

Introduction

For many food lovers, Indian Restaurants are a great way to relax and enjoy not only during weekends and holidays but also during weekdays. Indian Restaurants are like a one-stop destination for family gatherings. Most of the people in Delhi prefer Dumdar Indian Food to Western Foods. For retailers, the central location and the large crowd at the Restaurant provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more Restaurants to cater to the demand. As a result, there are many Restaurants in the city of New Delhi and many more are being built. Opening Restaurants allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new Restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the Restaurant is one of the most important decisions that will determine whether it will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of New Delhi, India to open a new Restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of New Delhi, India, if a property developer is looking to open a new Restaurant, where would you recommend that they open it?

Data

To solve the problem, we will need the following data:

- List of neighborhoods in New Delhi. This defines the scope of this project, which is confined to the city of New Delhi, the capital city of India.
- Latitude and longitude coordinates of those neighborhoods. This is required to plot the map and to get the venue data.
- Venue data, particularly data related to restaurants. We will use this data to perform clustering on the neighborhoods.

Data Extraction

This Wikipedia page (https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi) contains a list of neighborhoods in New Delhi, with a total of 9 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then, we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Indian Restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

Firstly, we need to get the list of neighborhoods in the city of New Delhi. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. Next, we will use the Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas Data Frame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of New Delhi. Next, we will use Foursquare API to get the top 200 venues that are within a radius of 5000 meters. We need to register a Foursquare Developer Account to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Indian Restaurants” data, we will filter the “Indian Restaurants” as venue category for the neighborhoods. Lastly, we will perform clustering on the data by using k-means clustering. 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 is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Indian Restaurants”. The results will allow us to identify which neighborhoods have higher concentration of Restaurants while which neighborhoods have fewer. Based on the occurrence of Restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new restaurants.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Indian Restaurants”,

• Cluster 0: Neighborhoods with moderate number of restaurants.

• Cluster 1: Neighborhoods with low number to no existence of restaurants.

• Cluster 2: Neighborhoods with high concentration of restaurants.

The results of the clustering are visualized in the map with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.

Discussion

As observations noted from the map in the Results section, most of the restaurants are concentrated in the central area of New Delhi, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no Indian restaurants in the neighborhoods. This represents a great opportunity and high potential areas to open new restaurant as there is very little to no competition from existing restaurants. Meanwhile, restaurants in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of restaurants. From another perspective, the results also show that the oversupply of restaurants mostly happened in the central area of the city, with the suburb area still have very few restaurants. Therefore, this project recommends property developers to capitalize on these findings to open new restaurants in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new restaurants in neighborhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of restaurants and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of restaurants, there are other factors such as population and income of residents that could influence the location decision of a new restaurant. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new restaurant. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results. Moreover, the data is very limited to make accurate predictions.

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 machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant.