Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Restaurants in Delhi,India

Author: Souvik Ghosh Roy Chowdhury

October, 2019

Introduction

Delhi is city, growing exponentially in Economical sector. With its population near about 1.9 crores it becomes one of the top market for investors. This notebook gives investors brief idea for opening new restaurants. For many visitors, food lovers, visiting Restaurants is a great way to relax and enjoy verities of foods themselves during weekends and holidays, even on weekdays after hectic office hours. As a market Delhi is great place where crowd is key for running a Restaurant Business. 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 Delhi, Capital of India, 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 Restaurants requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the Restaurants is one of the most important decisions that will determine whether the Restaurant 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 Delhi, India to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city Delhi, India, if a property developer is looking to open a new restaurant , where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new restaurants in the capital city of India i.e. Delhi. This project is timely as the demand of different types restaurants in city is increasing day by day. For example Japan curry chain to open 1st India restaurant in New Delhi suburb and many more to come in near future.

Data

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

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

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi) contains a list of neighbourhoods in Delhi, with a total of 135 neighbourhoods. 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 neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database 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 Shopping Mall 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). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Delhi. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi)We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful 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 DataFrame and then visualize the neighbourhoods 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 Delhi.

Next, we will use Foursquare API to get the top 50 venues that are within a radius of 1000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods 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 neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 analysing the “Restaurant” data, we will filter the “Restaurant” as venue category for the neighbourhoods.

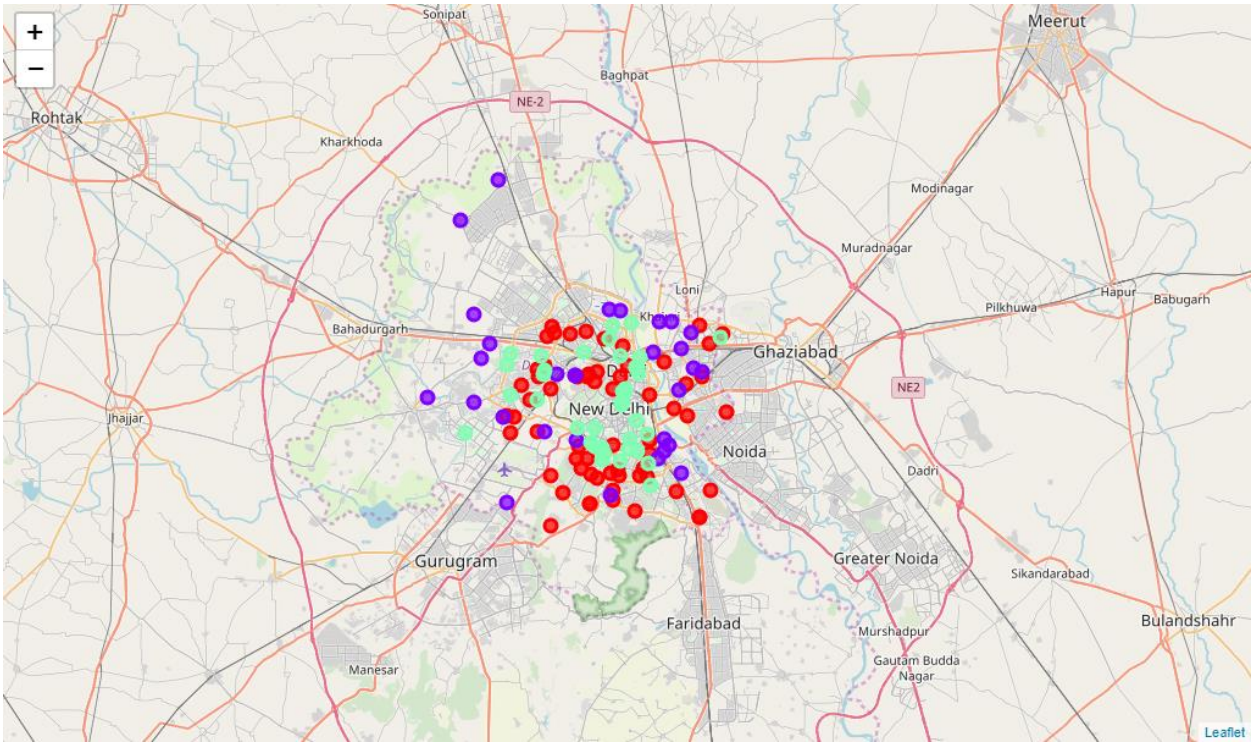
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Restaurants”. The results will allow us to identify which neighbourhoods have higher concentration of Restaurants while which neighbourhoods have fewer number of Restaurants. Based on the occurrence of Restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Restaurants.

Results

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

- Cluster 1: Neighbourhoods with moderate number of restaurants
- Cluster 2: Neighbourhoods with low number to moderate number of restaurants
- Cluster 3: Neighbourhoods with high concentration of restaurants

The results of the clustering are visualized in the map below with cluster 1 in red colour, cluster 2 in purple colour, and cluster 3 in mint green colour.



Discussion

As observations noted from the map in the Results section, we can see cluster 1 has moderate number of restaurants in neighborhoods,cluster 2 also has low to moderate number of restaurants while cluster 3 has large number restaurants.Hence to start a profitable business of opening restaurants,cluster 2 is best and competition is less here compared to other clusters.Cluster 1 is also good place to start business where competition is greater than cluster 2 and less than cluster 3.

In brief most of the restaurants in Delhi are concentrated in Central area,where North-East part is best to set up restaurants for investors.

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 restaurants. However, to the best knowledge of this researcher such data are not available to the neighbourhood 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 restaurants. 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.

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 restaurants. To answer the business question that was raised in the introduction section, the answer proposed by this project is: cluster 1 has moderate number of restaurants in neighborhoods,cluster 2 also has low to moderate number of restaurants while cluster 3 has large number restaurants.Hence to start a profitable business of opening restaurants,cluster 2 is best and competition is less here compared to other clusters.Cluster 1 is also good place to start business where competition is greater than cluster 2 and less than cluster 3.

References

Category:Neighbourhoods_in_Delhi. Wikipedia. Retrieved from (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi)

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs