Predicting Location in Delhi for a Mall

Satyendra

Introduction

Delhi is the largest commercial centre in northern India. As of 2016 recent estimates of the economy of the Delhi urban area have ranged from $167 to $370 billion (PPP metro GDP) ranking it either the most or second-most productive metro area of India. The nominal GSDP of the NCT of Delhi for 2016–17 was estimated at ₹6,224 billion (US$90 billion), 13% higher than in 2015–16.

Delhi has an attractive real estate market and is a preferred tourist destination. Owing to its location, connectivity and rich cultural history, Delhi has always been a prime tourist attraction of the country. Many shopping malls are present in Delhi and opening a new one requires certain factors in mind.

Problem:

Data that might contribute to determining Shopping mall opening location might include his area population and nearby shopping centres. This project aims to predict whether and how much a feasible will be to open the restaurant in the area.

Target Audience:

Various big brands, which invests in opening malls in Delhi. We can see in the below report that future is bright and companies are coming with the various projects.

Data

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

- List of neighbourhoods in New Delhi. This will elaborate the scope of data.
- For plotting the locations, we need longitudes and latitudes.
- 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/Neighbourhoods_of_Delhi) contains a list of neighbourhoods in New Delhi.
- We will use web-scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful soup 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 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 and the machine learning technique that is 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/Neighbourhoods_of_Delhi). We will do web scraping using Python requests and beautiful soup 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 data frame 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 co-ordinates data returned by geocoder are plotted correctly in the city of Delhi map.

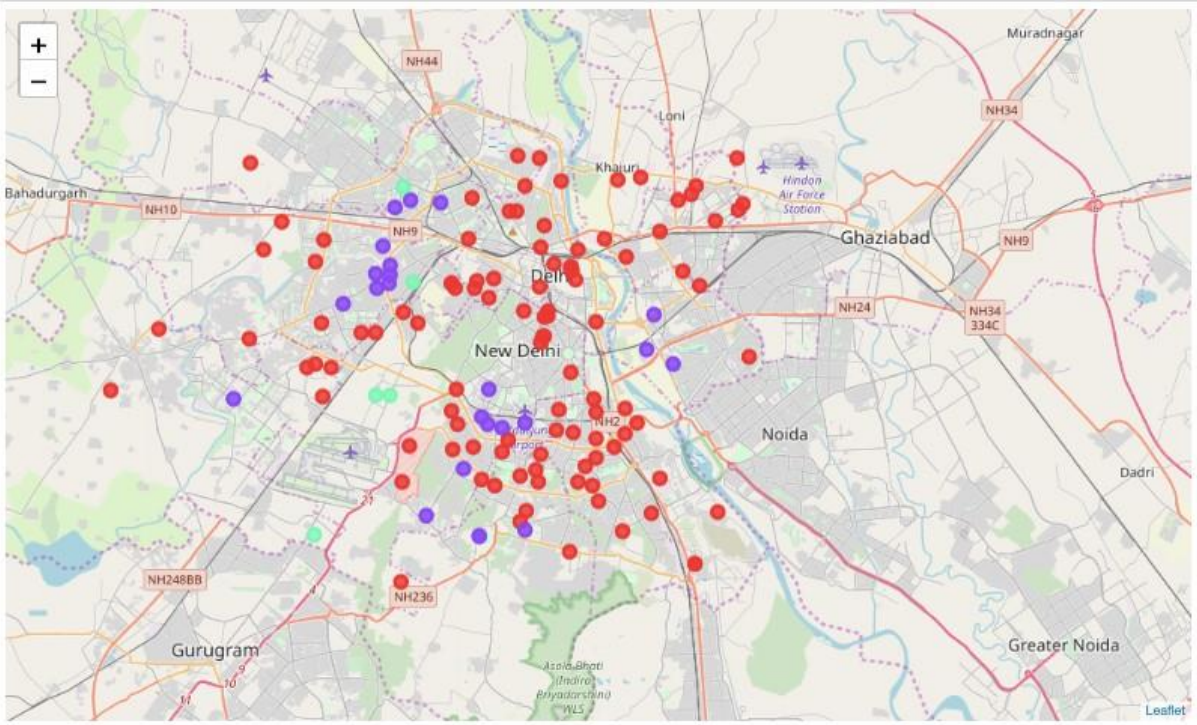
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods 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 location 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 "Shopping Mall" data, we will filter the "Shopping Mall" 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 three clusters based on their frequency of occurrence for "Shopping Mall". The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

Results:

The data is divided in three clusters, which represents different intensities of the shopping malls situated in the area

- Cluster 0: Moderate density of the shopping complex, it is represented by red colour.
- Cluster1: Low density of shopping complexes, it is represented by purple colour.
- Cluster 3: High density of shopping complexes, It is represented by mint green colour.



Discussion:

We can deduce the fact from the map that the most of the malls are located in south Delhi and central Delhi. These areas have moderate concentration but localities with moderate concentration is quite high. South and Central Delhi neighbourhoods are mostly in cluster 0. West and east Delhi have less no of mall concentration in cluster0. Cluster 1 neighbourhoods are mostly located in western and Southern part of Delhi. East Delhi have very few neighbourhoods in cluster 1. We can also notice that that north do not have areas in cluster 1.
West Delhi is the only region where the most no of cluster 2 neighbourhoods are present.

We can easily see from the visualisation that south-western part and eastern part of the capital has very less frequency of malls and can be invested after judging some local factors. Investors also must take a note that they should not try to invest near southern Delhi, as there is very high concentration of malls in this area. Some part of western areas very highly concentrated with shopping malls some regions in central part are also less concentrated.

Limitations and Suggestions for Future Research:

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls, there are other factors such as property value, local population, road infrastructure and income of residents that could influence the location decision of a new shopping mall. 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 shopping mall. 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 shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new shopping mall. 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 shopping mall.