COURSERA CAPSTONE PROJECT

Title: Opening a New Shopping Mall in Delhi, India

By: Nancy Mishra

INTRODUCTION

For many people, visiting shopping malls is a great way of relaxing and enjoying on weekends and holidays. People visit shopping malls to buy groceries, dine at restaurants, buy clothes and other fashion accessories and of course watch movies. For kids it is a place to play games and enjoy. Shopping malls are like a one stop destination for everyone to spend some time. For retailers, the central location and a large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers also take advantage of this trend to build shopping malls where there is a huge demand. Delhi, being the Capital of India, has a lot of great shopping malls, and many are being built. Opening shopping malls allows property builders to earn consistent rental income.

Business Problem

Opening a shopping mall is definitely not an easy task. Selecting a perfect location to build a shopping mall so it generates more revenue is one the most important factor. The objective of this project is to analyse and select the best locations in Delhi, India to open a new shopping mall. Using data science methodology and machine learning concepts like clustering, this project aims to provide solutions to answer the business question: In the city of Delhi, India, if a property builder is looking to open a new shopping mall, where would you recommend that they open it?

Target Audience

This project is particularly useful to property developers and investors who are looking to open new shopping malls or investing in them in the Capital city of India, i.e. Delhi.

DATA

To solve the problem, we need the following data:

1. List of neighbourhoods in Delhi. This defines the scope of this project which is limited to the city of Delhi, India.
2. Latitute and Longitude co-ordinates of the neighbourhoods. This required to plot the map and to get the venue data.
3. Venue data, particularly data related to shopping malls. We will use this data to perform clustering of neighbourhoods.

Sources of Data

The data of neighbourhoods of Delhi is taken from Kaggle.com (https://www.kaggle.com/shaswatd673/delhi-neighborhood-data). We will use web scraping technqiues to extract the data from there with the help pandas library of Python. After that we will use Foursquare API to get venues for those neighbourhoods. It will provide many categories of venue data, we particularly want the shopping mall venue data in order to solve our business objective. This project will make use of various data science techniques, from web scraping, data cleaning and wrangling, working with Foursquare API, machine learning

concepts(k-means clustering), and map visualization(Folium). In the next section we will present the methodology where we will discuss the steps taken, the data analysis done and machine learning techniques we used to solve the problem.

METHODOLOGY

Firstly, we need to get the list of neighbourhoods in the city of Delhi, India. The list is available in the Kaggle website database along with the latitudes and longitudes of the neighbourhoods. After adding the database of neighbourhoods along with its latitudes and longitudes we will populate the data into a pandas dataframe. We will have to do some data cleaning and wrangling so as to remove some unwanted values and also remove missing data. It will help in better data analysis and visualization. We will then visualize the data into a map using the Folium package. This will help in getting an idea of the neighbourhoods and to check that our data is correct.

Next we will use the 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 co-ordinates of neighbourhoods in a python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, category, latitude and longitude. With the data, we can examine different venues returned for each neighbourhood and curate the unique values. 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. Since we want to analyse the shopping malls venue, we will filter out the Shopping Mall data in another dataframe.

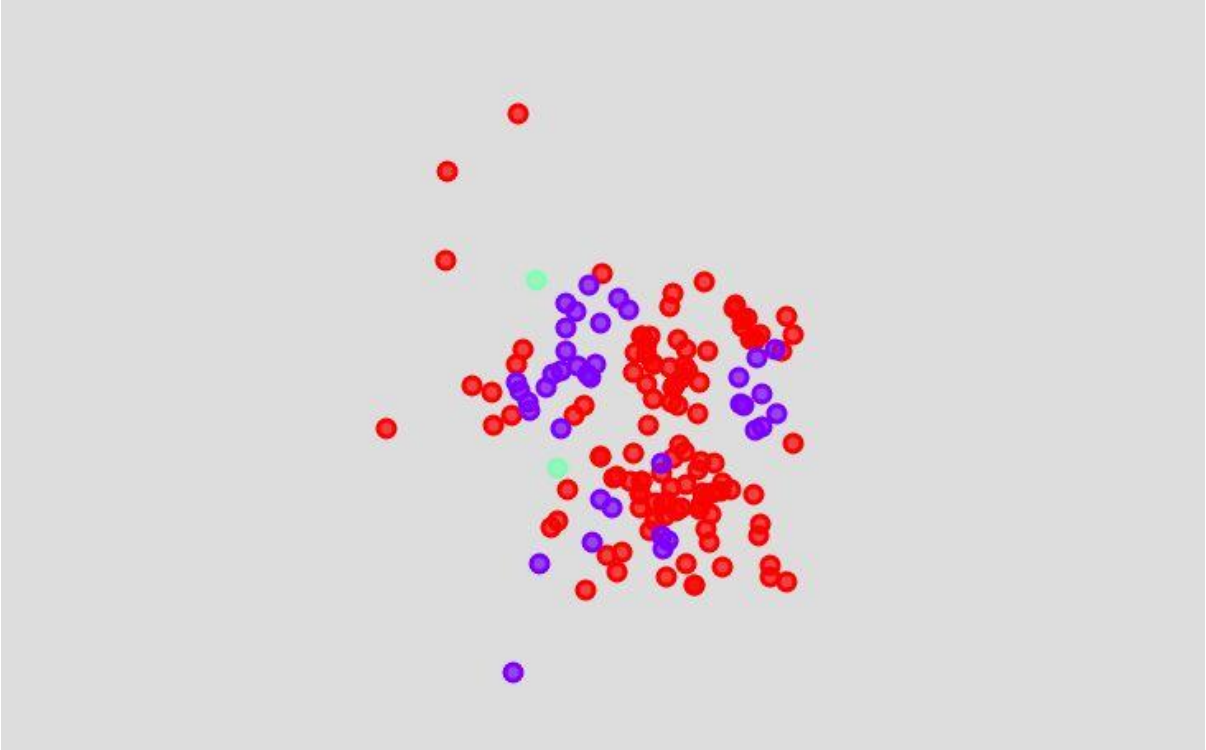
Lastly, we will perform clustering on the data using k-means clustering. K-means clustering algorithm defines k number of centroids and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 “Shopping Mall”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls and which have fewer shopping malls. Based on the occurrence of shopping malls in different neighbourhoods we can answer the question as to which neighbourhood is suited to open a new shopping mall.

RESULTS

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

- Cluster 0: Neighbourhoods with low number to no shopping malls.
- Cluster 1: Neighbourhoods with moderate number of shopping malls.
- Cluster 2: Neighbourhoods with high concentration of shopping malls.

The result of the clustering can be visualized in the following map with cluster 0 in red colour, cluster 1 in purple colour and cluster 2 in mint green colour.



DISCUSSION

As shown in the map, most of the neighbourhoods belong to cluster 0 where there are a very low number to no existence of shopping malls. So, these neighbourhoods are a great place to open up new shopping malls as there would be no competition with other existing malls. Meanwhile, neighbourhoods in cluster 2 have a high concentration of shopping malls which indicates that opening up a new shopping mall in these places would not be a great idea because of a big amount of already existing malls. Competition would be very high in these areas.

So, this project recommends property developers to open a new shopping mall in neighbourhoods that belong to cluster 0.

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 algorithm in clustering the data and lastly providing recommendations to the relevant stakeholders, i.e. property developers and investors regarding the best locations to open a new shopping mall.

The answer to the business question raised in the Introduction section of this report is: The best locations to open new shopping malls are the neighbourhoods of Delhi, India that are under the cluster 0. These neighbourhoods have a very few to no existence of shopping malls, so building a mall in the places would provide good benefits to property developers and investors.