The Battle of Neighborhoods – Study of best neighborhoods to open a new shopping mall in Mumbai, India.

Submitted by: -

Hardik Ruparelia

August 5, 2020

Introduction

A shopping mall is a place which has many modern things in it such as a complex of shops representing types of merchandisers and many walkways that enable the customer to go from unit to unit and to purchase some things. Mostly the shopping malls are been made in the urban areas with big buildings in it and it is a big area that is occupied by many small shops and convert it into a big shopping mall.

From the past there are been many development that has been done in the shopping malls as many entertainment venues like movie theatres restaurants and many things that are been added in a shopping mall as it is the single build structure many things are being constructed in it and it takes a lot of space and the area for the construction that is why it is mostly constructed in the big cities and the only is available in the urban sector.

Therefore, Mumbai being one of the most developed and busiest cities of India accommodates a large number of shopping malls across the city. Opening a shopping mall needs a serious amount of consideration. One of the vital factors for building a shopping mall is the location/area. Location plays major role in determining if a shopping mall will be a failure or success.

Business Problem

The aim of this capstone project is to analyze and select the best location in the city of Mumbai, India to open a new shopping mall.

Question: If a property developer is looking to build a shopping mall, where would you recommend that they open it in the city of Mumbai?

By using data science methodology and machine learning techniques like clustering, classification we will try to find the solution to above question.

Data

In order to solve the problem, we require following data:

1) **List of the neighborhoods in Mumbai**.
   Source: https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Mumbai

By the use of web scrapping technique, we extract the data from this webpage. We use python requests and beautifulsoup packages to achieve this.

2) **Latitude and longitude of each neighborhood**
   We use python geocoder package to obtain latitude and longitude of each neighborhood in Mumbai city.
3) **Obtain venue data for each neighborhood**
   We use Foursquare API to get the venues in each neighborhood.

In addition to that we do data cleaning and data wrangling. Also, we use k-mean clustering an unsupervised machine learning algorithm for clustering of data. Finally, we use Folium package to visualize map and to superimpose the different neighborhoods on map.

Methodology

Firstly, we need to get the list of neighborhoods in the city Mumbai. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Mumbai). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods 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 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 Mumbai.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 3000 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 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 “Shopping Mall” data, we will filter the “Shopping Mall” 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

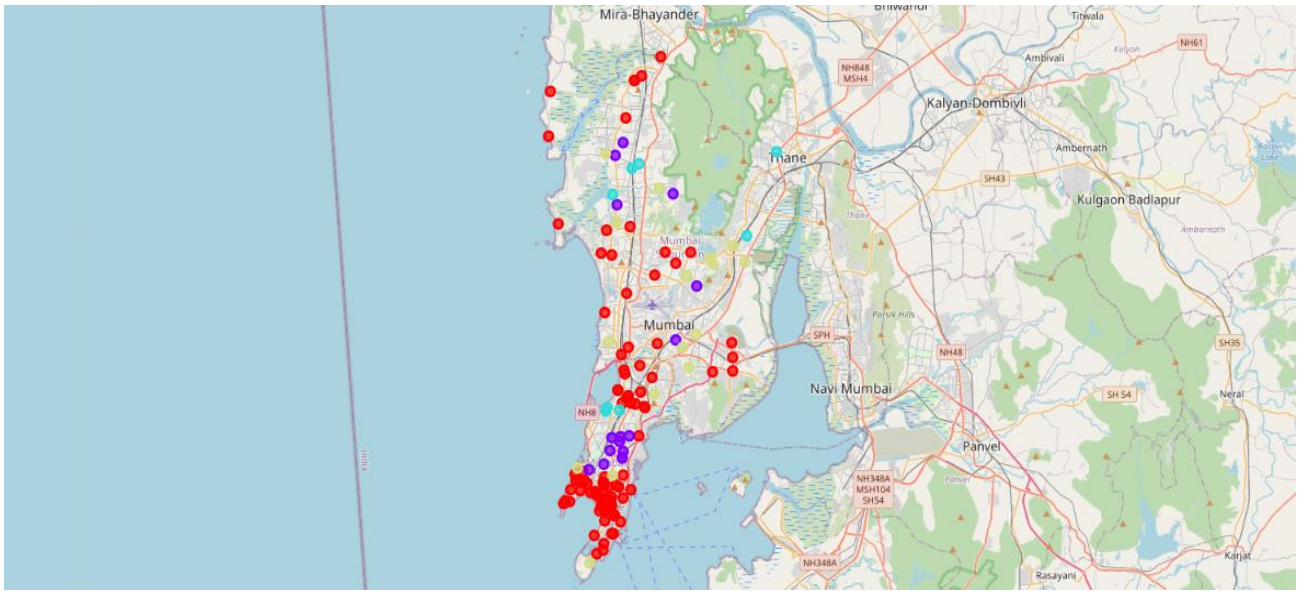
data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 4 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 4 clusters based on the frequency of occurrence for “Shopping Mall”:

- Cluster 1: Neighborhoods with highest number of shopping malls
- Cluster 2: Neighborhoods with moderate number of shopping malls
- Cluster 3: Neighborhoods with low number of shopping malls
- Cluster 4: Neighborhoods with moderate number of shopping malls

The results of the clustering are visualized in the map below with cluster 1 in red color, cluster 2 in purple color, cluster 3 in light yellow color and cluster 4 in light blue color.



Discussion

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the south-west area of Mumbai city, with the highest number in cluster 1 and moderate number in cluster 2 and cluster 4. On the other hand, cluster 3 has very low number of shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the south-west of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 3 with little competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 1 which already have high concentration of shopping malls and suffering from intense competition.

The list of neighborhoods in cluster 3 are shown in the following table:

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
17	Byculla	0.037736	2	19.13891	72.93817
59	JB Nagar	0.040000	2	19.19822	72.96171
57	I.C. Colony	0.037975	2	19.01639	72.82901
131	Western Suburbs (Mumbai)	0.030000	2	19.19021	72.85365
111	SPDC Colony	0.030000	2	19.01499	72.83845
123	Uran	0.051282	2	26.86634	80.93884
84	Malabar Hill	0.030000	2	19.18655	72.84842
12	Bandra Kurla Complex	0.030000	2	19.16785	72.83292
105	Powai	0.033898	2	19.01420	72.82776

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 4 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 neighborhoods in cluster 3 are the most preferred locations to open a new shopping mall