Coursera Capstone

IBM Applied Data Science Capstone

Upcoming Commercial Projects in Mumbai

By: Ajinkya V Pednekar

July 2020

Table of Contents

1. Introduction

The importance of the concept of sustainability is increasing every day; while the world population is increasing, the demand for shopping is increasing as well. For many people, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of Mumbai and many more are being built. Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall 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 Mumbai, 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 Mumbai, India if a property developer is looking to open a new shopping mall, 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 shopping malls in the economic capital city of India i.e. Mumbai. This project is timely as the city is currently suffering from shortage of decent malls. Data from the Business news released in 2015 showed that the India has only 77.6 million square feet of mall space, less than one tenth of U.S. levels, despite having nearly four times the population, with the shortage of attractive malls most acute in New Delhi and Mumbai.

2. Data

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

- List of neighborhoods in Mumbai. This defines the scope of this project which is confined to the city of Mumbai, the economic capital city of the country of India in South Asia.
- Latitude and longitude coordinates of those neighborhoods. 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 neighborhoods.

Sources of data and methods to extract them This Wikipedia page

(https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai) contains a list of neighborhoods in Mumbai, with a total of 93 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 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.

3. Methodology

Firstly, we will obtain list of neighborhoods in the city of Mumbai. Fortunately, the list is available on the Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai). We will do web scraping using Python requests and BeautifulSoup packages to extract the list of neighborhoods data. However, this is a list of names, locations and co-ordinates. We can also 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 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 5000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key and also access token. 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 3 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.

4. 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 “Shopping Mall”:

- Cluster 0: Neighborhoods with very low to no existence number of shopping malls
- Cluster 1,2,3: Neighborhoods with moderate number of shopping malls
- Cluster 4: Neighborhoods with high concentration of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, cluster 2 in blue color, cluster 3 in mint green color, and cluster 4 in orange color.

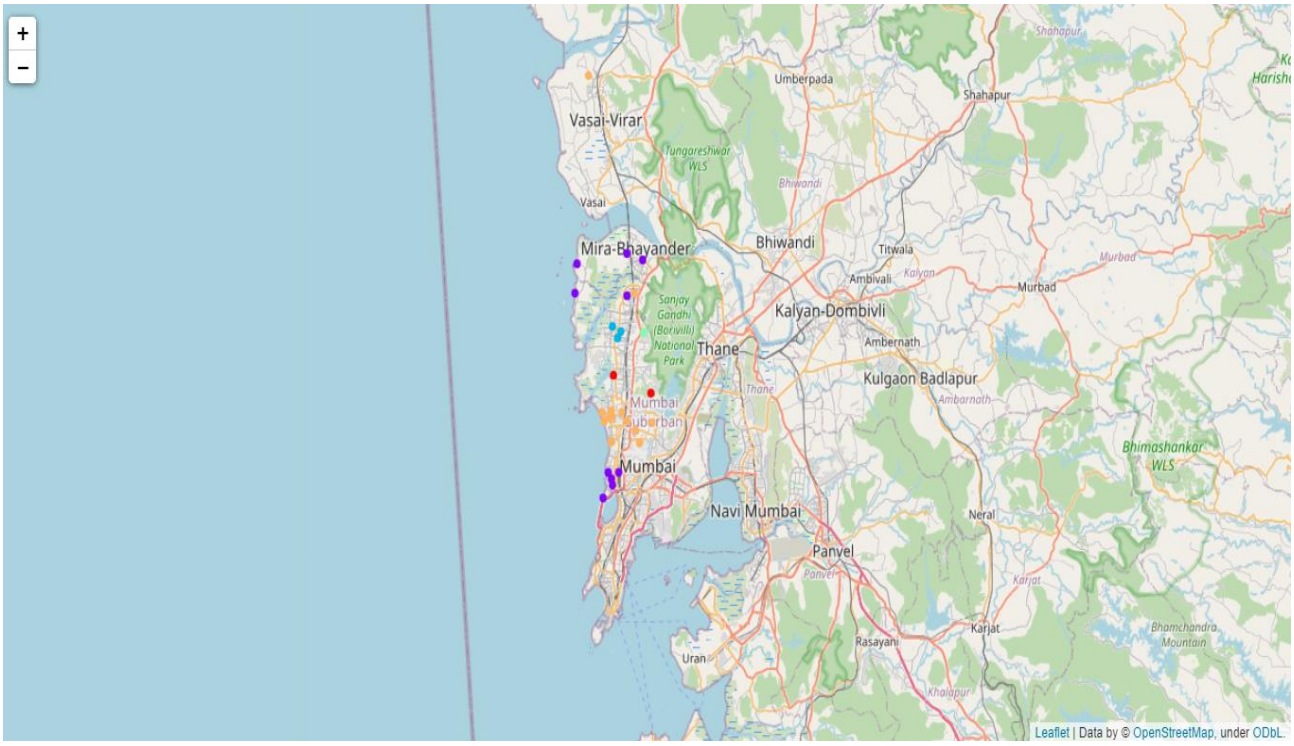

Fig 1. Cluster of places according to probability of mall

Examine the Clusters:

Cluster 0:

```
In [30]: mumbai_merged.loc[mumbai_merged['Cluster Labels'] == 0]
```

Out[30]:

	Neighborhood	Shopping Mall	Cluster Labels	Area	Latitude	Longitude
3	Goregaon,Western Suburbs	0.015	0	Bangur Nagar	19.167362	72.832252
3	Goregaon,Western Suburbs	0.015	0	Aarey Milk Colony	19.148493	72.881756

Cluster 1:

```
In [31]: mumbai_merged.loc[mumbai_merged['Cluster Labels'] == 1]
```

Out[31]:

	Neighborhood	Shopping Mall	Cluster Labels	Area	Latitude	Longitude
6	Khar,Western Suburbs	0.0	1	Khar Danda	19.068598	72.840042
2	Borivali (West),Western Suburbs	0.0	1	Gorai	19.250057	72.782021
2	Borivali (West),Western Suburbs	0.0	1	I.C. Colony	19.247039	72.84983
1	Bandra,Western Suburbs	0.0	1	Pali Hill	19.068	72.826
1	Bandra,Western Suburbs	0.0	1	Bandstand Promenade	19.042718	72.819132
1	Bandra,Western Suburbs	0.0	1	Kherwadi	19.0553	72.8314
7	Mira-Bhayandar,Western Suburbs	0.0	1	Bhayandar	19.29	72.85
7	Mira-Bhayandar,Western Suburbs	0.0	1	Uttan	19.28	72.785
7	Mira-Bhayandar,Western Suburbs	0.0	1	Mira Road	19.284167	72.871111
6	Khar,Western Suburbs	0.0	1	Pali Naka	19.062742	72.829396

Cluster 2:

Cluster 2

```
In [32]: mumbai_merged.loc[mumbai_merged['Cluster Labels'] == 2]
```

Out[32]:

	Neighborhood	Shopping Mall	Cluster Labels	Area	Latitude	Longitude
5	Kandivali West,Western Suburbs	0.01	2	Mahavir Nagar	19.211319	72.842737
5	Kandivali West,Western Suburbs	0.01	2	Charkop	19.216182	72.830575
5	Kandivali West,Western Suburbs	0.01	2	Poisar	19.204511	72.837639

Cluster 3:

Cluster 3

```
In [33]: mumbai_merged.loc[mumbai_merged['Cluster Labels'] == 3]
```

Out[33]:

	Neighborhood	Shopping Mall	Cluster Labels	Area	Latitude	Longitude
4	Kandivali East,Western Suburbs	0.02	3	Thakur village	19.210206	72.87298

Cluster 4:

```
In [34]: mumbai_merged.loc[mumbai_merged['Cluster Labels'] == 4]
```

Out[34]:

	Neighborhood	Shopping Mall	Cluster Labels	Area	Latitude	Longitude
8	Western Suburbs	0.00250	4	Chakala, Andheri	19.111388	72.860833
8	Western Suburbs	0.00250	4	Dahisa	19.250069	72.859347
8	Western Suburbs	0.00250	4	Jogeshwari West	19.12	72.85
8	Western Suburbs	0.00250	4	Juhu	19.1	72.83
0	Andheri,Western Suburbs	0.00375	4	Amboli	19.1293	72.8434
0	Andheri,Western Suburbs	0.00375	4	Versova	19.12	72.82
0	Andheri,Western Suburbs	0.00375	4	Seven Bungalows	19.129052	72.817018
0	Andheri,Western Suburbs	0.00375	4	Sahar	19.098889	72.867222
0	Andheri,Western Suburbs	0.00375	4	Marol	19.119219	72.882743
0	Andheri,Western Suburbs	0.00375	4	Lokhandwala	19.130815	72.82927
0	Andheri,Western Suburbs	0.00375	4	Four Bungalows	19.124714	72.82721
0	Andheri,Western Suburbs	0.00375	4	D.N. Nagar	19.124085	72.831373
8	Western Suburbs	0.00250	4	Virar	19.47	72.8
8	Western Suburbs	0.00250	4	Vile Parle	19.1	72.83

5. Discussion

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the western area of Mumbai city, with the highest number in cluster 4 and moderate number in cluster 2. On the other hand, cluster 1 has very low number to totally no 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 4 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, this also shows that the oversupply of shopping malls mostly happened in the western area 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 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 0,2 and 3 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 4 which already have high concentration of shopping malls and suffering from intense competition.

6. 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 population 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 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 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.

7. 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 5 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 0 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 making them decide to open a new shopping mall.