Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Shopping Mall in Mumbai, India

By: Niladri Deb

March 2021

Introduction

We know that Shopping Malls are not made just for the purpose of shopping. Shopping Malls provide much more than that, making it a great place to for people to hang out with their friends and families on Weekends and Holidays. Not only does it provide a vast variety of stores for purchasing a variety of clothes, electronics, groceries, and much more, it also provides entertainment facilities like Cinema Theatres for watching movies, Amusement Rides for young children, and many restaurants. Often a lot of retailers come to malls and set up events to promote products, movements, art. Sometimes the malls even decorate for the celebrations of festivals like Christmas, Diwali, St. Patrick's Day, etc. With so many benefits, property developers take advantage of these trends and go ahead with building more shopping malls, as it is also a source of rental income too. I live in Mumbai, the financial capital of India, and it is a home to some shopping malls. These are excellent malls with so many stores, services and various events going on throughout the year. High Street Phoenix is a home to 3 buildings, making it a shopping mall complex with many restaurants, stores, bars and a night club too. Though, it is considered to be the best shopping mall of the city as well as country, it's location could be an issue for many people to visit this stunning venue becomes a little inconvenient. For this reason, the property developers made another mall, but this time with name 'Phoenix Market City' at another region of Mumbai, but it was not very successful because the other malls that were quite good were really close to that region and there was a lot of competition. Opening a new mall requires a lot of considerations to be taken seriously and what makes it more complicated is that there are a lot of factors such as location to make a decision about opening new malls. Location plays a very important role and it will be very interesting to explore this.

Business Problem

The objective of this capstone project is simple – to analyse and then determine the best possible locations for opening a mall in Mumbai. By using machine learning and data science methods like clustering, this project aims to provide an answer to the following question: If a property developer is looking to open a mall in Mumbai, India, where would you recommend them to open one?

Target Audience of This Project

This project is going to be helpful for the property developers in Mumbai who are looking to open new shopping malls in Mumbai. I feel that this project is an appropriate one according to the time period because According to Economic Times, India is looking to add 100 more malls in the major cities of India such as Mumbai, Bangalore, Delhi, Chennai, Kolkata, etc. It is also mentioned in the article that according Anardock MD & CEO, the Western part of India (The region under which Mumbai comes) will get 36 new malls, with 18 of them that will be constructed at Mumbai. Additionally, with a median age of only 26.8, and with 50% of the population being under the age of 25, and with the job opportunities available in Mumbai, it will attract a large amount of the Indian population and new malls have a chance to succeed really well, provided they can be built in locations where their chances of success are higher.

Data

To do this project we will need the following data:

1. List of Neighborhoods in Mumbai. This will define the scope of the project – which is going to be limited to Mumbai, the financial capital of India.
2. The Latitude and Longitude co-ordinates of the neighborhoods. This will be very helpful in plotting the neighborhoods on the map and get the data of the venues.
3. Venue data, especially of shopping malls. This will be useful and important for the clustering part of the project.

Sources of Data and How to extract them

The Wikipedia page with the list of neighborhoods in Mumbai (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai) contains the list of the neighborhoods in Mumbai, nearly 95 of them. Since, the table of the longitude and latitude is already available, we will use geocoder to recheck the co-ordinates.

After that we will use the Foursquare API to get the venue data for those neighborhoods. Foursquare has a huge database of more than a 105 million places and is used by nearly 120,000 developers worldwide. While Foursquare provides a huge amount of venue data for neighborhoods, we will be specifically looking into the venue data for shopping malls in the neighborhoods of Mumbai. This project will be requiring some data science skills in terms of working with an API like Foursquare, data cleaning, data wrangling, machine learning in terms of K-Means Clustering and data visualization on maps using Folium.

Methodology

First, I had imported all the packages, and installed some of them. Then I needed the data on the list of the neighborhoods in Mumbai and I was lucky enough because I find this on a Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai) along with the latitudes and longitudes data. Next to extract the data was fairly simple – as a table was available, I used the knowledge of python where if you count in the negative order, the computer counts backwards (e.g., In loops) and extracted the table from the page, following which I stored it in a Data Frame. However, I didn't completely depend on the data for the co-ordinates on Wikipedia because numerical data can be more easily manipulated than data in the form of words, so I decided to find the co-ordinates of the neighborhoods using Geocoder. And on comparing I found that even though the data was less precise, it was a little different from that on the Wikipedia page (you can see it on the Notebook as some values were different due to the differences of the number from the 2nd or 3rd number after the decimal point). Then the co-ordinates that were extracted from the Wikipedia page were dropped and were replaced by the ones that were extracted using Geocoder. Then we plot this data on a map of Mumbai using the Folium package. This also helps us to check the co-ordinates of the neighborhoods are correctly stored in the Data frame and are correctly plotted on the map.

Now comes in our Foursquare API. We use this to retrieve the top 100 venues that are within a radius of 2 km. Before I used the Foursquare API, I had to first register a developer account and from there I made an app that gave me a 'Client ID' and 'Client Secret' which later helped me in making the API call using the co-ordinates of the neighborhoods. To pass the co-ordinates of all neighborhoods, I made a loop using Python and passed in the co-ordinates of the neighborhoods until there were no neighborhoods left in the Data Frame. Then the Foursquare API returned the data of the venues in the form of a JSON file. From this JSON file I extracted the name of the venue, its category and its co-ordinates. Now with this returned data, we can check the number of venues that were returned for each neighborhood and we can also check how many categories each of these returned venues could be categorized into. After this, I had created a Data Frame with each of those neighborhoods, the venue categories and the values of 'one-hot encoding' for each of those categories

for that neighborhood. After that, the mean of the frequencies of each of those categories in each of those neighborhoods were calculated and the Data Frame was adjusted accordingly. Since Shopping Malls are our priority for this project, we create a new Data Frame with the name of the neighborhoods, the values for 'Shopping Malls' category, and then we appended the values for the latitudes and longitudes after clustering the data (explained next).

Lastly, we cluster the data points using 'k-means clustering'. As I had explained in the notebook, 'k-means clustering' is an unsupervised machine learning technique for clustering data where we enter the number of desired centroids, or the value for k, based on which the each of the data points are grouped into groups with the nearest clusters and then the mean/ a new centroid for this cluster is calculated. This continues as a loop until the mean calculated for each of the clusters don't change. And as the name suggest, we don't supervise this technique. We apply it and then check the final results. Since I planned on keeping this project simple, I had declared k as 3 and the frequency of the 'Shopping Malls' data. This will be helpful in understanding which neighborhoods have a higher frequency of Shopping Malls, which have a moderate frequency and which have a lower frequency. This will help us identify the neighborhoods that would be suitable for building new malls.

Next, we plotted the clusters on a map of Mumbai using the Folium package again to visualize the data. Then the Data Frame was displayed for each of the clusters and based on which an 'Observation and (Brief) Analysis' was done.

Results

The following is the result of applying ‘k-means clustering’ based on the frequency value for ‘Shopping Malls’:

Cluster Label	Remark
0	No presence of Shopping Malls
1	Presence of moderate number of Shopping Malls
2	High presence of Shopping Malls

The above clusters have been plotted on the map below where Cluster 0 is indicated in red, Cluster 1 in purple and Cluster 2 in green.

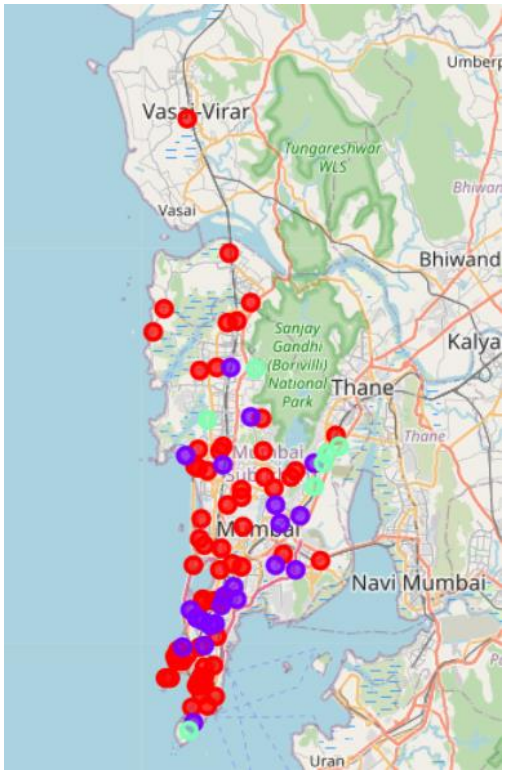

Figure 1 – Complete Map of Mumbai with all the clusters

As you can see Mumbai is a big city and how congested the map looks like. Hence, I have taken zoomed in, and taken 3 screenshots to show how spaced out the clusters are, starting with Figure 2 (Southern part of Mumbai), Figure 3

(a little higher than Southern part of Mumbai) and Figure 4 (Northern Part of Mumbai).

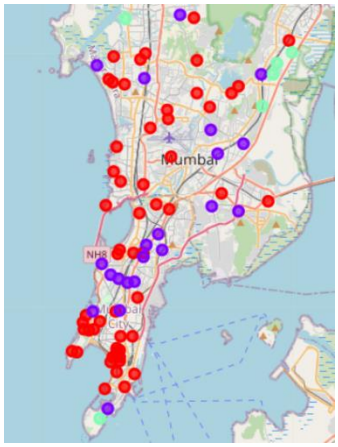

Figure 2 – Southern Part of Mumbai with all the clusters

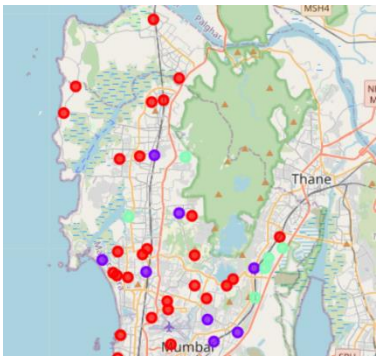

Figure 3 – A little to the North of Southern Part of Mumbai with all the clusters

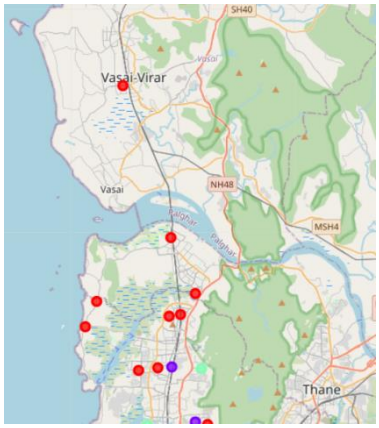

Figure 4 – Northern Part of Mumbai with all the clusters

Discussion

This part of the report is similar to the 'Observation and (Brief) Analysis' part of the Notebook. The areas concentrated with Shopping Malls in the city of Mumbai are scattered across the city. We can see that Clusters 1 and 2 show the presence of Shopping Malls in the neighborhoods while Cluster 0 shows no presence of Shopping Malls. Now let's discuss about the clusters for the relevant stakeholders, in our case property developers. The neighborhoods in Cluster 0 would be highly recommended for building new Shopping Malls. The absence of Shopping Malls in these neighborhoods would mean that there will be no competition and the chances of succeeding/gaining profits from a Shopping Mall would be higher. Next, though neighborhoods in Cluster 1 would not be highly recommended to build a new Shopping Mall, but it would depend on the features and some other factors of the Shopping Mall that would decide if it would succeed or fail in these neighborhoods. As I have kept this project simple by considering only the frequency of already built Shopping Malls in the neighborhood, I will not discuss about the other factors/features that could help a Shopping Mall succeed or fail. Lastly, it would be highly recommended to the property developers to avoid building a new Shopping Mall in the neighborhoods that fall under Cluster 2. There is a high frequency and many Shopping Malls present in these neighborhoods and the competition would be very high with the already built Shopping Malls.

Limitations and Future Scope

As I had kept this project a simple one so I just considered the frequency of the Shopping Malls in the neighborhoods or the location of other Shopping Malls for this project. But, when an idea for building a Shopping Mall in a neighborhood is put forward, along with the location and frequency of Shopping Malls in that neighborhood, a lot of other factors such as population of the neighborhood, the average income of the residents which will help us classify the neighborhoods as rich, poor, etc. Also, we must consider the age of the average age of the residents in the neighborhood as well as the lifestyle choices (whether one big place is preferred for shopping, or many small places are preferred for shopping), the features of the malls, the unique benefits of

the malls are some that come to my mind when I think of the idea of building new Shopping Malls in the neighborhoods. But of course, there are many more reasons and that leads to a lot of future scope. Unfortunately, it was difficult to find these data online because usually they are encrypted or kept in a manner such that it is not easily accessible to the general public. Further improvements that could be made is that one can do this project studying more than one variable, instead of only one factor like only location/frequency of other malls like I did. Also, one can do more plots like bar plots, line plots, box plots, etc. to make the data more visual and it could also help in understanding the data trend/pattern. Also, different methods of clustering could be used along with some different methods of classification and it could be seen how it affects the results. Last but not the least, I had used a free Foursquare API Developer's Account as well as a free IBM Cloud Account which inhibited the number of calls that were made using the API and the number of times the Notebook could be accessed in a month. If a paid account could be used, it could maybe yield more efficient results.

Conclusion

In this Capstone Project, the business problem was identified, for which we were able to specify the data was needed. Then the data was collected by extraction and preparation, followed by using an unsupervised method of clustering 'k-means clustering' to cluster the data depending on their similarities. Finally, the results and recommendations were conveyed to the relevant stakeholders, in our case property developers about the neighborhoods where a new Shopping Mall could be built. But to summarize the answer for the business question in the Introduction section: Neighborhoods in Cluster 0 will be highly recommended to build a new Shopping Malls. The relevant stakeholders or the property developers will be able to utilize the results that were found in this project and could build a new Shopping Mall in those neighborhoods which already have a lot and could avoid a lot of competition.

References

Neighborhoods in Mumbai:
https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai

Foursquare Developer Documents:
https://developer.foursquare.com/docs

Median Age of India:

https://www.statista.com/statistics/254469/median-age-of-the-population-in-india/#:~:text=The%20median%20age%20in%20India,38.1%20years%20old%20by%202050.&text=India%20has%20the%20second%20largest%20population%20in%20the%20world%2C%20after%20China.

Information on New Malls development in India:
https://economictimes.indiatimes.com/industry/services/property-/-cstruction/indian-cities-to-add-100-new-malls-by-2022-end-says-anarock/articleshow/73518677.cms?from=mdr