IBM Data Science Professional Capstone Project:

Applied Data Science

Battle of Neighbourhoods

Final Project Report

Alok Mishra

https://www.linkedin.com/in/mishralokk/

Table of Content

Executive Summary

Mumbai is known as the financial capital of the world's second most populous country India. The city is victim of long traffic snarls and congestion during peak hours. The onus can be given to the places causing huge gatherings like markets, business clusters and shopping malls.

This project picks up the plausible opening of shopping malls in Mumbai as a business problem where the efforts will be to find a solution using data science techniques and python.

This project report discusses the possibilities of opening a new shopping mall in Mumbai by using data science technologies and wonderful python packages. The methodology used in the project will help to locate a best region within Mumbai which is not only profitable by avoiding unnecessary competition and increase revenue for investors and developers but also help decongesting the traffic issues at Mumbai streets.

Introduction

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. Reason being their versatility covering shopping, eating, entertainment and other leisure activities under one roof.

Where at one hand shopping malls can be termed as a one-stop destination for all types of shoppers on other 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 with several under construction. Opening shopping malls allows Realty developers to earn consistent rental income.

As per live mint media, the recent order passed by Maharashtra government regarding 24x7 opening of mall may further attract realty developers. Though, 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 crucial decisions for the success of a mall.

Business Problem

The objective of this capstone project is to analyse 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 the city of 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 financial capital of India i.e. Mumbai. This project is also viable as it has the potential to help the traffic congestion in the city.

Parking problems is very common in India and vehicles parked outside the mall make the traffic situation only worse on Mumbai roads.

As per senior traffic official, *"Since all the shopping malls have food courts on the premises, the vehicles are parked right at the doors of the malls, congesting the roads adjacent to the shopping hubs, creating unnecessary traffic snarls."*

Data

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

1. **List of neighbourhoods in Mumbai:** defines the scope of this project which is confined to the city of Mumbai, the financial capital of the country India in Asia.
2. **Latitude and longitude coordinates:** of those neighbourhoods which is required in order to plot the map and also to get the venue data.
3. **Venue data:** specifically, the data related to the shopping malls. We will use this data to perform clustering on the neighborhood.

Sources of data and Methods of Extraction:

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Mumbai) contains a list of neighbourhoods in Mumbai, with a total of 42 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and **beautiful soup**[1] **package**.

Then we will get the geographical coordinates of the neighborhood using Python **Geocoder package**[2] which will give us the latitude and longitude coordinates of the neighbourhoods

Foursquare[3] **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 where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique(s) that was used in the project.

Methodology

Step by step methodology can be discussed here:

1. We will start with fetching the list of neighbourhoods in Mumbai. Fortunately, the list is available on the Wikipedia (https://en.wikipedia.org/wiki/Category:Suburbs_in_Mumbai). We will do web scraping using Python requests and beautiful soup packages to extract the list of neighborhoods data.
2. 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.
3. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using **Folium package**[4]. 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.
4. 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.
5. 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 neighbourhood and examine how many unique categories can be curated from all the returned venues.
6. 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 neighborhood.
7. Lastly, we will perform clustering on the data by using **k-means clustering**[5]. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Shopping Mall".

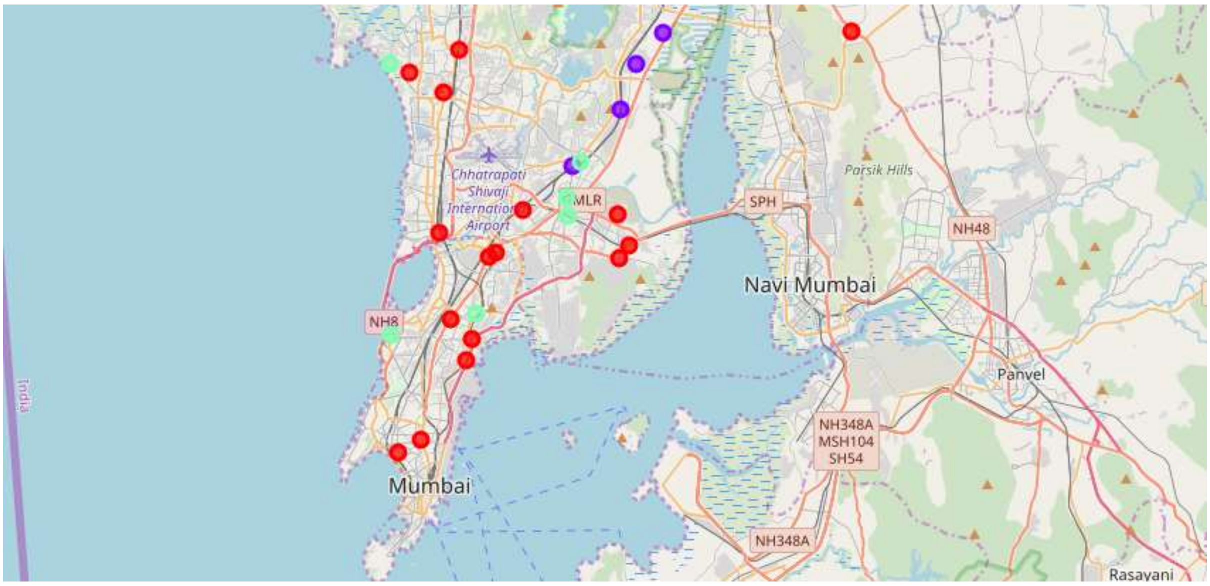
The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls.

Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls

Results

The results from the k-means clustering show that we can categorize the neighborhood into 3 clusters based on the frequency of occurrence of "Shopping Mall"

1. **Cluster 0**: Neighbourhoods with zero to negligible number of shopping malls
2. **Cluster 1:** Neighbourhoods with low number to no existence of shopping malls
3. **Cluster 2:** Neighbourhoods with high concentration of shopping malls



The results of the clustering are visualized in the map above with-

cluster 0 in RED COLOUR,

cluster 1 in PURPLE COLOUR, and

cluster 2 in MINT GREEN COLOUR.

Discussion

Most of the shopping malls are concentrated in the Westen and Southern areas of Mumbai city, with the highest number in cluster 1 followed by cluster 2. On the other hand, cluster 0 has zero to negligible presence of shopping malls in the neighborhoods. It opens the door of opportunities for 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 malls. From another perspective, this also shows that the oversupply of shopping malls mostly happened in the western, with the eastern and central areas still have very few shopping malls.

Therefore, this project recommends real estate developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 0 with practically negligible competition. Realty developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 2 with moderate competition.

Lastly, property developers are advised to avoid neighborhoods in cluster 1 which is already pretty crowded with shopping malls as it's not a great idea from investment point of view. Additionally, it will also worsen the existing traffic problem of Mumbai City.

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 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 better insights and 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. Realty developers and investors regarding the best locations to open a new shopping mall.

To answer the business question propounded in the introduction section, the answer proposed by this project is: The neighbourhoods 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 in their decisions to open a new shopping mall.

References

Navalkar P., (August 2019). Freepressjournal.in [online] Available at https://www.freepressjournal.in/mumbai/no-parking-outside-malls-for-app-based-food-delivery-boys [Accessed 30 Jan. 2020].

Python Package Index. [online] Available at https://pypi.org/ [Accessed 30 Jan. 2020].

Sapam B., (January 2020). [online] Available at https://www.livemint.com/news/india/mumbai-s-enhanced-nightlife-fails-to-enthuse-city-s-mall-owners-11580018616872.html [Accessed 30 Jan. 2020].

Wikipedia.org [online] Available at https://en.wikipedia.org/wiki/Category:Suburbs_in_Mumbai [Accessed 30 Jan. 2020].

Appendix:

1. **Beautiful Soup** is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping
2. **Geocoder Package** is a Simple and consistent *geocoding* library written in Python
3. **Foursquare** has one of the largest databases of 105+ million places and is used by over 125,000 developers.
4. **Folium Package** *folium* builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the Leaflet.js library.
5. **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.