IBM Applied Data Science Capstone Project

Opening a New Restaurant in Mumbai, India

Introduction

For many people, dining at a nice restaurant is a great way to relax and enjoy themselves during weekends and holidays. They can dine at restaurants, do grocery shopping, shop at the various fashion outlets, watch movies and perform many more activities. For restaurant owners, the central location and the large crowd at a junction provides a great oppurtunity to make good business. Property developers are also taking advantage of this trend to build more restaurants to cater to the demand. As a result, there are many restaurants in the city of Mumbai and many more are being built. Opening different types of cusines allows people to earn consistent income. Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

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 restaurant. 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 person is looking to open a new restaurant, 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 restaurant in the business capital of India i.e. Mumbai. Being the business capital of India it attracts people from all over India, which in itself is a coutry of diverse culture and tastes, and abroad. With lots of different people comes lots of different eating habits, which means a single restaurant can never be the best at all the items that it serves. Hence we have different types of restaurants which are particulary famous for their special type of food like continental, chinese, etc.

Data

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

- List of neighbourhoods in Mumbai. This defines the scope of this project which is confined to the city of Mumbai, the capital city of Maharashtra, India.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to restaurants. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This html page (https://mumbai7.com/postal-codes-in-mumbai/) contains a list of neighbourhoods in Mumbai. We will use web scraping techniques to extract the data from the html page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database 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 different Restaurant categories 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, 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.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Mumbai. Fortunately, the list is available in the html page (https://mumbai7.com/postal-codes-in-mumbai/). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods 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 neighbourhoods 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 500 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. 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. As there would different categories of restaurants like Fast Food Restaurant, Chinese Restaurant, Vegetarian Restaurant, etc we will be replacing all the venue categories in the data frame which have a string ‘restaurant’ in them with ‘Restaurant’. 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 “Restaurant” data, we will filter the “Restaurant” as venue category for the neighbourhoods.

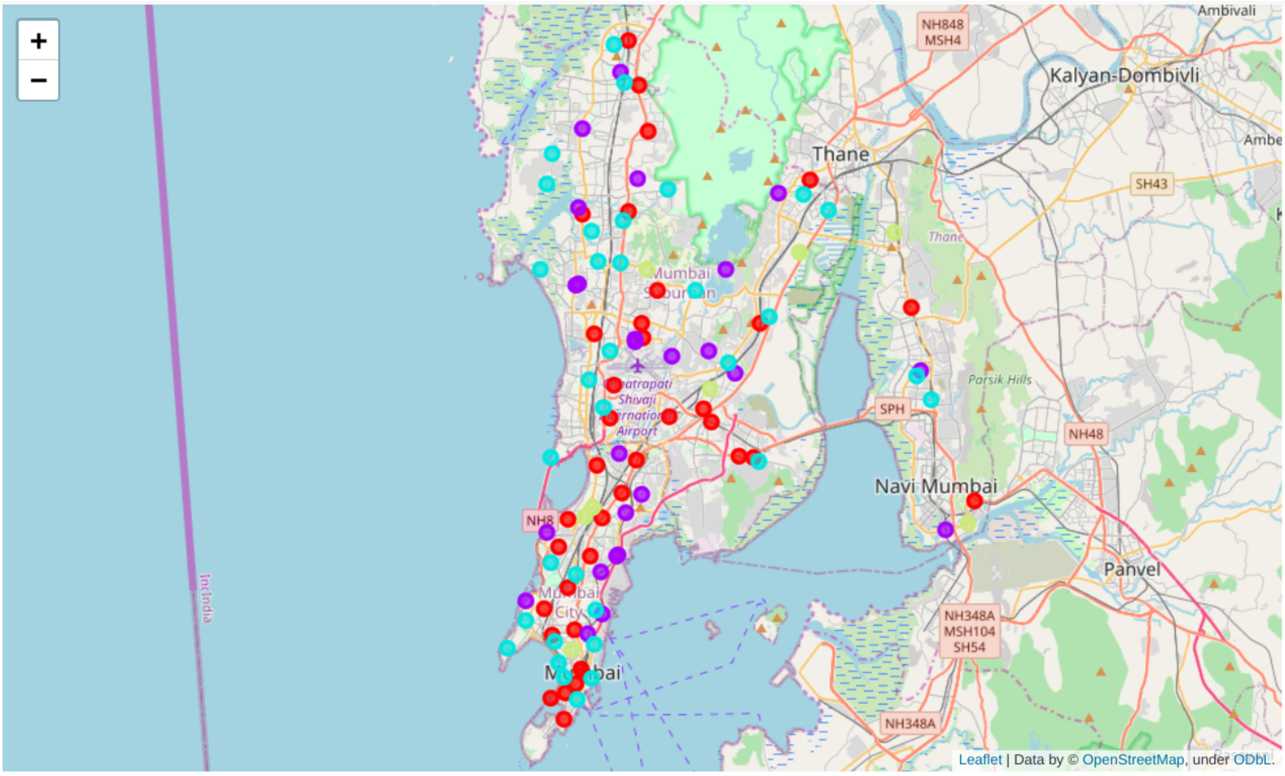
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 4 clusters based on their frequency of occurrence for “Restaurant”. The results will allow us to identify which neighbourhoods have higher concentration of restaurants while which neighbourhoods have fewer number of restaurants. Based on the occurrence of restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new restaurants.

Results

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

- Cluster 0: Neighbourhoods with 30 to 50 percent of venues being restaurants
- Cluster 1: Neighbourhoods with low number to no existence of restaurants
- Cluster 2: Neighbourhoods with 15 to 30 percent of venues being restaurants
- Cluster 3: Neighbourhoods with high concentration of restaurants

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, cluster 2 in light blue colour and cluster 3 in dark yellow color.



Discussion

As observations noted from the map in the Results section, most of the restaurants are concentrated in the central area as well as area near to the sea in Mumbai city, with the highest number in cluster 3 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no restaurant in the neighbourhoods. This represents a great opportunity and high potential areas to open new restaurants as there is very little to no competition from existing restaurants. From another perspective, the results also show that most of the restaurants are in the central area of the city, with the suburb area still have very few restaurants. Therefore, this project recommends people to capitalize on these findings to open new restaurants in neighbourhoods in cluster 1 with little to no competition. People with idea to built great dining place to stand out from the competition can also open new restaurants in neighbourhoods in cluster 2 with low competition. Lastly, people are advised to avoid neighbourhoods in cluster 3 which already have high concentration of restaurants and suffering from intense competition.

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 restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new restaurant. 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 restaurant.