Coursera Capstone

IBM Applied Data Science

Opening A New Chineese Restaurant in Mumbai,India

By :Sumanta Kumar Muduli

28th June 2020

Introduction:

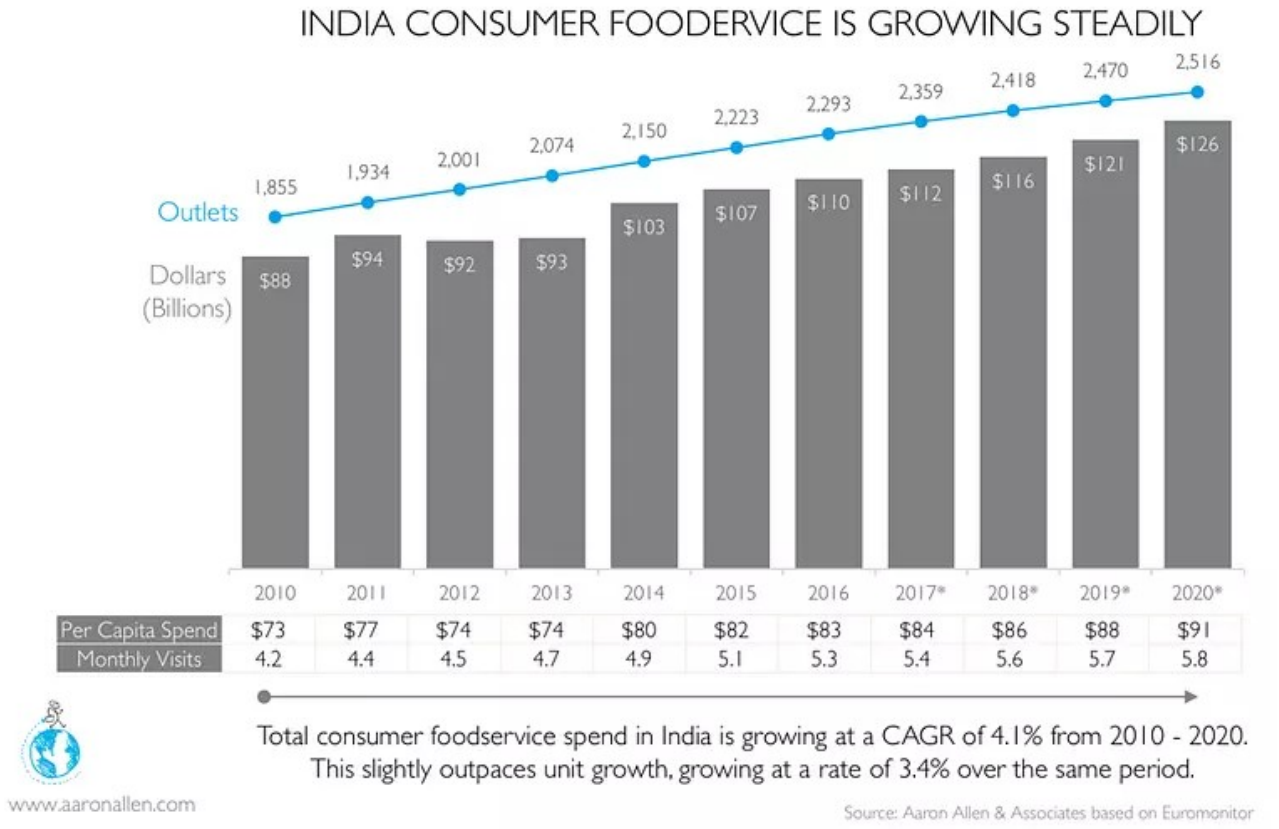
Mumbai is the financial capital of India.It has highest hotel and restaurant industry in the country due to presence of measure financial activities in the city like different production industry,Bollywood industry,Educational institutions,Bank industry.This industry is also expanding at a very fast rate due to the high profit involvment and increasing demand.Property developers are also taking advantage of this increasing demand trend to build more restaurants to cater the demand of investors.As a result of which there are many restaurants in the city and many more are being built.Of course like many businesses, opening a new restaurant requires serious considerations.This is a very complicated problem than it seems.Particularly the selection of location is one one of the most important decision which will decide whether the restaurant is a success or a failure.

Business Problem:

The objective of this capstone project is to analyse and select the best location to open a restaurant in Mumbai city,india.Using machine learning techniques like clustering this project aims to provide the solution to business question :In the city of Mumbai ,India some one is looking to open a new restaurant, where would you recommend him to open it?¶

Target audiance of this project:

This project is basically usefull for the investors who are looking to set a new restaurant in Mumbai city.



As per the trend shown in above figure published by AARONALLEN and ASSOCIATES a global restaurant consultancy company , india's consumer food service is growing steadily.

Data:

To solve the problem,we need the foolowing data-
1.list of the neighbourhoods in Mumbai city.
2.latitude and longitude coordinates of those neighbourhoods.This is required in order to plot the map and also to get the venue data.
3.venue data,particularly data related to restaurants.we will use this data to perform clustering on neighbourhoods.
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 neighbourhoods of Mumbai,India. We will use web scrapping techniques to extract the data from wikipedia website.As in that page all the neighbourhoods with their coordinates exists in tabular form ,we could directly extract these with pandas library.
After that we will use foursquare API to get the venue data of those neighbourhood .Foursquare API will provide many categories of venue data,We particularly interested in restasurants data category in order to help us in solving the business problem.
This is a project that will make use of various data science skills from web scrapping ,data cleaning,working with Foursquare API,data wrangling,data visualization(folium) to machine learning skills.

Methodology:

First we need to get the list of neighbourhoods of Mumbai city,India.Fortunately,it is available in the wikipedia page(‘https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai’) with respective coordinates.We now have to do web scrapping to covert that table into pandas data frame.We need to get the geographical coordinates in the form of latitude and longitude in order to use Foursquare API.To do so we will use wonderfull 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 pandas data frame and then visualize the neighbourhoods in a map using folium package.This allow 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.

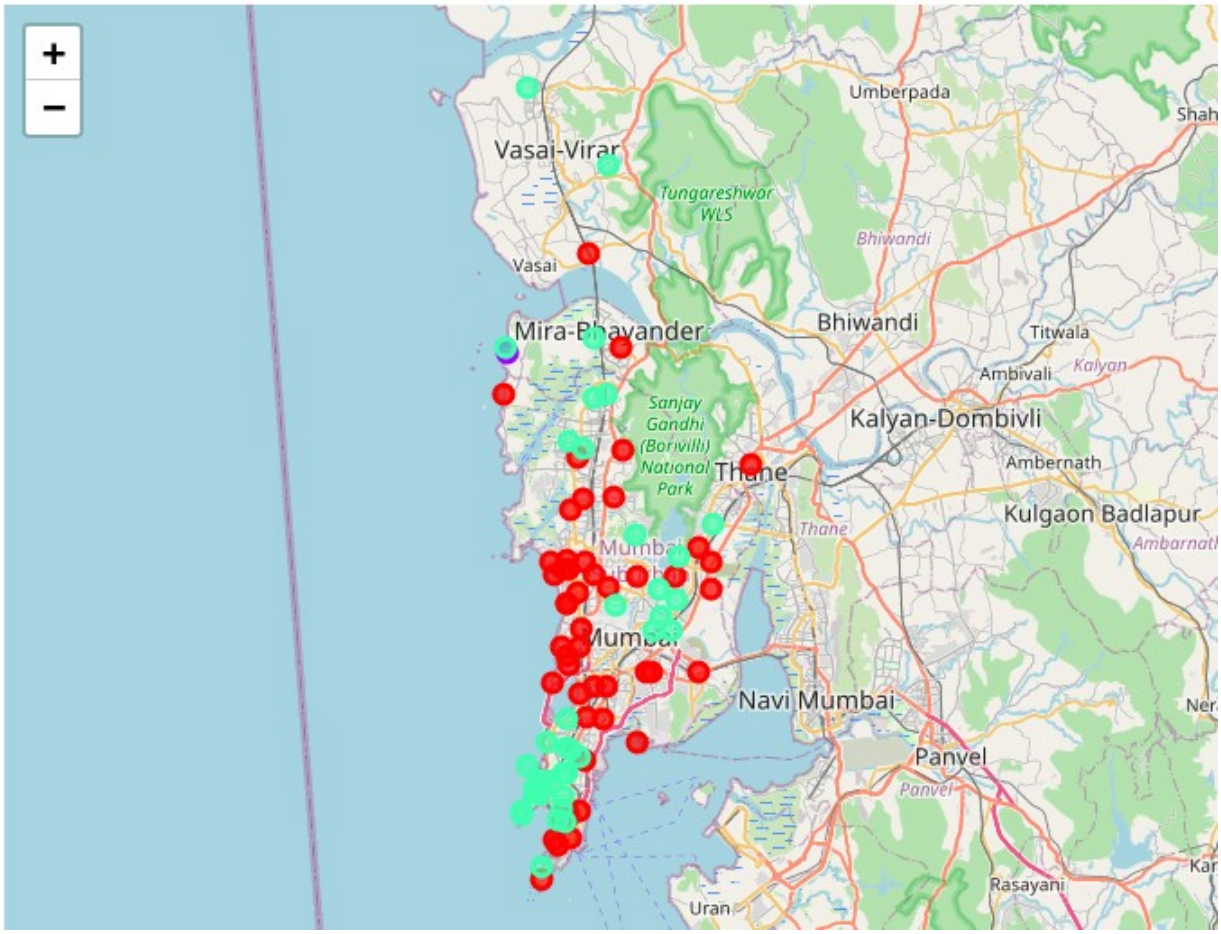
Next,we will use Foursquare API to get the top 100 venues that are within a radious of 2000 meters .We need to resister a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key.We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a pyhon loop.Foursquare will return the venue data in JSOn format and we will extract the venue name,venue

category,venue latitudeand longitude.With the data we can check how many venues are returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues.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 coing so,we are also preparing the data for use in clustering.Since we are analysing ‘Restaurant’ data,we 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 3 clusters based on their frequency of occurrence for “Restaurants”. 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 neighbourhoods, 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 neighbourhoods into 3 clusters based on the frequency of occurrence for “Restaurant”: • Cluster 0: Neighbourhoods with low number of Restaurant • Cluster 1: Neighbourhoods with high concentration of Restaurant • Cluster 2: Neighbourhoods with moderate number of Restaurant .The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion:

As observations noted from the map in the Results section, most of the Restaurant are concentrated in the north area of Mumbai city, with the highest number in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has very low number to no restaurant in the neighbourhoods. This represents a great opportunity and high potential areas to open new Restaurant as there is very little to no competition from existing malls. Meanwhile, Restaurant in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of Restaurant . Therefore, this project recommends property developers to capitalize on these findings to open new Restaurant in neighbourhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new Restaurant in neighbourhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 1 which already have high concentration of Restaurant and suffering from intense competition.

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 more 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. 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 0 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 over crowded areas in their decisions to open a new Restaurant.

References:

Neighbourhood data is retrieved from https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai

Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs

India consumer food service growth data(www.aaronallen.com)