Coursera Capstone Project

IBM Applied Data Science Capstone

Opening a New 'Café' in Mumbai

Introduction

Cafes are one of the most important aspects of humans. Be it a date, a meeting, if you want to enjoy with friends, if you want to eat some snacks, these are the most favourite places of people. As a result, there are many cafes in the city of Mumbai and many more are required to build considering increasing popularity. Even though these businesses require less capital and space, they can fetch large profits. But what matters is the location. The location plays a vital role in deciding whether this business will fail or not.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Mumbai to open a new cafe. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question.

Target Audience of this project

This project is particularly useful to small businesses and individuals looking to open or invest in new cafes in Mumbai. Since, food industry is one of the industry which is everlasting and always in demand, many people are interested in either investing in such business or actually put their money and do this business.

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 Mumbai, the financial capital city of the country of 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 cafes. We will use this data to perform clustering on the neighbourhoods

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_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 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. 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.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Mumbai. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_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 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. 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. 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 “cafes” data, we will filter the “cafes” as venue category for the neighbourhoods.

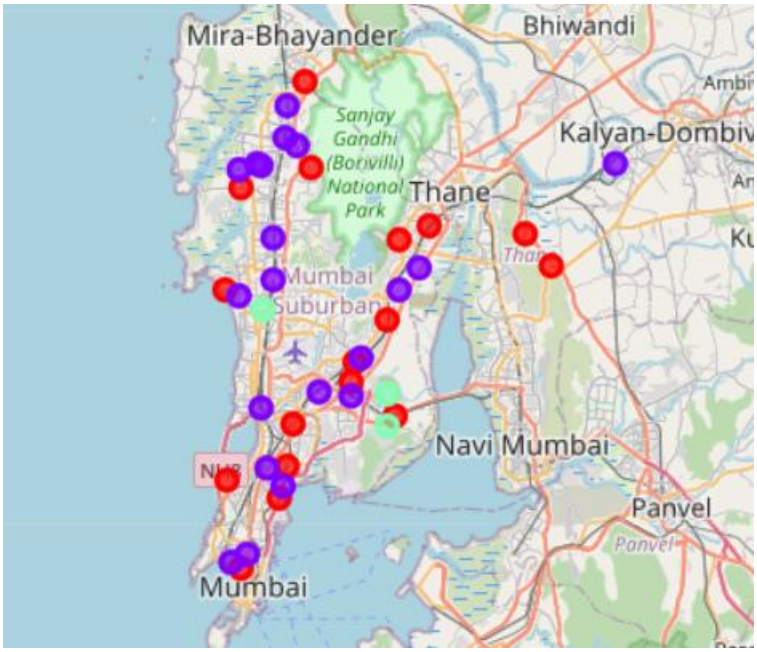
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 “cafes”. The results will allow us to identify which neighbourhoods have higher concentration of cafes while which neighbourhoods have fewer number of “cafes”. Based on the occurrence of cafes in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new cafes.

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 “Cafes”:

- Cluster 0: Neighbourhoods with moderate number of cafes
- Cluster 1: Neighbourhoods with highest number of cafes
- Cluster 2: Neighbourhoods with low or close to zero number of cafes

The results of the clustering are visualized in the map below with cluster 0 in r purple ed colour, cluster 1 in red colour, and cluster 2 in mint green colour.



Discussion

As we can see, most of the cafes are concentrated in cluster 1 with highest number. Cluster 0 comprises of moderate number of cafes. Cluster 2, on the other hand, has very low number of or no cafes at all. This represents a great opportunity and high potential areas to open new cafes as there is very little to no competition from existing cafes. Also, with unique ideas or strategies, there is a chance of growth in cluster 0 places, but the new cafes will definitely face moderate competition from other cafes. Lastly, investors and new businesses are advised to avoid neighbourhoods in cluster 1 which already have high concentration of shopping malls 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 cafes, there are other factors such as population and income of residents that could influence the location decision of a new cafes. 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 cafe. 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 shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to open a new cafe. 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 cafe.