BATTLE OF NEIGHBOURHOODS

CAPSTONE PROJECT

IBM APPLIED DATA SCIENCE CAPSTONE

Opening an Indian Restaurant in the suburbs near Mumbai, India

Pradnyesh Choudhari

April 2020

Introduction

For many people, these days dining in restaurants has become very common. They can enjoy themselves and the trends have completely changed, the restaurants get crowded not only during holidays or weekends but remain crowded over the complete year. For Indians, it is very common to prefer Indian Cuisine than any other more occasionally. Suburbs near Mumbai include all those places which surround Mumbai which has undergone a lot of development and is becoming center of attraction for many brands and major developers worldwide. These places are a big and a never-ending market for all those who are interested in providing services for the same. For opening a new restaurant considering a location that has less competition and for the developers, they have to consider even the other factors such as rent and all. Particularly, the location of the restaurant is a major reason for the success and failure of the project.

Business Statement

The objective of the capstone project is to analyze the neighborhoods using the techniques of data science and a few machine learning techniques such as clustering, to select the best location for opening an Indian restaurant in the suburbs surrounding to Mumbai, India.

If a restaurant chain wants to open a new restaurant in the suburbs near Mumbai where should you open it?

Target Audience

The target audience for this project can be major food chains or the investors and developers who want to invest their money in the restaurant business in Mumbai, India. It can also help the developers in the municipal corporation of plan the city and bound certain spaces to be only invested in this sort. Indian restaurant business is never ending and in a city like Mumbai where the population is becoming denser and denser and the lifestyle changing day by day this analysis may be helpful for all those who are interested in the restaurant business and this can also be applied for the all other accepts of development.

Data

- List of the Suburbs near Mumbai, India. The list is the most important as it defines the scope of the project and the area to which it is confined.

- Latitude and Longitude coordinates of these cities. These are required to plot the map and using the Foursquare API to get the data for the different surrounding places and the data related to the venue.

The Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai) contains the data of a complete list of suburbs near Mumbai. We scrap the data from the Wikipedia page where using the **beautifulsoup** library package of python. The **python geocoder** package gives the latitude and longitude data of these places. Further using the **Foursquare API** we get the venue data of the surrounding. **Folium** library is used to plot the map using the location data and finally after data cleaning using the machine learning algorithm **k-means clustering** we segment the data and cluster it into groups. After the clusters are formed, we get a group of places where we can open the restaurant and the problem statement with which we started this project gets fulfilled.

Methodology

The Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai) contains the data of a complete list of suburbs near Mumbai. By using the python package beautifulsoup we scrap the data from the Wikipedia page but just a list of the names of suburbs is not sufficient to get the desired results. So we further use the python geocoder package gives the latitude and longitude data of these places and then by using pandas and numpy libraries we create a data frame and add the longitude and latitude data into the data frame.

Then by using the folium library of python is used to plot the map of the Mumbai, India and further mark these suburbs on it. Foursquare API is used to get the data of the venues nearby these suburbs within a radius of 3000 m. We get the data from Foursquare API but in JSON format which we have to extract the venue name, venue longitude and latitude. This all data is further added into the data frame which we had created earlier by combining the data based on the location data. Then we analyse each neighborhood and taking the mean of the frequency of the occurrence of each venue category.

By doing this we create the data for apply the k-means clustering, a ML algorithm to segment the data into different clusters. We further cluster the data into 3 clusters based on the “Indian Restaurant” data available with us. The results allow us to get the data to identify which neighborhoods have higher concentration of Indian Restaurants and which neighborhoods don’t have much concentration of Indian Restaurants. Based on the these results, we can answer the problem statement that

we created earlier and find the suitable location for opening a new restaurant which attract more crowd.

Results

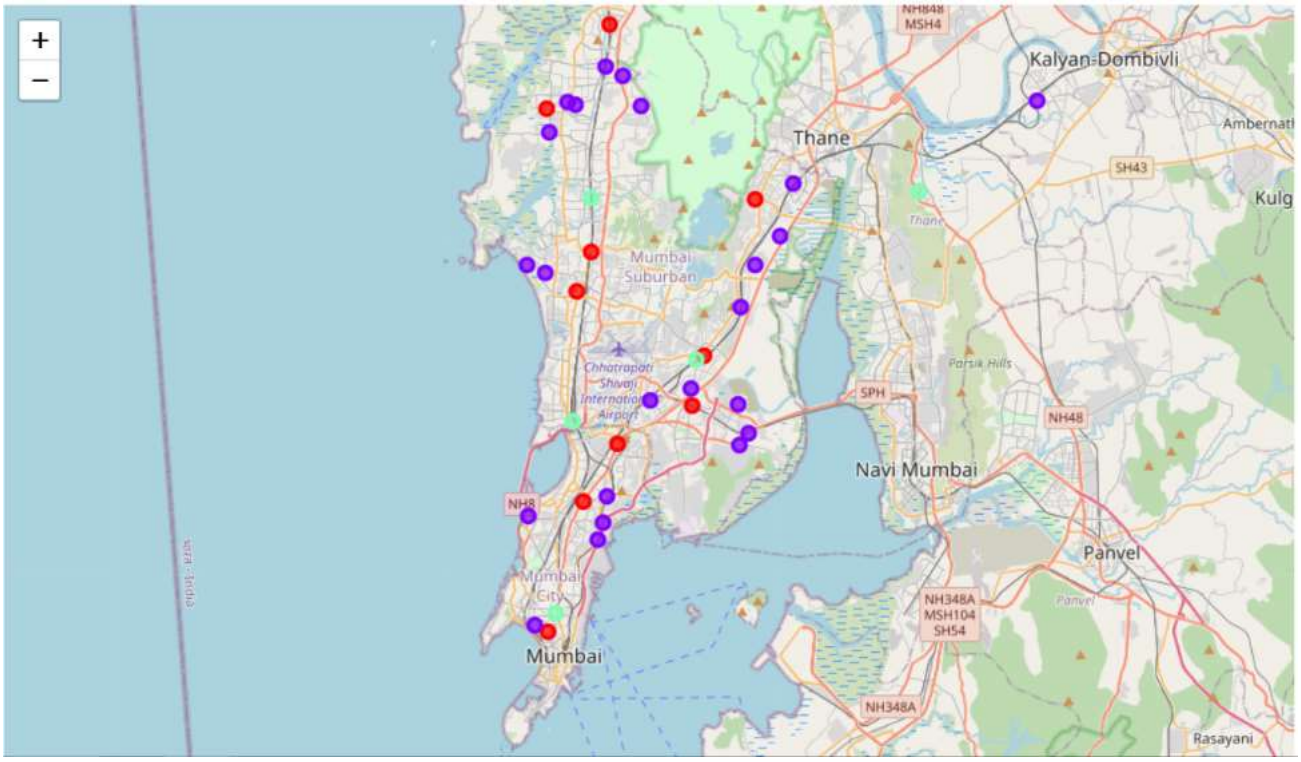
After applying the k-means cluster we obtain a result which we have again plotted on a map using the folium library.

We have clustered the data for Indian Restaurants in 3 clusters

Cluster 1: Red Colour: Neighborhood with moderate number of Indian Restaurants.

Cluster 2: Purple Colour: Neighborhood with low number of Indian Restaurants.

Cluster 3: Mint Green Colour: Neighborhood with high number of Indian Restaurants.



Inference

As the map in the result section suggests and from the observations there are very less places across the suburbs which have very high number of Indian Restaurant and most of the places have moderate number of restaurants from which we can easily find that there is a lot of scope in the business for opening restaurant. The cluster 2 gives the list of all those places where one can think about opening a new restaurant and those places which fall in the cluster 3 should be avoided while choosing a location for the new restaurant. If these results are taken into consideration, cluster 1 and 2 are good options for opening a new restaurant and cluster 3 are not.

Limitations and Further Scope

Due to lack of data many suburbs were not even considered during this project. The other major problem is that we have only considered the number of restaurants around that place and which is only a single factor while considering the facts. There are a lot of other factors such as population density, economic conditions of the location. However, due to lack of the data for all such factors could not be considered while applying the ML model. The API used in the project is a basic and free version of the Foursquare API and there were restrictions to the data and number of calls that we make on daily basis. For the further research if this topic is chosen then all these factors should be taken into consideration and the paid version of the Foursquare API can be used.

References

1. Wikipedia Page https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai
2. Foursquare API https://foursquare.com/developers/apps