IBM APPLIED DATA SCIENCE CAPSTONE

APPROPRIATE LOCATIONS TO OPEN A NEW SHOPPING MALL IN MUMBAI, INDIA

INTRODUCTION

Shopping malls are a one-stop destination, where shoppers can do various activities, ranging from shopping, eating, gaming and watching movies. It is a great place to visit, and is always buzzing, especially during holidays. For, property developers, it is a great way to make profit out of catering to the demands of the public. However, a lot of thought and consideration goes into opening a shopping mall. One of the most important factors to be considered, is the location.

BUSSINESS PROBLEM

The aim of this project is to help property developers in choosing the ideal location for opening a shopping mall, in Mumbai, India, using data science methodology and machine learning techniques like clustering.

TARGET AUDIENCE

This project aims to help property developers in opening new shopping malls around Mumbai, India. It will help them choose ideal locations minimizing competition and maximizing profit.

DATA

- List of neighborhoods in Mumbai, India
- Latitude and longitude coordinates of those neighborhoods, in order to plot the map
- Data related to shopping malls, in order to perform clustering on the neighborhoods

First, we extract the neighborhoods in Mumbai, using web scraping, from the page: https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai .

Then, we use the Geocoder library to extract the coordinates of each neighborhood.

Then, we use the Foursquare API to get the venue data for each of the neighborhoods. It provides us with a lot of venues, but we are interested in Shopping malls only, to solve our problem.

We also make use of machine learning techniques, such as K means clustering and map visualization using Folium.

METHODOLOGY

First, we extract the neighborhoods in Mumbai, using web scraping with the help of Python requests and beautifulsoup packages, from the page: https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai .

Next, we need to get the geographical coordinates of each of these neignborhoods, in order to be able to use it in Foursquare API. We do this with the help of Geocoder library. We then convert all of the data we have into a pandas dataframe, and visualize the data using a map with the help of Folium.

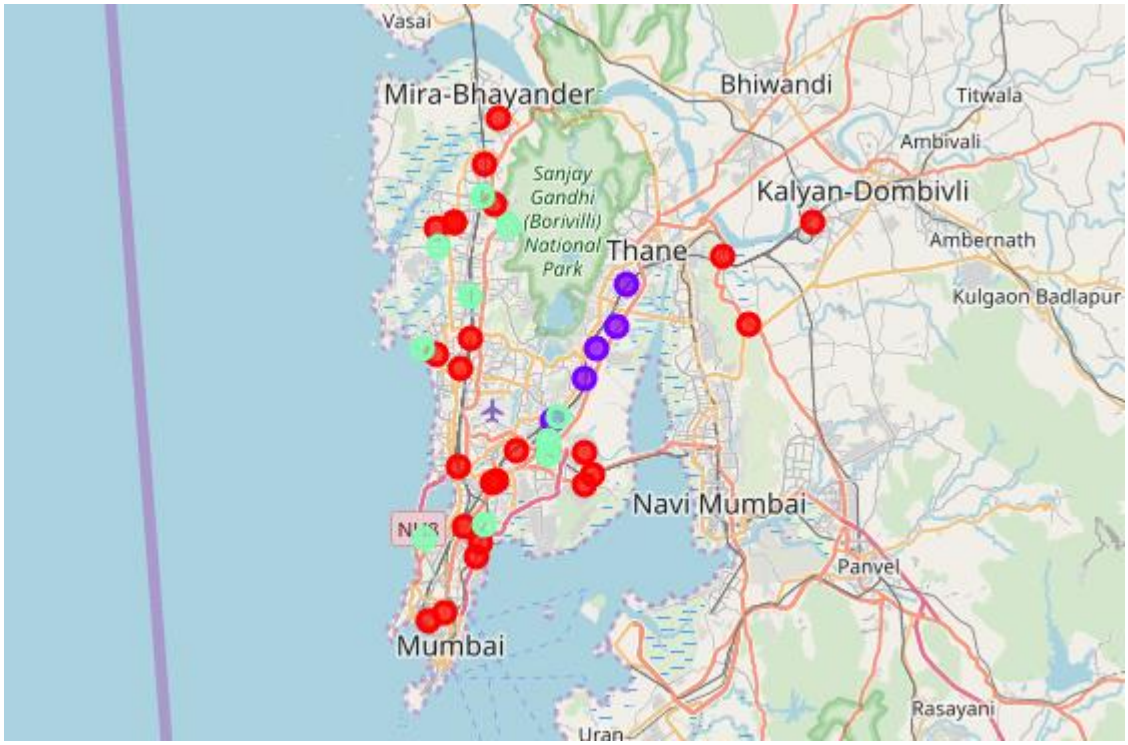
We then use Foursquare API to find the top 100 venues within a radius of 2 kms. We need to provide the Foursquare ID and Foursquare secret key. We then execute a loop to make API calls to Foursquare passing in the coordinates of the neignborhoods. Foursquare returns the data in JSON format, and we extract the venue name, venue category, venue latitude and venue longitude. We, then, analyse each neighborhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. Since, our business problem is relevant to only shopping malls, we filter “Shopping mall” as venue category for the neighborhoods.

We, then perform k-means clustering. It identifies k centroids and allocates every data point to the nearest centroid, thereby forming a cluster. We cluster the neighborhoods into 3 clusters based on their frequency of occurrence of shopping malls. This helps us identify which clusters have low, moderate, and high number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, we can identify which neighborhoods are most suitable to open shopping malls.

RESULTS

The results from the K-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence of “Shopping mall”:

Cluster 0 (in red) has no shopping malls.
Cluster 1 (in purple) has the most number of shopping malls.
Cluster 2 (in cyan) has moderate number of shopping malls.



DISCUSSION

Cluster 0 (in red) has no shopping malls. Cluster 1 (in purple) has the most number of shopping malls. Cluster 2 (in cyan) has moderate number of shopping malls.

Therefore, it is advised to open new shopping malls in neighborhoods belonging to cluster 0, where there is no competition. This will also provide the people living in these neighbourhoods with a means of livelihood.

Property developers can consider opening a shopping mall in neighborhoods belonging to cluster 2 if they believe they can stand out from the other malls in the locality and can fight the moderate competition.

Property developers are advised to avoid neighborhoods in cluster 1 which already have high concentration of shopping malls and are suffering from intense competition.

CONCLUSION

We have provided a solution to the business problem by using data science methodology and machine learning techniques. We extracted the data from a webpage and prepared it. We performed machine learning by clustering the data. And we helped relevant stakeholders identify what the best location is to open a shopping mall. From the findings of this project, we advise property developers to open shopping malls in neighborhoods belonging to cluster 0.