IBM Data Science Capstone

Identification of Prospective Location for New Shopping Mall in Bengluru, India

By:

Deepro Sengupta

Table of Content

Contents

Introduction

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can shop for everything ranging from groceries, clothes; electronics as well as Fashion related articles such as accessories, etc. Furthermore, they also provide other entertainment options like fast food, fine dining, movie theatres etic all under one roof. For retailers, the large crowd that it attracts enables them to sell their products and realize their investments quickly as well as have a very less consumer acquisition cost. Additionally, shopping malls also provide a great investment opportunity for property developers as well. They can choose to sell each shops within the Mall or rent it out. Therefore, opening a shopping mall is a good business decision. However, like any business decision, it should be supported with serious consideration. Particularly, location of a mall can be the difference between its success or failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Bengaluru, India for new shopping mall. Using data science methodology and machine learning algorithm like clustering, this project aims to provide solutions to answer the business question: In Bengaluru, what should be the location for a new shopping mall?

Target Audience

This project is particularly useful to property developers and investors looking to open or invest in a new shopping mall in the city of Bengaluru in Karnataka, India. This project, however, can be adapted for other purposes as well with minor code modification. This is due to the fact that this project already collects massive amount of data related to the top 100 local venues in Bengaluru before directing our attention towards shopping malls in these neighbourhoods. Therefore, if one wanted to identify a location to open a new food joint, for example, it can be done just by a few minor changes on the code.

Data

In order to work on our problem, we will need the following data:

* Names of neighbourhoods in Bengaluru.

* The geographical co-ordinates of each neighbourhood.

* All shopping malls in Bengaluru

In order to acquire these data, we have used the following sources:

* Foursquare API

* Google Maps Reverse Geocoding

Data Source, extraction and usage

We used the Python Pandas library's pandas.read_html to extract all tables containing a list of all neighbourhoods in Bengaluru, India. Then we used the Python Geoencoder package to get the latitude and longitudinal coordinates of the neighbourhoods.

Following this, we will use Foursquare API to get the data for the top 100 venues in those neighbourhoods. In order to do so, we will use the latitudinal and longitudinal data which we have acquired using the Geocoder package and use the Request library to long with our Foursquare API credentials to get the necessary data as a JSON file. Foursquare has one of the largest database of 105+ million places and is used by over 1,25,000 developers. Although Foursquare API provides many categories of venue data, we will direct our attention to the Shopping Mall category in order to help us solve the aforementioned business problem.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Bengaluru, Karnatake. Fortunately, the list is available in the Wikipedia page. We will acquire the list of neighbourhoods names using Pandas read_html library and then get the latitude and longitude. 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 Bengaluru.

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 "Shopping Mall" data, we will filter the "Shopping Mall" as venue category for the neighbourhoods.

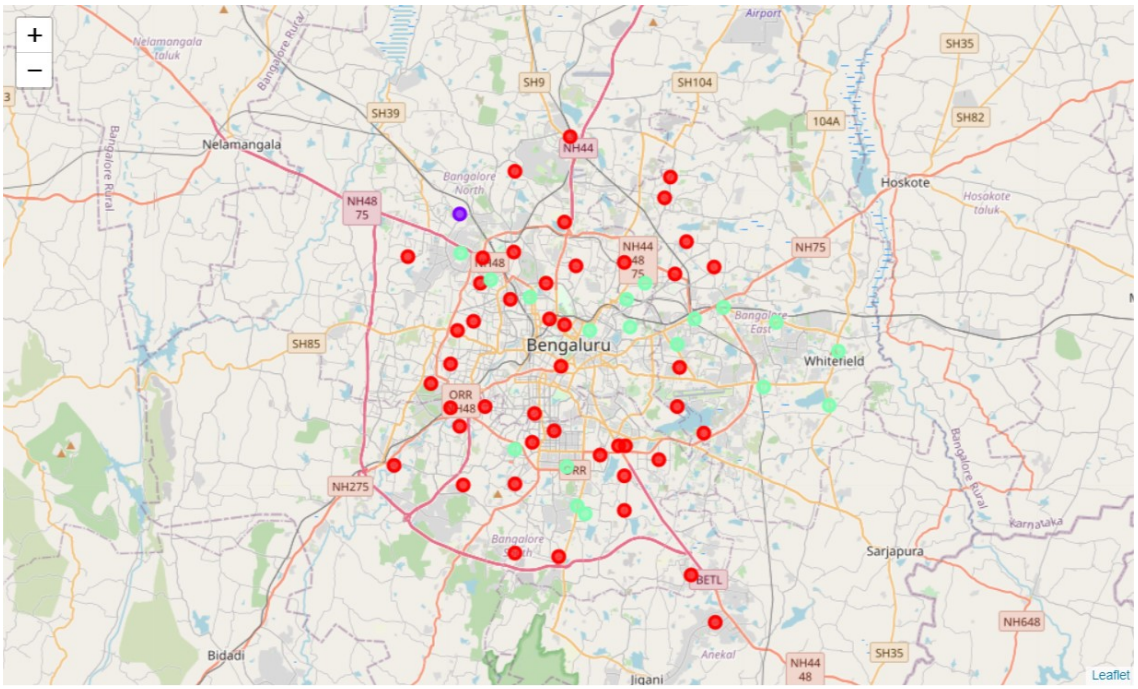
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 "Shopping Mall". 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 “Shopping Mall”:

- Cluster 0: Neighbourhoods with moderate number of shopping malls
- Cluster 1: Neighbourhoods with low number to no existence of shopping malls
- Cluster 2: Neighbourhoods with high concentration of shopping malls

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 shopping malls are concentrated in the central area of Bengaluru city, with the highest number in cluster 0 and moderate number in cluster 2. On the other hand, cluster 1 has just one shopping mall in the neighbourhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighbourhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 0 which already have high concentration of shopping malls 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 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 1 are the most preferred locations to open a new shopping mall. 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 shopping mall.