Coursera Capstone Project

IBM Applied Data Science Capstone

Identify new place for a shopping mall in Bangalore, India

By : *Ravi Teja Jalagam*

April 2020

Introduction :

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of Bangalore. Opening shopping malls allows property developers to earn consistent rental income. Ofcourse, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping malls is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem :

The objective of this capstone project is to analyse and select the best locations in the city of Bangalore, India to open a new shopping mall. Using Data Science methodology and Machine Learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Bangalore, India if a property developers is looking to open a new shopping mall, where would you recommend that they open it?

Target Audience of the project :

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in the capital city of Karnataka i.e Bangalore. This project timely helps as the city is currently suffering from oversupply of shopping malls.

Data :

To solve the problem, we need the following data

1. List of Neighbourhoods in Bangalore city.
2. Latitude and Longitude of those Neighbourhoods. This is required to plot the map and also to get the venue data.
3. Venue data, particularly data related to shopping malls. 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:Neighbourhoods_in_Bangalore contains a list of neighbourhoods in Bangalore, with a total of 67 neighbourhoods. We have already gathered data of Neighbourhoods and their geographical coordinates in a csv file. We will be creating a dataframe from a csv file.

After that, we will use FourSquare API to get venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping mall category in order to help us to solve the business problem put forward. 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 Bangalore. Fortunately, the list is available in the Wikipedia page
(https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore) 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 addresses 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 visualise the neighbourhoods in a map using the Folium package. This allows us to perform a sanity check to make sure that geographical coordinates data returned by Geocoder are correctly plotted in the city of Bangalore.

Next, we will use Foursquare API to get theto p100 venues that are within radius of 2000 Meters.
We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare Authentication 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 venue 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 the use in clustering. Since we are analyzing the "Shopping Mall" data, we will filter the "Shopping Mall" as a venue category for the neighbourhoods.

Lastly, we will perform clustering in 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 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":

1. Cluster 0: Neighbourhoods with high concentration of shopping malls
2. Cluster 1: Neighbourhoods with moderate number of shopping malls
3. Cluster 2: Neighbourhoods with low number or no existence shopping malls

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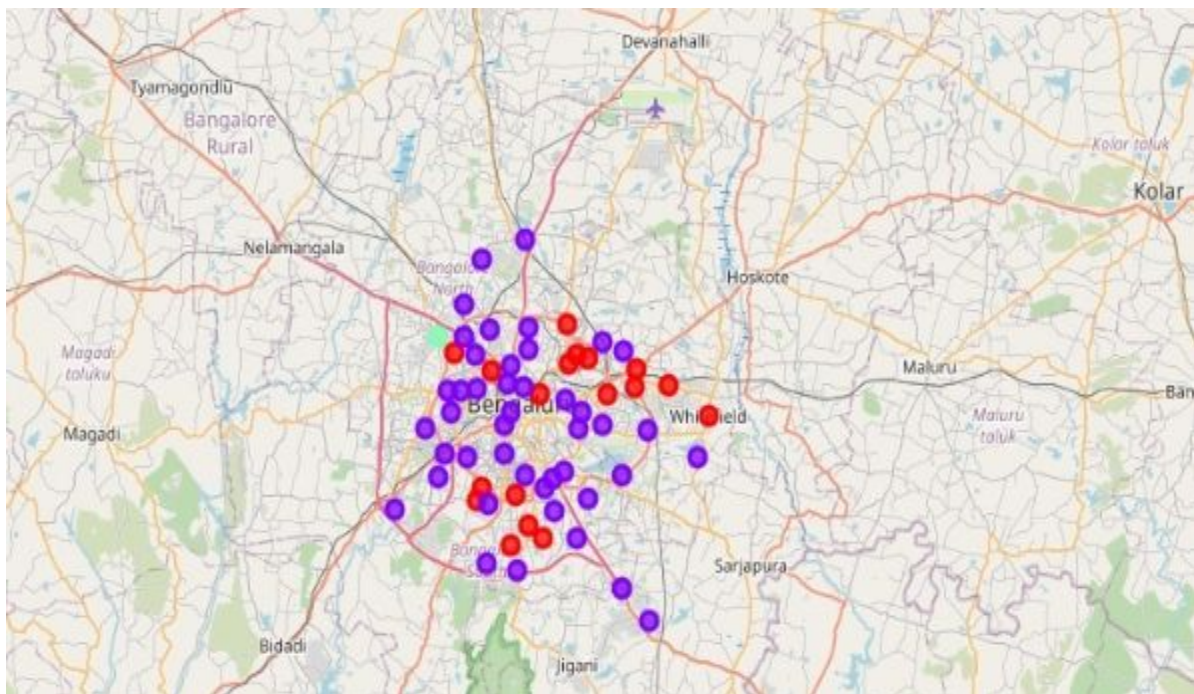
Discussion

As observations noted from the map in the Results section, Most of the shopping malls are concentrated in the Eastern part of Bangalore city, with the highest number in cluster 0 and moderate number in cluster 1. On the other hand, cluster 2 has a very low number of totally no shopping malls in the neighborhoods. 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, this also shows that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still having very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 2 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 0 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 a factor i.e frequency of occurrence of shopping malls in Neighbourhoods. But there are other factors such as population and income of residents that could influence the location decision of a shopping mall. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future projects 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.

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 neighbourhood in cluster 0 are 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 decision to open a new shopping mall.

References

Category: Neighbourhoods in Bangalore. Wikipedia. Source
https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore

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