Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Coffee Shop in Bangalore, India

By: Albert Abraham

October 2019

Contents

Introduction

For many people, a coffee shop is a great place to relax and enjoy themselves, hangout with friends, have casual meetings and/or go for a date. Property developers are also taking advantage of this trend to building more coffee shops to cater to the demand. As a result, there are many coffee shops in the city of Bangalore and many more are being built. Opening coffee shops allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new coffee shop requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the coffee shop is one of the most important decisions that will determine whether the shop 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 coffee shop. 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 developer is looking to open a new coffee shop, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new coffee shops in the silicon valley of India, Bangalore. This project is timely as the city is currently suffering from oversupply of coffee shops.

Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Bangalore.
- 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 coffee shops. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

Government of India provides datasets that are publically available at : data.gov.in.
There is a dataset containing All India Pincode directory with contact details along with Latitude and longitude.
We will filter the dataset to retrieve only data belonging to Bangalore. Latitude & Longitude columns are mostly missing values, so we will compute these using geocoder based on pincode of the location.
After that, we will use the Foursquare API to get the 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 Coffee Shop 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: data cleaning, data wrangling, working with API (Foursquare), 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, there is All India Pincode directory with contact details along with Latitude and longitude. We will wrangle the data to get data corresponding to Bangalore with only required variables. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use the Foursquare API. To do so, we will use custom geocoder code, 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 Bangalore.

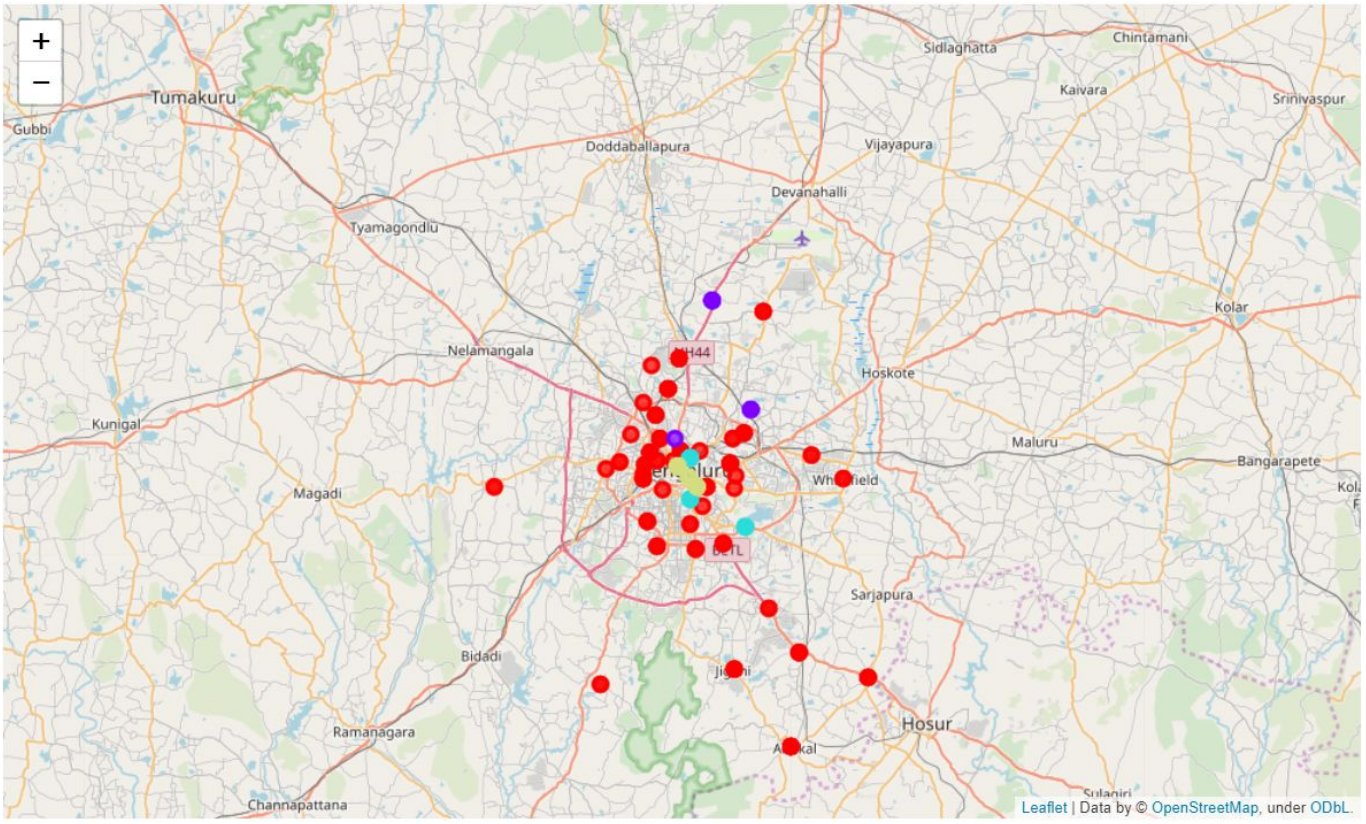
Next, we will use the Foursquare API to get the top 100 venues that are within a radius of 500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the 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 “Coffee Shops” data, we will filter the “Coffee Shop” 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 most popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 4 clusters based on their frequency of occurrence for “Coffee Shop”. The results will allow us to identify which neighbourhoods have higher concentration of coffee shops while which neighbourhoods have fewer number of coffee shops. Based on the occurrence of coffee shops in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new coffee shops.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 4 clusters based on the frequency of occurrence for “Coffee Shop”:
Cluster 1: Neighbourhoods with no coffee shops
Cluster 2: Neighbourhoods with high concentration of coffee shops
Cluster 3: Neighbourhoods with moderate concentration of coffee shops
Cluster 4: Neighbourhoods with low concentration of coffee shops

The results of the clustering are visualized in the map below with cluster 1 in red colour, cluster 2 in purple colour, cluster 3 in cyan colour, and cluster 4 in yellow colour.



Discussion

As observations noted from the map in the Results section, Most of the coffee shops are concentrated, with the highest number in cluster 2 and moderate number in cluster 3. On the other hand, cluster 4 has very low number of coffee shops and cluster 1 has totally no coffee shop in the neighborhoods. This represents a great opportunity and high potential areas to open new coffee shops as there is very little to no competition from existing coffee shops. Meanwhile, coffee shops in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of coffee shops. From another perspective, this also shows that the oversupply of coffee shops mostly happened in the central area of the city, with the suburb area still have very few coffee shops. Therefore, this project recommends property developers to

capitalize on these findings to open new coffee shops in neighborhoods in cluster 1 and 4 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new coffee shops in neighborhoods in cluster 3 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of coffee shops 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 coffee shops, there are other factors such as population and income of residents that could influence the location decision of a new coffee shop. 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 coffee shop. 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 4 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 coffee shop. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 and 4 are the most preferred locations to open a new coffee shop. 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 coffee shop.

References

- All India Pincode directory with contact details along with Latitude and longitude. Retrieved from https://data.gov.in/resources/all-india-pincode-directory-contact-details-along-latitude-and-longitude
- Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs