Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Ice-Cream shop in Bangalore, India

By: Nikhil Agarwal

April 9, 2020

Introduction

Ice-Cream has always been favourite choice for everyone when it comes to food during entertainment, trips or on day outings. And so do business, franchises and eateries shops. Now a days there are ice-cream franchises, local shops made on a big scale. As the trend of fancy shops have increased so has the investments in it and with big investments come big risks. So, before opening a shop one has to be aware about its surrounding like competition it is going to get. So, Location becomes very important for any business. Because audience is everything.

Business Problem

The objective of this capstone project is to analyse and select all neighbourhood areas in the city of Bangalore, India to open a new Ice-Cream parlour. 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 ice-cream parlour, 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 ice-cream parlour in the capital city of Bangalore, India. Bangalore is huge IT hub and people here lives for the weekend and lot of people here loves to spend life eating and celebrating. Now, to make more profit we need to find an area which is famous or good but has less number of ice-cream parlour so that we can get minimum competition. So, before an investment results of this project can be really helpful.

Data

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

- • List of neighbourhoods in Bangalore: This defines the scope of this project which is confined to the city of Bangalore, the metropolitan city of the country of India in South Asia.
- • 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 existing Ice-cream shops. 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/List_of_neighbourhoods_in_Bangalore) contains a list of neighbourhoods in Kuala Lumpur, with a total of 65 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and pandas packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database 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

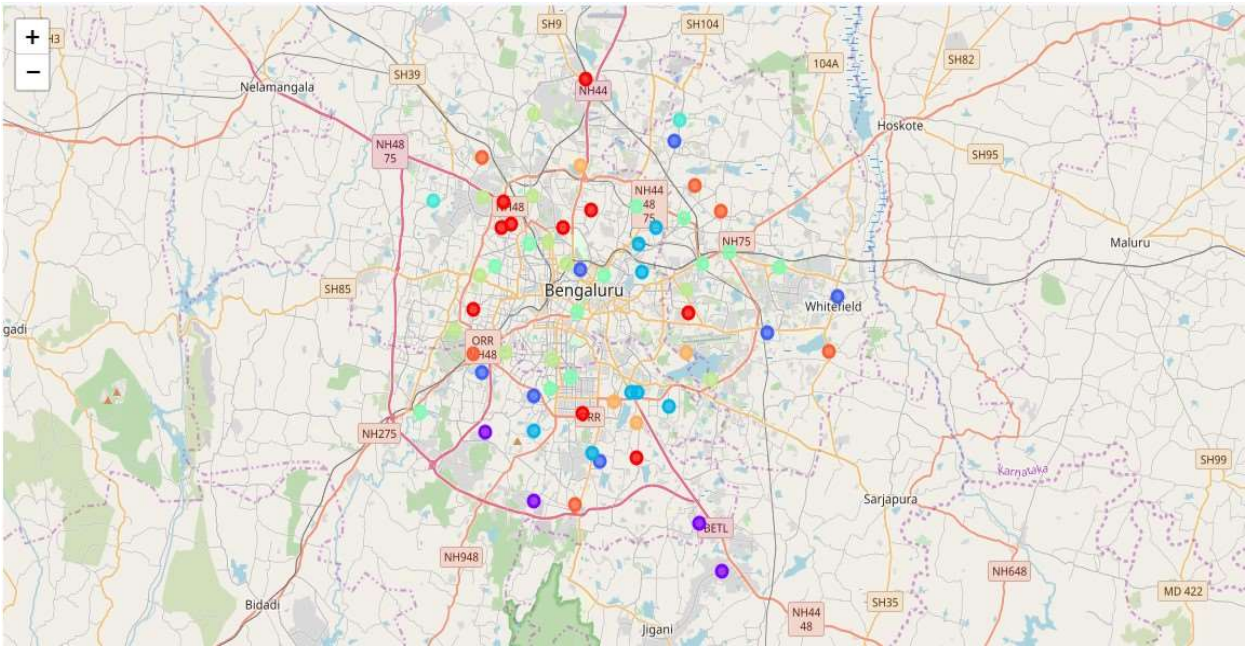
- Firstly, we need to get the list of neighbourhoods in the city of Bangalore. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore). We will do web scraping using Python requests and pandas packages to extract the list of neighbourhoods 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 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 Foursquare API to get the top 300 venues that are within a radius of 5000 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 “Ice Cream Shop” data, we will filter the “Ice Cream 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 popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 9 clusters based on their frequency of occurrence for “Ice Cream shop”. The results will allow us to identify which neighbourhoods have higher concentration of Ice Cream Shops while which neighbourhoods have fewer number of Ice Cream Shops. Based on the occurrence of Ice Cream Shops in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open Ice Cream Shop.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 9 clusters based on the frequency of occurrence for “Ice Cream Shops”:

- Cluster 2 and 5 is area with **least** number of ice cream shops,
- Cluster 3, 6, 9 has **moderate** number of ice-cream shops, and
- Cluster 1,4,7,8 has **maximum** number of ice-cream shops

So, it is more likely that one should choose 1 out of 7 neighbourhood areas among cluster 2 and 5 for opening a new ice-cream shop.



The 7 Neighbourhoods of cluster 2 and 5 are:

- Uttarahalli
- Electronic City
- Anjanapura
- Bommasandra
- Cantonment area
- Vijayanagar
- Indiranagar

Discussion

As observations noted from the map in the Results section, this project has been successful in dividing Bangalore into 9 cluster of areas.

Based on frequency of ice creams shops we have determined 4 clusters (1, 4, 7, 8) to have adequate number of shops and opening new shop in these areas would not be a very good idea.

3 clusters (3, 6, 9) have moderate number of shops and one should consider these areas for business only when it offers some really good options (land, cheaper, crowdie).

2 clusters i.e. cluster 2 and 5 has very less number of shops and it contains in total 7 areas. So, these areas would be better for setup of a new shop.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of ice-cream shops, there are other factors such as population ,shopping mall and income of residents that could influence the location decision of a new ice-cream shops. 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 ice-cream 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 9 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 7 neighbourhoods in cluster 2, 5 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.

References

Category:

List of Neighbourhoods in Bangalore. *Wikipedia*. Retrieved from

https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore

Foursquare Developers Documentation. *Foursquare*. Retrieved from

https://developer.foursquare.com/docs