Capstone Project

IBM Applied Data Science

By – Abhishek Mishra

August, 2020

BUSINESS PROBLEM

- Opening a new Indian restaurant in the city of Bangalore, India.

INTRODUCTION

Bangalore is the IT city of India, there is a major population of youngsters working in thousands of IT firms coming from all over the country. Thus, there is a huge market of restaurants and fast food joints as the young population does not believe much in cooking food. After work a lot of people take dinner at restaurants and then go back home. This induced the opening of new restaurants in Bangalore. A restaurant provides a huge earning if it runs well, also it's success or failure depends upon a lot of factors from recruiting a good management team and chefs to picking up a location for opening a new restaurant. Although the location one of the most important factor among the others.

BUSINESS PROBLEM

The objective of this project is to analyze and select the best location in Bangalore, India to open a new Indian restaurant.

Using data science methodology and unsupervised machine learning techniques like clustering, this aims to provide the best solution to answer the business problem: In the city of Bangalore, India, if an investor is willing to open a new Indian restaurant, where you would recommend that they open it?!

DATA DESCRIPTION

To solve this problem we need the following data

- List of post offices and their pin codes in Bangalore, IN
- Location coordinates of these postal areas
- Venue data, we will use this data to perform the clustering technique.

SOURCES OF DATA / DATA EXTRACTION

This page "https://finkode.com/ka/bangalore.html" contains a list of post offices and their pin code in the city of Bangalore, IN. We will be using web scraping techniques to extract the data from the html code using python requests and Beautiful soup package. Then we will get the coordinates of these locations using the python geocoder package, which will give us the latitude and longitude values of each post office location.

After that, we will use the Foursquare API to access the venue data for these post office locations. Foursquare will provide many categories of venue data, we are particularly interested in the restaurant venues to proceed further in our analysis.

Methodology

Firstly, we need to get the list of areas in the city of Bangalore, IN. Generally, post office code is a way to analyze the areas of a city. On searching online we found a page "https://finkode.com/ka/bangalore.html" where we got our data very efficiently and in a proper way.

We have to perform web scraping using python requests and beautiful soup packages to extract the list of areas and their pin codes. However, this is just the list of names of the areas in the city of Bangalore, we will be needing the geospatial coordinates to proceed further in our analysis and to use Foursquare API calls. The data extracted is stored in a pandas data frame, to which the coordinates are to be added.

Next, to get the coordinates of these locations we used the Geocoder package available in python. Postal codes have been passed as arguments in the "geocoder.location" function to get their geospatial coordinates. This function has been used inside a loop to get the coordinates of all the locations. Upon fetching the geospatial coordinates we have to store this corresponding to their location in our data frame. After populating the data frame we mapped the location using folium package available in python for better visualization.

After this, venue data is required! For getting the venue data we have to use the Foursquare API to access the location data. For this we have to register on the Foursquare website and we have to define the credentials to make API calls which will enable us to access the venues for our analysis.

Lastly, we will perform the clustering technique on our prepared data set. We are using the k-means clustering algorithm this will cluster the areas depending upon the type of most occurring venues at that location. We will cluster the areas into 4 clusters and the cluster which have the least occurrence of Indian restaurants will be our target cluster.

Results

The results showed us the 4 clusters:

- Cluster 1 - The most common venue was ATM and restaurants are moderately common.
- Cluster 2 - The most common venue in this cluster is Indian restaurant.
- Cluster 3 - In this cluster the most common venues are fast food joints and there close to none Indian restaurants.
- Cluster 4 - The most common venue in this cluster is Indian restaurant.

Conclusion

Most of the Indian restaurant are concentrated in the cluster 2 and 4 in Bangalore city, with the highest number in cluster 4 and moderate number in cluster 2. On the other hand, cluster 3 has very low number of Indian restaurant in the area. This represents a great opportunity and high potential areas to open new Indian restaurants as there is very little competition from existing Indian restaurants, as there is a scarcity of Indian restaurants in this cluster only fast food joints are present. Meanwhile, Indian restaurants in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of restaurants. Therefore, this project recommends the investor to capitalize on these findings to open new Indian restaurants in neighborhoods in cluster 3 with little to no competition.