Predicting Hotels in India

Marina Sruthi M

1.Introduction

1.1 Background

In India, Hotels have caught a tsunami of middle class expansion over the past decade, as more Indians earn enough extra cash to develop a taste for tourism, and business travel has also increased. Entrepreneurs are searching for a best place to open hotels. We can also utilise software engineers to create apps and an operating system designed to overcome every hurdle associated with running thousands of hotels. They are used by customers, cleaners, auditors and reception and sales staff.

1.2 Problem

The idea behind this project is that hotels are always a bigger opportunity in india. It might present a great opportunity for an entrepreneur. The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new hotel in india. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In India, if an entrepreneur wants to open a hotel, where should they consider opening it?

2. Data Acquisition and Cleaning

2.1 Data Sources

The datasets are gathered by web scraping from here. The data thatvwe are looking for has the following details.

- List of cities in India.
- Latitude and Longitude of these cities.
- Venue data related to Hotels. This will help us find the cities that are most suitable to open a hotel
- Getting Latitude and Longitude data of these cities via Geocoder package
- Using Foursquare API to get venue data related to these cities

2.2 Data Cleaning

- Remove null values
- The data had postal column which is a string of city,state, country. We split them and store it in separate columns.

3. Methodology

The list of cities in India is gathered as said in data section. Inorder to achieve it a web scraping is done using request and beautiful soup apo can prettify it and also it is easier and more convenient to pull tabular data directly from a web page into dataframe. We can get the latitude and longitude from the geocoders api or the above web scraping provides all the details.

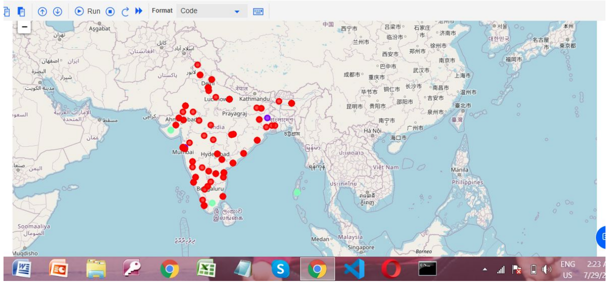
Once the latitude and longitude is obtained we perform data cleaning. Then we can visualized the map of India using Folium package to verify whether these are correct coordinates.

Next, we use the Foursquare API to pull the list of top 100 venues within 500 meters radius. In order to obtain account ID and API key to pull the data from foursquare a developer account is created. From Foursquare, we will be able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues.

Then, we analyze each cities by grouping the rows by cities and taking the mean on the frequency of occurrence of each venue category. This is to prepare for the clustering to be done later.

Lastly, we perform the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, 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 it is highly suited for this project as well. I have clustered the cities in india into 3 clusters based on their frequency of occurrence for “Hotel”. Based on the results (the concentration of clusters), we will be able to recommend the ideal location to open the hotel.

4. Visualization



The results from k-means clustering show that we can categorize Indian Cities into 3 clusters based on how many Thai restaurants are in each neighborhood:

- Cluster 0: Cities with high number of Hotels
- Cluster 1: Cities with no hotels
- Cluster 2: Cities with no or less number of hotels

The results are visualized in the above map with Cluster 0 in green color, Cluster 1 in red color and Cluster 2 in light purple color.

5. Prediction :

Most of Hotels are in Cluster 2 which is around Gujarat and Tamil Nadu and lowest (close to zero) in Cluster 0 areas which are west bengal and Maharashtra. Looking at nearby venues, it seems Cluster 1 might be a good location as there are not a lot of Hotels in these areas. Therefore, this project recommends the
entrepreneur to open an hotel in these locations with little to no competition.

6. 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 the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.