Opening A New Restaurant in Pune, India

Vatsal Unadkat

1. Introduction

1.1 Background

Being one of the IT hubs in India, Pune is one of the fastest growing cities in India with a growing population of 31.2 lakhs (2011). This has people from all over the country and abroad coming to Pune with a diverse food taste. This project focuses on Restaurant, food truck and franchise owners looking to open and/or expand their business. Of course, as with many business decisions, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the outlet is one of the most important decisions that will determine whether the business will be a success or a failure.

1.2 Problem

The objective of this capstone project is to analyze and choose the best locations in the city of Pune, India to open a restaurant. 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 Pune, if a business man is looking to open/expand a restaurant/franchise, where would you recommend that they open it?

1.3 Interest

Restaurant, food truck and franchise owners looking to expand their business would be very interested in accurate prediction on where they should open new outlets for competitive advantage, business value and to make the maximum profit.

2. Data acquisition and cleaning

2.1 Data sources

We will require a list of the areas in Pune. The coordinates of these areas (Latitude and Longitude). Finally, we will need the venue data related to all categories of restaurants. This is will useful in finding segments in the city which contains pertinent concentration of food outlet/restaurant types.

2.2 Data Extraction

The webpage of 'MakeMyTrip' alphabetically lists down all areas within Pune city. We will use web scraping techniques to extract the data from the webpage, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. 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 restaurants category 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 (MakeMyTrip webpage), 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.

3. Methodology

I started with how to get the list of areas/neighborhoods in Pune, India. This was made possible by extracting the list of areas from MakeMyTrip page ("https://www.makemytrip.com/hotels/hotels-pune-area-list.html").

I performed the web scraping by using the 'beautifulsoup' library in Python. The data was not in a tabular format; therefore, the extraction was done using a function which iterates over the area links.

The initial scraping only retrieves the list of areas in Pune. I then had to get their latitudes and longitudes by utilizing Foursquare API to pull the list of restaurants near these areas. When I tried to use the geopy package, it only worked intermittently which led me to compile a CSV file consisting of areas and their coordinates. After gathering all these coordinates, I visualized the map of Pune using the Folium package to check if the coordinates were correct or not.

I then used the Foursquare API to pull the list of top restaurants within 500 meters radius. I had to create a Foursquare developer account to obtain an account ID and API key to pull the data. From Foursquare database, I could now pull the names, categories, latitude, and longitude of the venues. With this data in hand, I could also check how many unique categories I could get from these venues. Then, I analyzed each area by grouping the rows by areas and taking the mean on the frequency of occurrence of each venue category. This was at preparation step for the clustering to be done later. I specifically used "restaurant" as a query to search in an area. Concentration of specific type of restaurant in a part of a city signifies that the restaurants/food outlets belonging to a category is congested in that region.

For my final step, I did clustering using k-means. 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 it is highly suited for this project as well. I have clustered the neighborhoods in Pune into 3 clusters based on their frequency of occurrence for categories of restaurants. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open a restaurant outlet.

4. Results

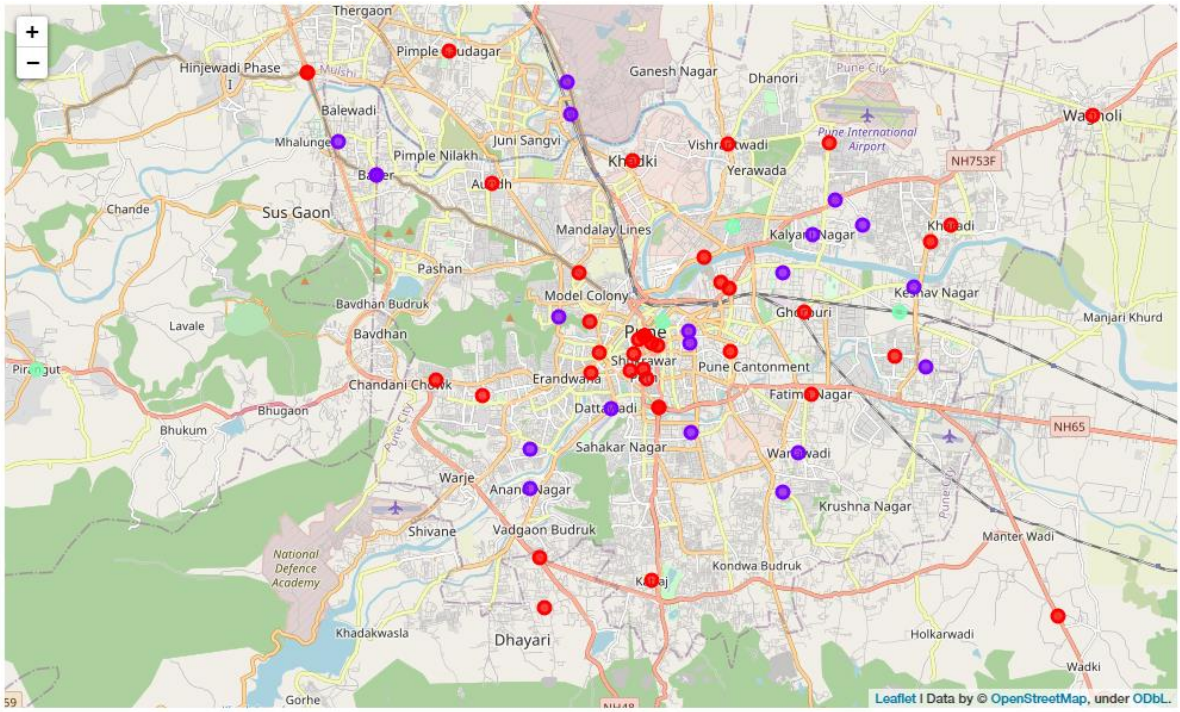
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Resturant”:

- Cluster 0: Areas with high number of Indian restaurants
- Cluster 1: Areas with high number of snacks/breakfast outlets
- Cluster 2: Areas with high concentration of food trucks/Food courts/Fast food

The results of the clustering are visualized in the map below with:

- Cluster 0: Red color
- Cluster 1: Purple color
- Cluster 2: Green color

Out[127]:



5. Discussion

As observations noted from the map in the Results section, most Indian restaurants are located around the central and northern parts of Pune. There is a lot of potential to open Indian Restaurants in the south and south east part of the city with little to no competition. Another opportunity is with food trucks, as they are rare in the north-west part of the city.

This project considers only one factor: the occurrence of the type of restaurant in each area. There are a lot of other factors that should be taken into consideration such as the population distribution, income of the residents in that area, etc. These factors also play a major role in influencing the opening location. As this project was done in a short timespan and due to the lack of datasets, it was not possible to take these factors into consideration. Future research can take place based on these factors.

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 machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. restaurant/franchise owner regarding the best locations to open a new outlet. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods south and south east part of the city are the most preferred locations to open a new outlet. 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.

7. References

- List of areas in Pune: https://www.makemytrip.com/hotels/hotels-pune-area-list.html
- Foursquare Developer Documentation: https://developer.foursquare.com/docs
- Code and documents for the project can be found here: https://github.com/VatsalUnadkat/IBM_Data_Science_Capstone_Project