Coursera IBM Data Science Capstone Project

Opening new catering service outlets in Pune, India

Prepared by: Aditya Mahale

Introduction

For the Capstone project, I have prepared a hypothetical scenario for a food business owner who wants to explore the opening of catering delivery service outlets in different parts of a new city (Pune). The motivation of this project evolves from the fact that Pune is one of the fastest-growing cities in India. Being an IT hub, the city attracts crowds from several states having different food palette. For a food enterprise owner who is aware of the aforementioned fact but oblivious of areas/regions within the city suitable for a specific outlet, the information about specific parts of the city best for opening a particular outlet is of paramount importance. With that thing in mind, finding locations to open such outlets is one of the most important decisions for the owner and I am designing this project to help him find the most suitable locations.

Business Problem

The purpose of this capstone project is to find the most suitable location for the delivery service owner to open limited delivery outlets in different parts of the Pune city. Although several locations can be triangulated based on distance, the owner wants to find the most popular types of food consumed across different parts of the city to set up relevant outlets in relevant parts of the city. Also, there is a budgetary restriction to the opening of the number of outlets. By using data science tools and machine learning models such as clustering, this project aims to provide solutions to answer the business question: In Pune city, if a food business owner wants to open outlets, which suitable locations should he consider opening it?

Target Audience

The delivery service owner who wants to explore best locations to setup a catering outlet in a new city(Pune).

Data

To solve this problem, following data is needed:

1) List of areas in Pune, India.

2) Latitude and Longitude of these areas.

3) 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.

Extracting the data

1) Web scraping to retrieve list of areas in Pune city. Website “makemytrip” alphabetically lists down all areas within Pune city.

2) Extracting Latitude and Longitude data of these areas using geopy package

3) Using Foursquare API to get venue data related in relevant areas.

Methodology

First, I need to get the list of areas/neighborhoods in Pune, India. This is 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 is not in a tabular format; therefore the extraction is done using a function which iterates over the area links.

First, scraping only retrieves the list of areas in Pune. I will have to get their latitudes and longitudes by utilizing Foursquare API to pull the list of restaurants near these areas. To extract the coordinates, I tried using the geopy package but it worked intermittently. Therefore I compiled 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 verify whether these are correct coordinates.

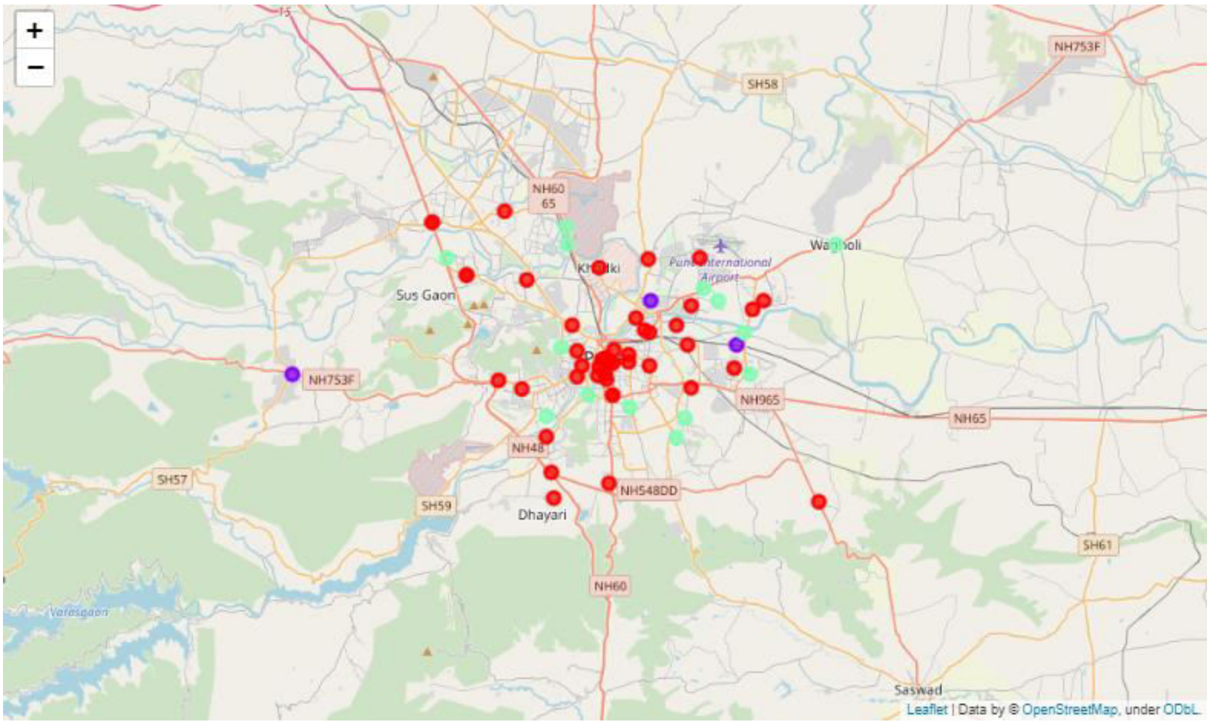
Next, I use Foursquare API to pull the list of top restaurants within 500 meters radius. I have created a Foursquare developer account to obtain an account ID and API key to pull the data. From Foursquare, I can pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each area by grouping the rows by areas and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

I specifically used "restaurant" as a query to search in a particular area. Concentration of specific type of restaurant in a part of a city signifies that the restaurants/food outlets belonging to a particular category is congested in that region.

Lastly, I did the 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 the delivery outlet

Results

Clusters:



The results from k-means clustering show that we can categorize Pune areas into 3 clusters based on types of restaurants in each area:

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

Recommendations

Indian restaurants are mostly present in cluster 0 situated around central and northern parts of Pune. There is a good opportunity to open Indian restaurants in the southern and south-east corner of the city with little competition. Also, food trucks and snack places are rare in the north-west part of the city. Therefore, this project recommends the entrepreneur to open two delivery outlets one located in the south-east side of the city with an Indian menu and another outlet focussing on fast food and food truck items in the north-west part of the city.

Limitations and Suggestions for Future Research

In this project, I only consider one factor: the occurrence of the type of restaurants in each area. Many factors can be taken into consideration such as population density, the income of residents that could influence the decision to open a new catering delivery outlet. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration these factors. Besides, I am relying on the existence of restaurants/food outlets only for this project but future research can take into consideration other variables ratings and popularity of a restaurant.

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.

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/Aditya250892/Coursera_Capstone/tree/master/Final%20Project