COURSERA CAPSTONE

IBM Applied Data Science

Neighborhood Recommendation to open a new Indian Restaurant in Ahmadabad City of India

By: I S Kushagra

May 2020

INTRODUCTION

Food is the most essential part of any living beings life. Food provides us our daily fuel to work, walk, sleep , play and what not. Humans, being an advanced species has always been fond of food and we don't miss to try a different cuisine from time to time. Restaurants are our one stop solution for food whenever we are outside, on vacation or just bored of home made food. The restaurants focused around a particular cuisine offers are always famous among people. Indian food is one such cuisine, which people love to have.

But, selecting a perfect location for a Restaurant is the most important and mind boggling decision. This analysis will help define the features of the city and provide a suggestion of neighborhoods where one might open an Indian Restaurant to stay profitable.

BUSINESS PROBLEM

The main objective of this capstone project is to analyze and select the best location in the city of Ahmedabad, India to open a new restaurant which server Indian Cuisines. Using the methodology and tools of data science and machine learning technique like clustering, this project aims to provide solutions to answer the one major question which is, If in the city of Ahmedabad, a person is looking to open a new Restaurant, where would one recommend that they open it?

DATA

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

- List of neighborhoods in Ahmedabad. This defines the scope of this project which is confined to the city of Ahmedabad, a city in the state of Gujarat in India
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to restaurants. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This webpage (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Ahmedabad) contains a list of neighborhoods in Ahmedabad, with a total of 81 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful soup 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.

Following which, we will use Foursquare API to get the venue data for those neighborhoods. 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 Indian Restaurants category in order to help us to solve the business problem put forward.

Upon receiving the data about the top venues such as museums, shops, restaurants, atms etc from foursquare for every neighbourhood defined in the city of Ahmedabad. Based on the details we can identify the areas where there is a high footfall, i.e. more number of people visit that neighborhood(eg: shopping malls, museums, parks etc), and would most probably be the best location to open the restaurant at.

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.

Example :

Assume neighborhoods “A” and “B” respectively. Upon analysis of these neighborhood we find that neighborhood A has 2 shopping malls, 10 restaurants and 1 museum. While neighborhood “B” has 2 shopping malls, 3 restaurants and no museums.

In this case, it might seem that the footfall at “A” would be higher than ‘B’, as there are 3 visiting locations, but there are already 10 restaurants in the vicinity, which will increase the competition by a lot. Hence, to stay profitable, opening a restaurant at neighborhood “B” would be more logical.

METHODOLOGY

The first step in this project was to get a list of neighbourhoods in the city of Ahmedabad, India. This was achieved by scraping the Wikipedia Page(https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Ahmedabad) using BeautifulSoup. The list of neighborhoods gathered were stored in a pandas dataframe. Next step, was to get the geographical co-ordinates in the form of latitude and longitude for the following neighborhoods. To do so, google's places API was used, and the co-ordinates of the neighborhoods were appended to the existing dataframe. Using the geographical data the neighborhoods were marked by superimposing these locations on a map generated by Folium. This also helps us to second check the obtained data for its correctness.

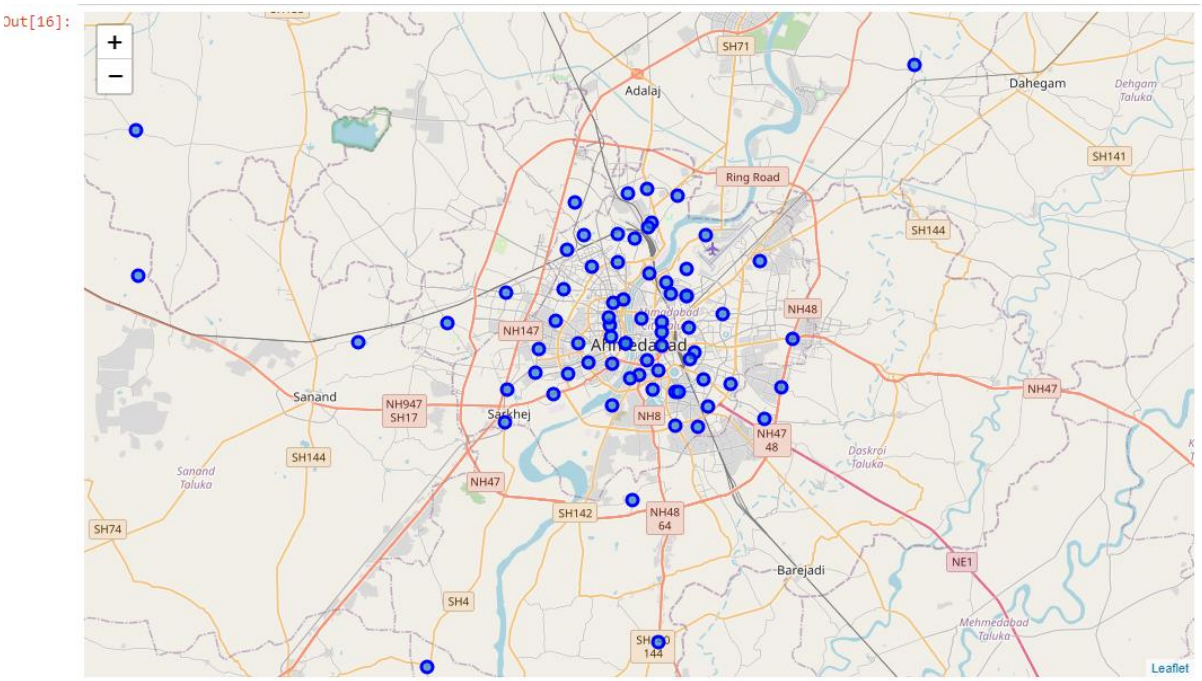

Figure : Neighborhoods in Ahmedabad

Next, using Foursquare API the top 100 venues that were within a radius if 3000 meters were retrieved for each of the neighborhoods. To do this, an API call with required parameters such as, Client Id, Client Secret, Radius, Search query etc were posted, whose reply was stored in a json file. Using this Json file, the required information were filtered and the venue name, venue category, venue latitude, venue longitude was stored into a separate dataframe. The details such as reviews, ratings etc were ignored as such data was not useful for this scenario.

With this data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood 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 analyzing the "Indian Restaurants" data, we will filter the "Indian Restaurant" as venue category for the neighborhoods.

Lastly, a clustering was performed on te data by using k-means clustering. This algoritm identified the number of centroids, and then allocated every data to the nearest cluster, while minimizing the inter cluster distance and maximizing the intra cluster distance. It's a form of

unsupervised machine learning algorithm and is particularly suited to solve the kind of problems similar to this project. The neighborhood were clustered into 3 groups, based on their frequency of occurrence of 'Indian Restaurants'. The results will allow us to identify which neighborhoods have higher concentration of Indian Restaurants while which neighborhoods have fewer number of Indian Restaurants. Based on the occurrence of Indian Restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open a new Indian Restaurants.

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 "Indian Restaurant":

- Cluster 0: Neighborhoods with moderate number of Indian Restaurant
- Cluster 1: Neighborhoods with low number to no existence of Indian Restaurant
- Cluster 2: Neighborhoods with high concentration of Indian Restaurant

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.

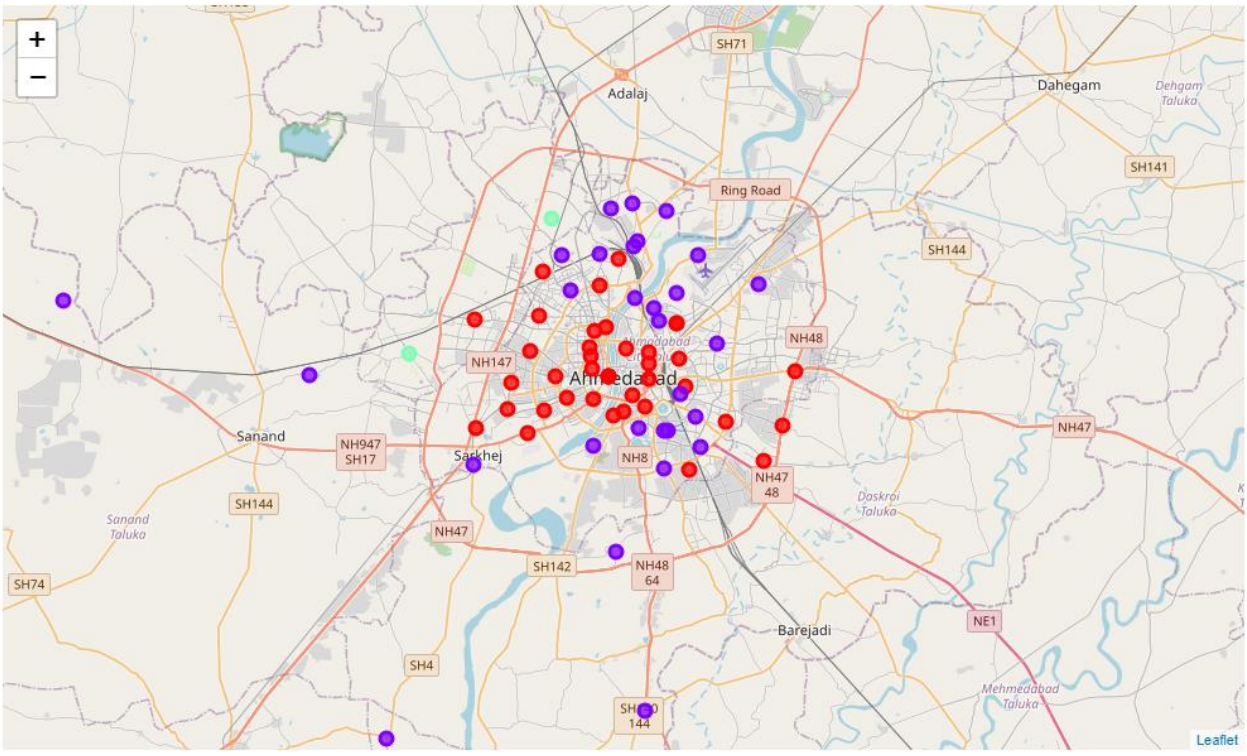

Figure: Output from K-Means Clustering

We can see that most of the Indian restaurants are concentrated around the centre of the city.

DISCUSSION

From the visual Map of the clusters we can observer that the concentration of Indian Cuisine serving restaurants is very well spread around the central areas of the city. However we can observer that cluster 2 contains the highest concentration of Indian Restaurants, represented by the Mint colours circles, are located towards the outskirts of the city, while cluster 1 has a very low number of restaurants to none in these neighborhoods, which is represented by purple colour markers. Cluster 0 shows a moderate number of Indian Restaurants represented by Red color markers.

Referring to the above observations we can say that Cluster 2 would be the least preferable choice to open an Indian Restaurant as it contains the highest number of Indian Restaurants which would lead to an intense competition, and it might take a lot of time for the restaurant to create awareness and image of itself in the market.

Neighborhoods Located in cluster 1 would be the best choice to open a new Indian Restaurants as there is little to no competition in these locations, and it would also be a great opportunity to capitalise the market in these areas.

However, if a person is looking to expand, and open his/her chain of restaurant at a new location. He/she can go with either cluster 1 or 0 assuming that the restaurant already has a known image ad popularity in the community. For such owners, cluster 0 could serve as a good opportunity as opening a restaurant which already has a brand recognition and reputation, can benefit immensely from the existing competition, as consumers may prefer their Indian Restaurant more over the others.

Limitations and Suggestions for Future Research

In this project, we only consider the frequency of occurrence of Indian Restaurants in suggesting the preferable location to open an Indian Restaurant. Meanwhile, many other factors such as population, footfall, population density etc may affect the results widely.
Factors such as the details of restaurants which serve food of different cuisines, and number of shopping malls, parks, public places might also enhance the analysis, as, opening an Indian Restaurant at a public place where there are less restaurants which serve Indian food will be highly recommended.

As Foursquare is not famous in my city, hence details such as people’s reviews, ratings, etc are missing, hence, using a well known provider such as Google’s places API will generate a well defined analysis and accurate results.

CONCLUSION

In this project we went though the processes 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 the recommendations to the relevant stakeholders.

From the details analysis of the problem and geographical data, we can conclude that cluster 1 are the most preferred locations to open a new Indian Restaurant as there is little to no competition, and it will be a good opportunity to capitalize on the market in these areas.