RESTAURANT RECOMMENDER SYSTEM IN INDIA THE PLACE CHENNAI

17/07/2021

IBM Capstone project

Restaurant recommender system in India the place Chennai is a machine learning model, developed to demonstrate as a capstone project to IBM through coursera. It recommends restaurants based on user's likes and dislikes and his previous interest data.

Table of contents

1. Introduction :

Problem background:

Chennai is the capital and largest city of the Indian state of Tamil Nadu. With a population of over 82007976 (as of January 2020), Chennai is located on the south–eastern coast of India in the north–eastern part of Tamil Nadu on a flat coastal plain known as the Eastern Coastal Plains. Its average elevation is around 6.7 meters (22 ft), and its highest point is 60 m (200 ft).Chennai is 4th largest Indian city but it is the 36th metropolitan city the world over.

The diversity of the cuisine available is reflective of the social and economic diversity of Chennai. Roadside vendors, tea stalls, South Indian, North Indian, Muslim food, Chinese and Western fast food are all very popular in the city. Udupi restaurants are very popular and serve predominantly vegetarian cuisine. The Chinese food and the Thai food served in most of the restaurants are can be customized to cater to the tastes of the Indian population. Chennai can also be called a foodie's paradise because of its vast variety of foods and edibles with a touch of Chennai's uniqueness and tradition.

Problem description:

Suppose I travel and keep changing places very frequently. This is very hectic and plus i get to experience very different types of environment, of which I do not have much knowledge about. In such situation, food can be an important factor for decided how you rate your trips and plus also recommending it to the people. Food can also attract people around to world to try it out if it were to be the best. In such scenarios, we need to find the right place, at reasonable cost, to serve us the best possible way. So there are few questions that must be addressed, such as:

1. How many types of foods are available in the restaurant?
2. Which is the most nearest to me with good rating?
3. How many "similar" restaurants are available nearby me?
4. Do the "similar" restaurants cost more? If so, what specialty do those have?

To address such question, ABC Company’s manager decides to allocate this project to me not just to find out solutions to the questions but also build a system that can help in recommending new places based on their rankings compared to the previously visited by me.

Expectations from this recommender system are to get answer for the questions, and in such a way that it uncovers all the perspective of managing recommendations. It is sighted to show:

1. What types of restaurants are present in a particular area?
2. Where are the similar restaurant present based on a preference to particular food?
3. How do different restaurants rank with respect to my preferences?

Target Audience:

Target audiences for this project does not limit to a person who keeps travelling but everyone. People could simply decide to look for a similar restaurant all the time because they are addicted to a specific category of food. People who rarely use restaurants would prefer to have the most rated restaurants nearby them and all this could be easily handled by our recommender system. So target for this project is basically everyone who is exploring different places or similar places.

Success rate:

With restaurants evolving, new food categories emerge, hybrid food starts to be more popular, we need a system that could help us access vast number of food varieties. It is impossible for a person to ask each and every one about their visit to a particular place and also not everyone remembers everything. On the other hand, Computers are good at remembering things, and with Machine learning to its peak, it high time technology will by our personal guidance and help us personally based on our likes and dislikes. So people would care about this project as their personal assistance and success rate could certainly increase with time.

2. Data :

Data requirements:

To find a solution to the questions and build a recommender model, we need data and lots of data. Data can answer question which are unimaginable and non-answerable by humans because humans do not have the tendency to analyze such large dataset and produce analytics to find a solutions.

Let's consider the base scenario:

Suppose I want to find a restaurant, then logically, I need 3 things:
1. Its geographical coordinates (latitude and longitude) to find out where exactly it is located.
2. Population of the neighborhood where the restaurant is located.
3. Average income of neighborhood to know how much is the restaurant worth.

Let’s take a closer look at each of these:
1. To access location of a restaurant, it’s Latitude and Longitude is to be known so that we can point at its coordinates and create a map displaying all the restaurants with its labels respectively.
2. Population of a neighborhood is very important factor in determining a restaurant's growth and amount of customers who turn up to eat. Logically, the more the population of a neighborhood, the more people will be interested to walk openly into a restaurant and less the population, less number of people frequently visits a restaurant. Also if more people visit, better the restaurant is rated because it is accessed by different people with different taste. Hence is very important factor.
3. Income of a neighborhood is also very important factor as population was. Income is directly proportional to richness of a neighborhood. If people in a neighborhood earns more than an average income, then it is very much possible that they will spend more however not always true with very less probability. So a restaurant assessment is proportional to income of a neighborhood.

Data collection:

1. Collecting geographical coordinates is not difficult but after googling for more than 2 days, it was not available on open source data websites such as Wikipedia, India gov website, census report websites etc. So I decided to use Google maps API to fetch latitude and longitude but google API has limited number of calls that I could make with my free

account. So it would take around 15 - 20 days to fetch location of all the neighborhoods in Chennai.

Initially I scrapped list of neighbor's using beautifulSoup4 from [wikipedia](https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai). The table headings becoming the boroughs and data becoming the neighborhoods. Chennai has 8 boroughs and 64 neighborhoods. So i manually googled each neighborhood to find its corresponding latitude and longitude. After doing so, I produced the following data frame.

Borough	Neighborhoods	Latitude	Longitude
South and East Chennai	Adambakkam	12.9880	80.2047
South and East Chennai	Adyar	13.0012	80.2565
South and East Chennai	Alandur	12.9975	80.2006
West Chennai	Alapakkam	13.0490	80.1673
West Chennai	Alwarthirunagar	13.0426	80.1840
West Chennai	Ambattur	13.1143	80.1548
West Chennai	Aminjikarai	13.0698	80.2245
West Chennai	Anna Nagar	13.0850	80.2101
West Chennai	Annanur	13.1184	80.1246
West Chennai	Arumbakkam	13.0724	80.2102

2. Population by neighborhood is again easy to find out given that it's readily available. But in case of Chennai, it is again not the case. i was able to find population data for few cities. [Here is the link](https://geoiq.io/places/Tamil-Nadu/eUfUxMrlBs). Rest other neighborhood population is assumed and may be inaccurate but since this is a demonstrating project, the main idea to get the working model. The data frame for Chennai neighborhood population looks like:

	Borough	Neighborhoods	Population
0	South and East Chennai	Adambakkam	82848
1	South and East Chennai	Adyar	147143
2	South and East Chennai	Alandur	47983
3	West Chennai	Alapakkam	527874
4	West Chennai	Alwarthirunagar	893629

3. Income by neighborhood is again easy to find out given that it's readily available. But in case of Chennai, it is again not the case. i was able to find Income data for main city. [Here is the link](https://www.indiacensus.net/states/tamil-nadu). Neighborhood Income is

assumed and may be inaccurate but since this is a demonstrating project, the main idea to get the working model. The data frame for Chennai neighborhood population looks like:

	Borough	Neighborhoods	AverageIncome
0	South and East Chennai	Adambakkam	18944.099792
1	South and East Chennai	Adyar	56837.022198
2	South and East Chennai	Alandur	41991.817435
3	West Chennai	Alapakkam	6667.447632
4	West Chennai	Alwarthirunagar	53270.063892

4. Foursquare API:

Use of foursquare is focused to fetch nearest venue locations so that we can use them to form a cluster. Foursquare API leverages the power of finding nearest venues in a radius (in my case: 500mts) and also corresponding coordinates, venue location and names. After calling, the following data frame is created:

	Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adambakkam	South and East Chennai	12.9880	80.2047	Venkateshwara Super Market	12.986320	80.205168	Department Store
1	Adambakkam	South and East Chennai	12.9880	80.2047	Ibaco	12.988729	80.205646	Dessert Shop
2	Adambakkam	South and East Chennai	12.9880	80.2047	Deepam Restaurant	12.985380	80.205281	Indian Restaurant
3	Adambakkam	South and East Chennai	12.9880	80.2047	ibaco Adambakkam	12.987358	80.200504	Ice Cream Shop
4	Adyar	South and East Chennai	13.0012	80.2565	Zha Cafe	12.999730	80.254806	Café

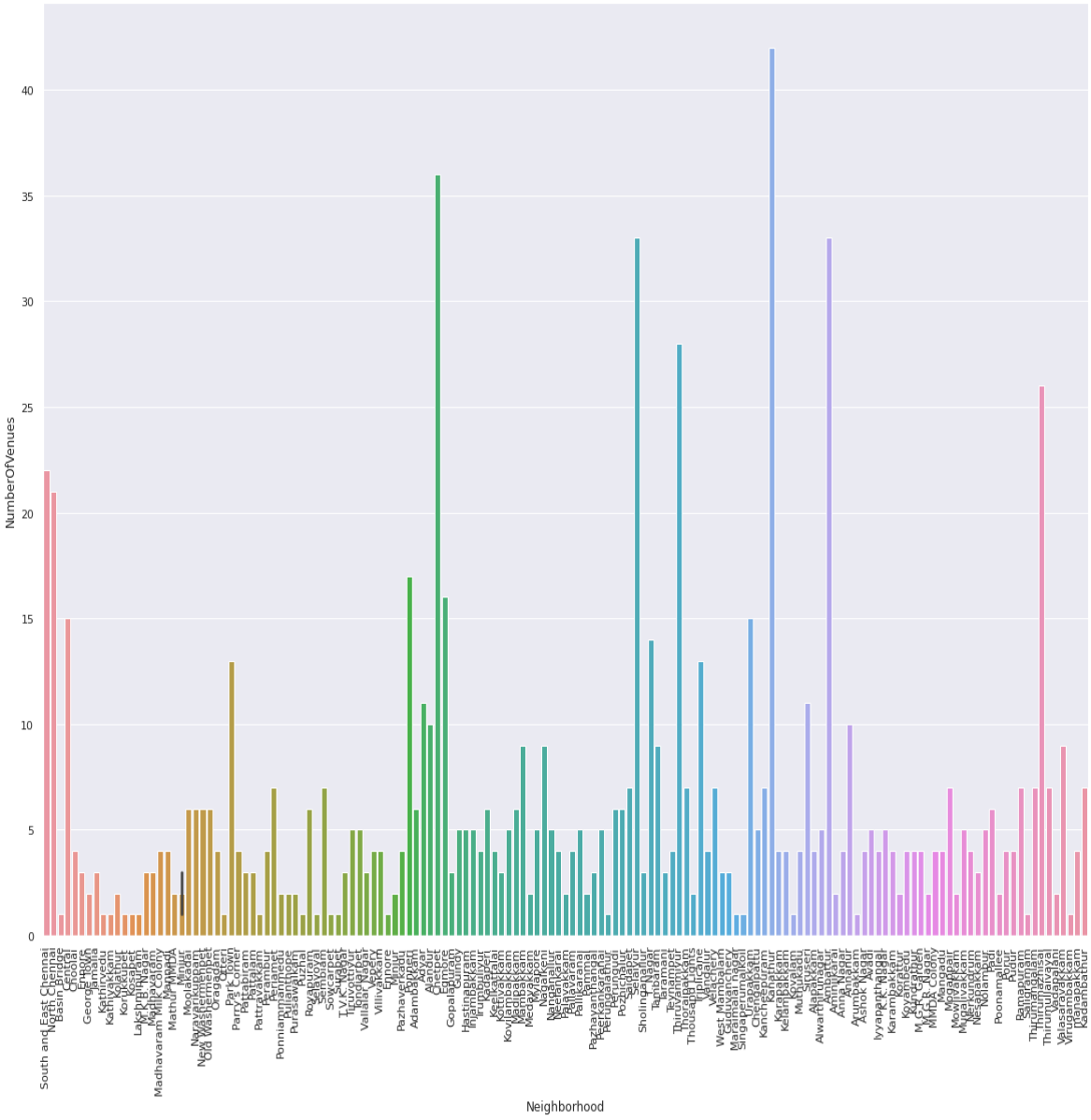
3. Methodology :

Exploratory analysis:

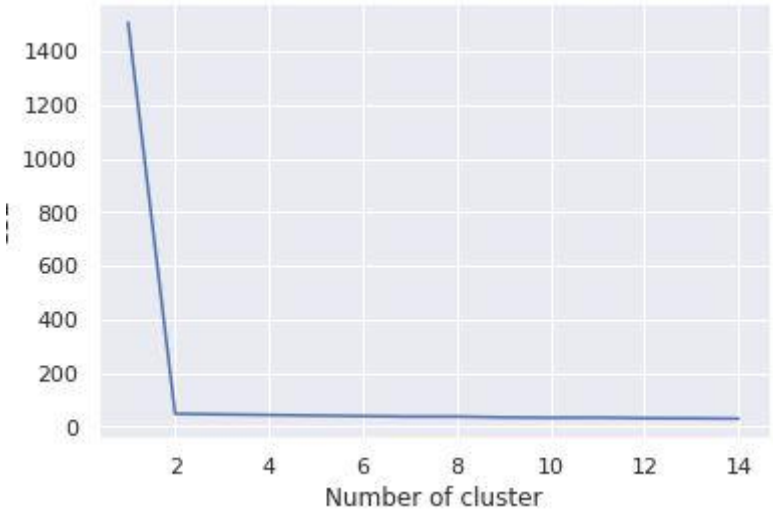
Scrapping the data from different sources and then combining it to form a single-ton dataset is a difficult task. To do so, we need to explore the current state of dataset and then list up all the features needed to be fetched.

Exploring the dataset is important because it gives you initial insights and may help you to get partial idea of the answers that you are looking to find out from the data.

While exploring the dataset, I found out that Kanathur has most number of venues while Basin Bridge has the least.



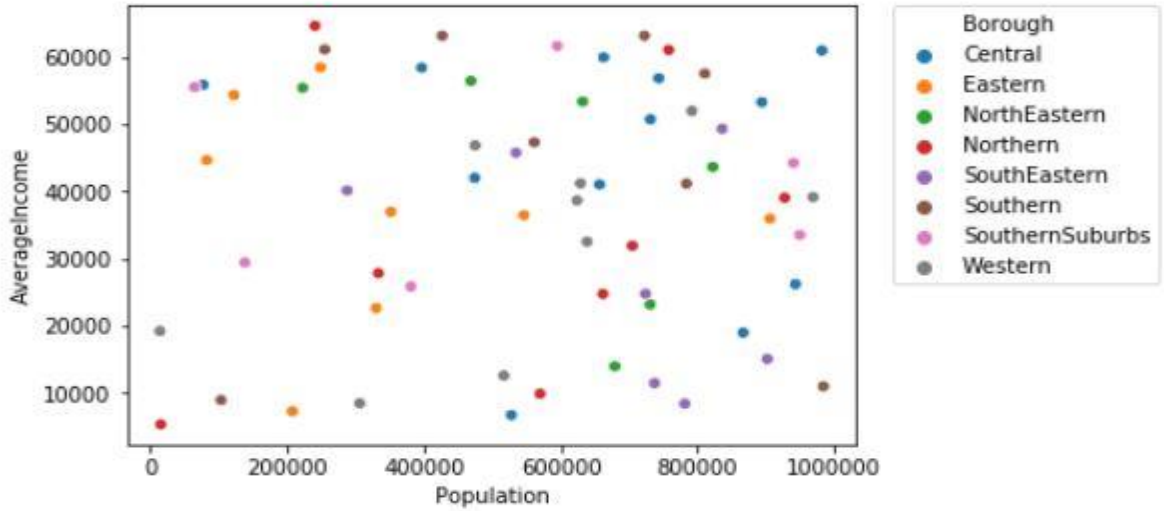
Also while producing graph for number of cluster, I produced a graph to explore all the values for n_clusters and then finding the best by exploring the elbow graph.



Inferential analysis:

Most important factors while building the recommender system were population and income. They are the most import factor because they have a nonlinear relationship according to our dataset.

It needed to make some inferential analysis to understand this nonlinear relationship. As the amount of population increases, it does not necessarily mean that average income of a neighborhood will also increase. It is true to most of the case but also many cases differ to follow this trend. Similarly, a neighborhood with less number of people may not necessarily have less average income. It is possible to have less number of people and more income and vice versa. This can be inferred from the following graph:



4. Result :

The result of the recommender system is that it produces a list of top restaurants and the most common venue item that the user can enjoy. During the runtime of the model, a simulation was done by taking ‘Porur’ as the neighborhood and then processed through our model so that it could recommend neighborhoods with similar characters as that of ‘Porur’.

The following image shows the result:

```
top3_df.head(3)
```

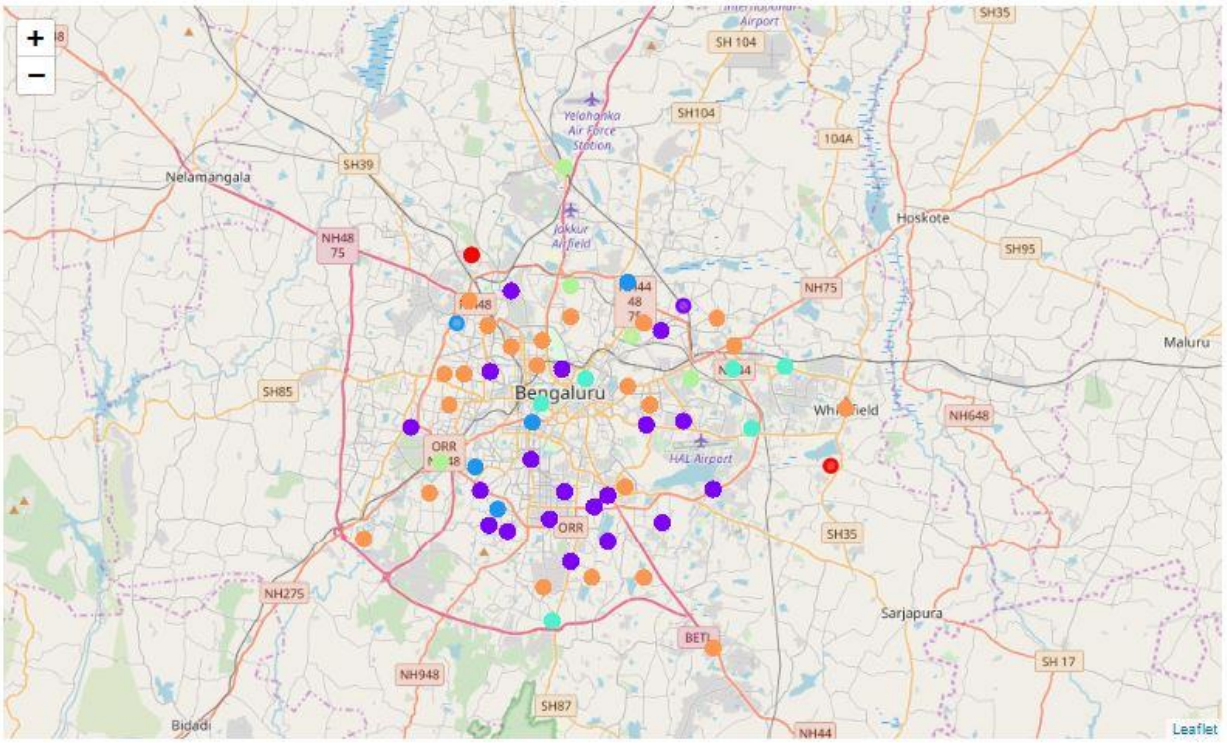
	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Adambakkam	Venue Category_Ice Cream Shop	Venue Category_Indian Restaurant	Venue Category_Department Store	[0.2447784110101273]
1	Adyar	Venue Category_Indian Restaurant	Venue Category_Pizza Place	Venue Category_Café	[0.48269740096662883]
2	Alandur	Venue Category_Indian Restaurant	Venue Category_Train Station	Venue Category_Breakfast Spot	[0.35180044626827367]

5. Discussion :

Since there was a nonlinear relationship between income and population, it can be concluded that we must always perform inferential approach to find relationship among different set of features. Also during clustering, similar neighborhoods must be dumped into the right cluster.

The following graph shows the clusters:

Out[61]:



Another observation that we can make is that choosing number of clustering could produce very diverse results. Some may be over fitted or some may be under fitted. Hence analysis of number of clusters must be done. Ref elbow_graph in the Methodology section.

6. Conclusion :

The recommender system is a system that considers factors such as population, income and makes use of Foursquare API to determine nearby venues. It is a powerful data driven model whose efficiency may decrease with more data but accuracy will increase. It will help users to finish their hunger by providing the best recommendation to fulfil all their needs.