RESTAURANT RECOMMENDER SYSTEM

15/07/2020

Coursera IBM Data Science Professional Capstone project

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

Table of contents

1. Introduction :

Problem background:

Pune, also known as Poona is a city in the Indian State of Maharashtra and the eighth most populous city in India, with an estimated population of about 7.4 million as of 2020. Pune is a sprawling city in the western Indian state of Maharashtra. It was once the base of the Peshwas (prime ministers) of the Maratha Empire, which lasted from 1674 to 1818. It's known for the grand Aga Khan Palace, built in 1892 and now a memorial to Mahatma Gandhi, whose ashes are preserved in the garden. The 8th-century Pataleshwar Cave Temple is dedicated to the Hindu god Shiva.

It has been ranked as one of the most livable cities in India several times. Due to its growing demand and rise in the number of students for quality education, it is also known as oxford of India. It has a population of 7.27 million and is ranked 8th in the list of metropolitan cities in India. The rise and demand for cuisine is reflection of the cultural diversity present at the moment. Various cuisines available is reflective of the social and economic diversity of Pune. Roadside vendors, tea stalls, South Indian, North Indian, Western fast food are all very popular in the city. Kolhapuri restaurants are very popular and serve authentic spicy cuisines. Pune can also be called a foodie's paradise because of its vast variety of foods and edibles with a tinge of Marathi masalas. Vada Pav, Misal Pav, Bhakri are one of must try vegetarian fast food dish native to the state of Maharashtra.

Problem description:

It is hectic if someone travels a lot and keeps on changing places frequently. Plus the experience vary with changes in environment if the new location being travelled to is unknown. Under such scenarios, food plays an important factor for deciding how one rates his/ her trips and also recommmends it to the people who plan to visit those places or restaurants. Food can also attract people around the world to try it out if others have been to those places and recommend it to be a good one. Under such scenarios, we need to find the right place, at reasonable cost, to serve us the best possible way. Questions that must be addresed while looking for such problem are:

- How many types of foods are available in the restaurant ?
- Which is closest to user with good rating ?
- How many "similar" restaurants are available nearby ?
- Do the "similar" restaurants cost more ? If so, what speciality do that have ?
- To address such question, XYZ company's manager decides to allocate this project 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 is to get answer for the questions, and in such a way that it uncovers all the perspective of managing recommendations. It sights to show us:

- What types of restaurants are present in a particular area ?
- where are the other restaurants that are present based on a preference to a particular food ?
- 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 the most rated restaurants nearby them and all this could be easily handled by our recommender system. So, target for this project would be everyone who is willing to exploring different or similar places based on their choice and interest.

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 everyone about their visit to a 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:

In order to find a solution to the questions and build a recommender model, we need data. Data is the only thing that can answer questions which we are unable to answer by humans because as human beings, we do not have the tendency to analyses such large dataset and produce analytics to find solutions.

Let's consider a scenario:

Suppose we want to find a restaurant then, logically we need to look at 3 things :

- Its geographical coordinates(latitide and longitude) to find our where exactly it is located.
- Population of the neighborhood where the restaurant is located.
- Average income of neighborhood to know how much is the restaurant worth.

Let's take a closer look at each of these:

- To access location of a restaurant, its 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.
- 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 visit a restaurant. Also if more people visit, better the restaurant is rated because it is accessed by different people with different taste. Hence it is very important factor.
- 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's assessment is proportional to income of a neighborhood.

Data collection:

- Collecting geographical coordinates is not difficult but after intensive search, 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. Initially I scrapped list of neighbor's using beautifulSoup4 from wikipedia. The table headings becoming the boroughs and data becoming the neighborhoods. Pune has 5 boroughs and 102 neighborhoods. After having googled each neighborhood to find its corresponding latitude and longitude, the following dataframme came out as the output.

	Borough	Neighbourhoods	Latitude	Longitude
0	Pune Municipal Corporation	Ambegaon	19.144350	73.729103
1	Pune Municipal Corporation	Aundh	18.558130	73.807281
2	Pune Municipal Corporation	Baner	18.562120	73.802544
3	Pune Municipal Corporation	Bavdhan Khurd	18.472270	73.884640
4	Pune Municipal Corporation	Bavdhan Budruk	18.456620	73.844100
5	Pune Municipal Corporation	Balewadi	18.574739	73.770889
6	Pune Municipal Corporation	Shivajinagar	18.526110	73.844131
7	Pune Municipal Corporation	Bibvewadi	18.483370	73.855614
8	Pune Municipal Corporation	Bhugaon	18.488890	73.741562
9	Pune Municipal Corporation	Bhukum	18.498380	73.719390

- Population by neighborhood is again easy to find out given that its readily available. But in case of Pune, it is again not the case. i was able to find population data for few cities. Here is the link. 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 dataframe for Pune neighborhood population looks like:

	Borough	Neighbourhoods	Population
0	Pune Municipal Corporation	Ambegaon	54791
1	Pune Municipal Corporation	Aundh	17899
2	Pune Municipal Corporation	Baner	84725
3	Pune Municipal Corporation	Bavdhan Khurd	83399
4	Pune Municipal Corporation	Bavdhan Budruk	24639
5	Pune Municipal Corporation	Balewadi	25542
6	Pune Municipal Corporation	Shivajinagar	51600
7	Pune Municipal Corporation	Bibvewadi	40332
8	Pune Municipal Corporation	Bhugaon	23005
9	Pune Municipal Corporation	Bhukum	37229

- Income by neighborhood is again easy to find out given that its readily available. But incase of Pune, it is again not the case. i was able to find Income data for main city. Here is the link. Neighborhood Income is assumed and may be inaccurate but since this is a demonstrating project, the main idea to get the working model. The dataframe for Pune neighborhood population looks like:

|	Borough	Neighbourhoods	Average Income
0	Pune Municipal Corporation	Ambegaon	21554.136889
1	Pune Municipal Corporation	Aundh	14995.002708
2	Pune Municipal Corporation	Baner	29959.393522
3	Pune Municipal Corporation	Bavdhan Khurd	14971.799874
4	Pune Municipal Corporation	Bavdhan Budruk	14986.677169
5	Pune Municipal Corporation	Balewadi	14927.853254
6	Pune Municipal Corporation	Shivajinagar	14992.906561
7	Pune Municipal Corporation	Bibvewadi	14970.687088
8	Pune Municipal Corporation	Bhugaon	14914.758979
9	Pune Municipal Corporation	Bhukum	14993.989437

- FourSquare API: Use of foursquare has been used 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 dataframe is created:

	Neighbourhoods	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Aundh	Pune Municipal Corporation	18.55813	73.807281	Starbucks	18.556595	73.809153	Coffee Shop
1	Aundh	Pune Municipal Corporation	18.55813	73.807281	Crosswords	18.556177	73.809131	Bookstore
2	Aundh	Pune Municipal Corporation	18.55813	73.807281	Naturals Icecream	18.556192	73.809169	Ice Cream Shop
3	Aundh	Pune Municipal Corporation	18.55813	73.807281	Taareef	18.556077	73.809120	Indian Restaurant
4	Aundh	Pune Municipal Corporation	18.55813	73.807281	Westend mall	18.561814	73.807220	Shopping Mall

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 Wakad has most number of venues while Wadgaon sheri and Talawade had 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.
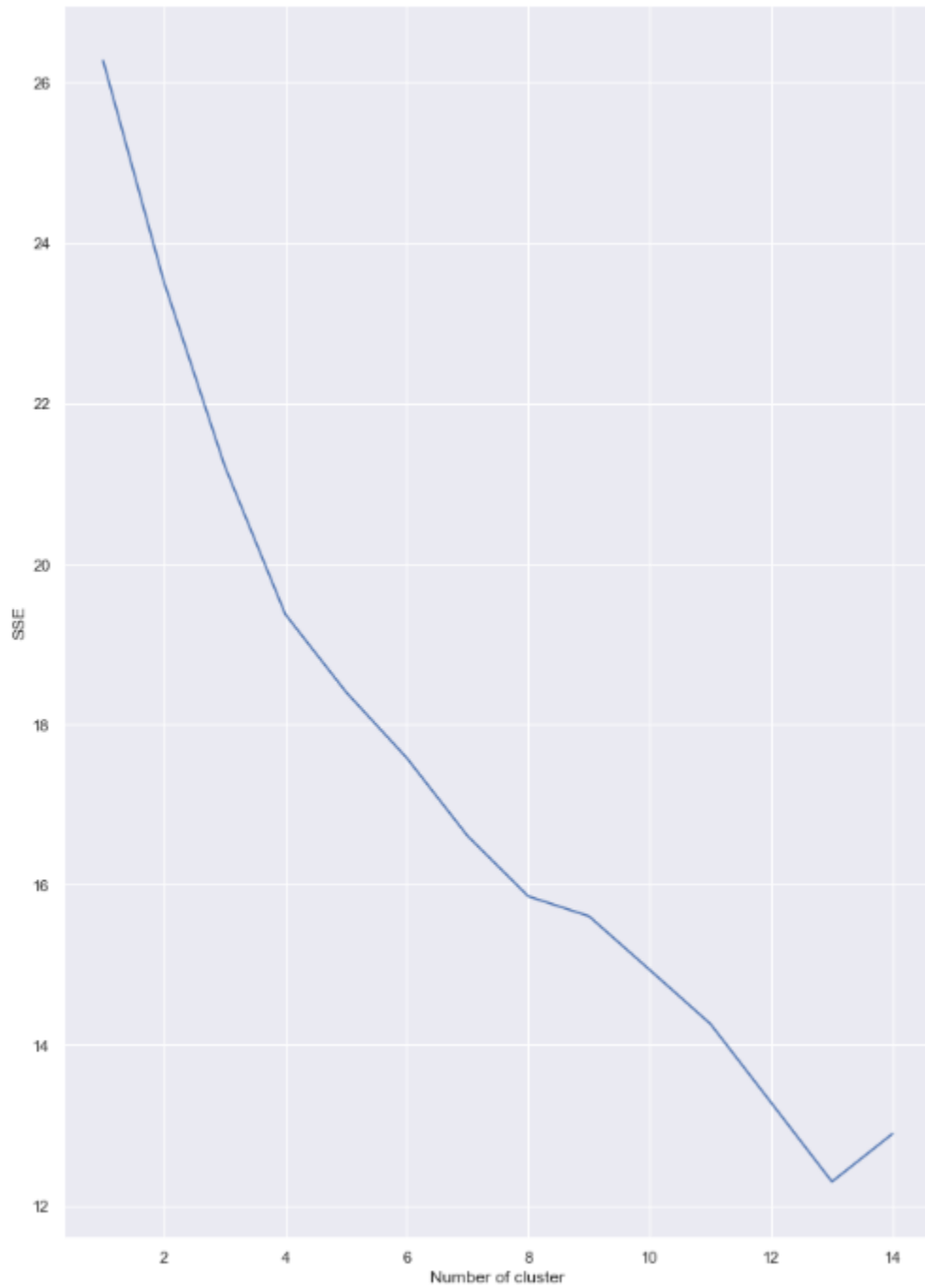
The optimal value for cluster is 5 from the graph

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.

3. 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 **'Hinjawadi'** as the neighborhood and then processed through our model so that it could recommend neighborhoods with similar characters as that of **Hinjawadi**.

The following image shows the result:

	Neighbourhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Aundh	Venue Category_Indian Restaurant	Venue Category_Dessert Shop	Venue Category_Bakery	[0.33080879075424324]
1	Balewadi	Venue Category_Asian Restaurant	Venue Category_Tea Room	Venue Category_Restaurant	[0.42512906930448313]
2	Bhavani Peth	Venue Category_ATM	Venue Category_Furniture / Home Store	Venue Category_Diner	[0.37365876153498023]

4. 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:
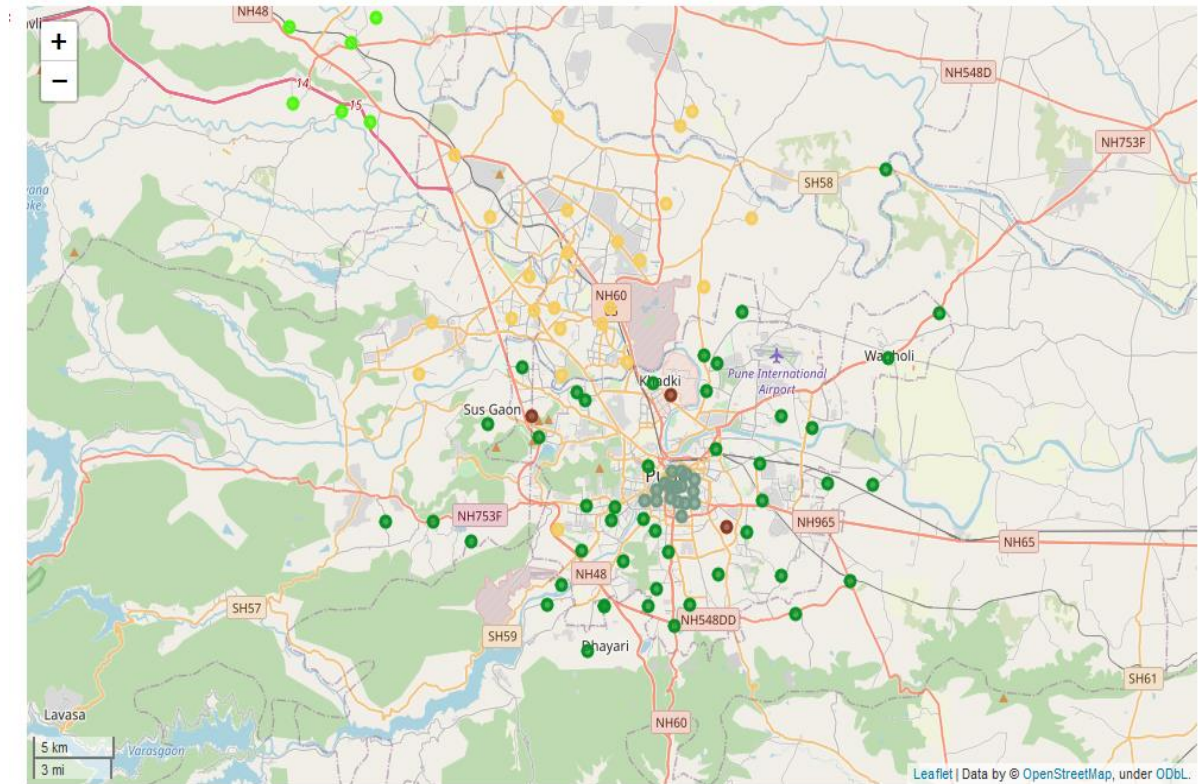
Another observation that we can make is that choosing number of clustering could produce very diverse results. Some may be over fitted or some may be under fitted. Hence analysis of number of clusters must be done. Ref elbow_graph in the Methodology section.

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.