Restaurant Recommender System

IBM Capstone Project

Authored by: D. Vivek Nayudu

Logo
Name

1. Introduction

Problem Background:

Nagpur is the third largest city and the winter capital of the Indian state of Maharashtra. It is the 13th largest city by population in India. According to an Oxford Economics report, Nagpur is projected to be the fifth fastest growing city in the world from 2019-2035 with an average growth of 8.41%. It has been proposed as one of the Smart Cities in Maharashtra and is one of the top ten cities in India in Smart City Project execution. It is a major commercial and political centre of the Vidarbha region of Maharashtra.

The Vidharbha region has its own distinctive cuisine known as the Varhadi cuisine or Saoji cuisine. Nagpur is also famous for its oranges, which have some typical qualities have recently begun to attract international attention. Numerous beverages are made out of the oranges. Santra Barfi is also a famous dish, arising from orange which is produced locally in Nagpur. Mominpura is a majority Muslim area of the city and it is famous for its Mughal dishes and Biryani. The city is also famous for rare black chickens called Kadaknath Chicken which are cooked in varhadi style. Nagpur is also famous for tarri poha, a variety of flattened rice, and has many food joints; each having their own way of preparing and serving it. Samosas are also famous in Nagpur and is available at many restaurants and food spots. Another famous food is Patodi and Kadhi.

Problem Statement:

With growing state of affairs the city several restaurants have cropped up in city. It has become a hub for road side eateries and some old food joints and have rose into prominence. This puts students, visiting government officials and tourists in a state of dilemma when comes to find good places to eat.

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 addresed, such as :

1. Which cuisines are available in the restaurant?
2. Which restaurant nearest to me with good ratings?
3. How many "similar" restaurants are available near by me?
4. Do the "similar" restaurants cost more? if so, what is their speciality?

To address such questions, ABC company decides to allocate this project to me not to just find out answers 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 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 students or tourists 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 handed 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 everyone 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:

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

Collecting geographical coordinates is not difficult but after googling for few hours, it was not available on open source data websites such as wikipedia, india gov website, census report websites etc. Initially i scrapped list of localities using beautifulSoup4 from wikipedia. The table headings becoming the boroughs and data becoming the neighborhoods. Localities in Nagpur do not have boroughs. I have randomly classified Nagpur into 5 boroughs and 42 neighborhoods. So i manually googled each neighborhood to find its corresponding latitude and longitude. After doing so, i produced the following dataframe.

	Borough	Neighborhood	Latitude	Longitude	Population	Averge Income
0	Eastern	Mahal	21.144979	79.105782	23000	18450
1	Eastern	Sitabuldi	21.152451	79.080559	45435	34545
2	Eastern	Dhantoli	21.131060	79.077120	67456	83476
3	Eastern	Itwari	21.149870	79.109741	56867	64356
4	Eastern	Mominpura	21.151280	79.103550	45345	85769

Population by neighborhood is again easy to find out given that its readily available. But in case of Nagpur, it is again not the case. I was able to find population data for few areas. 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.

Income by neighborhood is again easy to find out given that its readily available. But in case of Nagpur, it is again not the case. i was able to find Income data for main city. 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 Nagpur neighborhood looks like :

	Borough	Neighborhood	Latitude	Longitude	Population	Normalized Population	Normalized Population	Averge Income	Normalized Income
0	Eastern	Mahal	21.144979	79.105782	23000	-1.803267	-1.803267	18450	-1.725391
1	Eastern	Sitabuldi	21.152451	79.080559	45435	-0.749180	-0.749180	34545	-1.015375
2	Eastern	Dhantoli	21.131060	79.077120	67456	0.285455	0.285455	83476	1.143169
3	Eastern	Itwari	21.149870	79.109741	56867	-0.212059	-0.212059	64356	0.299709
4	Eastern	Mominpura	21.151280	79.103550	45345	-0.753409	-0.753409	85769	1.244323
5	Eastern	Dharampeth	21.139700	79.063100	78888	0.822576	0.822576	45790	-0.519313
6	Eastern	Ramdaspeth	21.135800	79.076500	65656	0.200884	0.200884	26547	-1.368199
7	Eastern	Shraddhanand Peth	21.124100	79.058400	54656	-0.315941	-0.315941	23478	-1.503585
8	Western	Sadar	21.163000	79.073500	55667	-0.268440	-0.268440	87587	1.324522
9	Western	Civil Lines	21.152800	79.074700	32545	-1.354804	-1.354804	56779	-0.034543
10	Western	Gandhibagh	21.150400	79.102600	43546	-0.837933	-0.837933	83459	1.142420
11	Western	Nandanvan	21.138800	79.127600	68787	0.347990	0.347990	74895	0.764627
12	Western	Kalamna	21.164700	79.146900	87674	1.235378	1.235378	64756	0.317354
13	Western	Wardhaman Nagar	21.150000	79.137600	67735	0.298563	0.298563	86787	1.289231
14	Western	Seminary Hills	21.166000	79.055000	23467	-1.781325	-1.781325	85768	1.244279
15	Western	Police Line Takli	21.177200	79.067700	76754	0.722312	0.722312	56762	-0.035293
16	Western	Mankapur	21.184600	79.078800	43557	-0.837416	-0.837416	34576	-1.014008
17	Western	Pachpaoli	21.162000	79.105800	53457	-0.372275	-0.372275	37486	-0.885636
18	Western	Vayusena Nagar	21.161600	79.015000	34523	-1.261870	-1.261870	84379	1.183004
19	Northern	Ravi Nagar	21.149100	79.055000	78788	0.817878	0.817878	47578	-0.440437
20	Northern	Byramji Town	21.173300	79.082200	78784	0.817690	0.817690	57687	0.005512
21	Northern	Chaoni	21.146633	79.088860	54645	-0.316457	-0.316457	38457	-0.842801

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 dataframe is created:

|	Borough	Neighborhood	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Eastern	Mahal	21.144979	79.105782	karmaveer book depot	21.145136	79.105995	Bookstore
1	Eastern	Mahal	21.144979	79.105782	Ghate Milk Center	21.147249	79.103875	Indian Restaurant
2	Eastern	Mahal	21.144979	79.105782	raam bhandar	21.147359	79.108110	Breakfast Spot
3	Eastern	Mahal	21.144979	79.105782	Ajra	21.146498	79.102303	Asian Restaurant
4	Eastern	Sitabuldi	21.152451	79.080559	Inspire Apple Store	21.153496	79.078711	Electronics Store

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 Dharampeth has most number of venues while Jaripatka has the least.
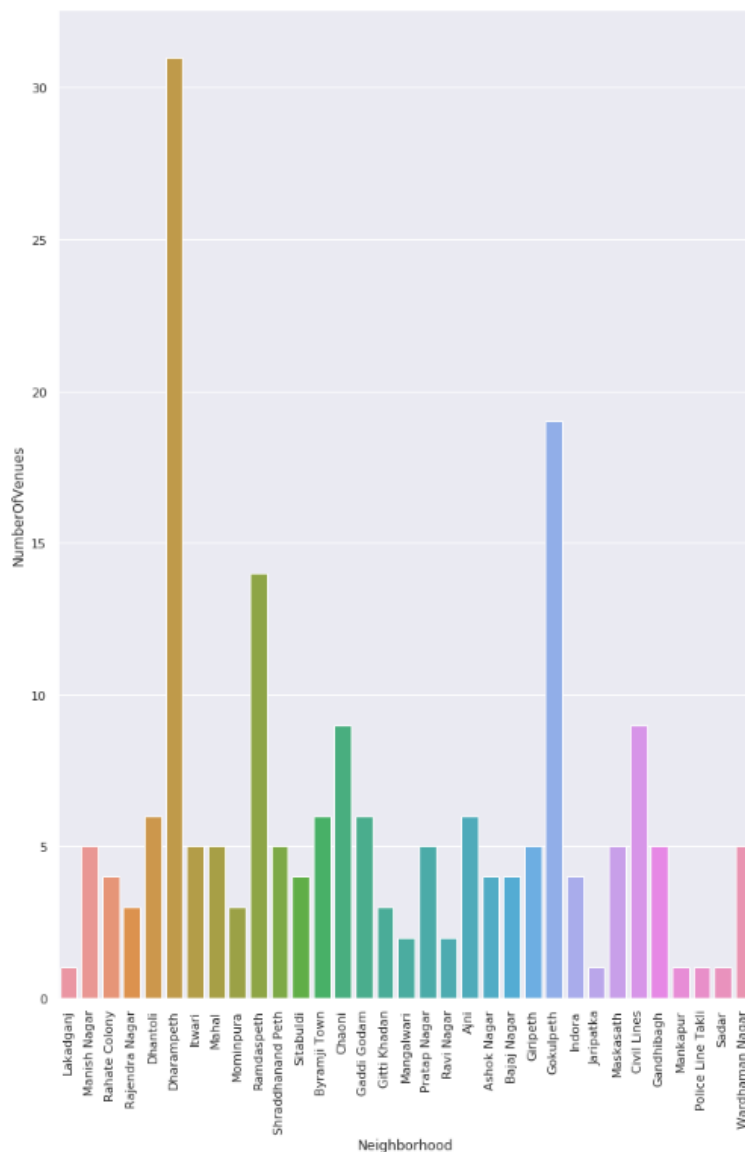
Also while producing graph for number of cluster, I produced a graph to explore all the values for n_clusters and then finding the best by exploring the elbow graph.
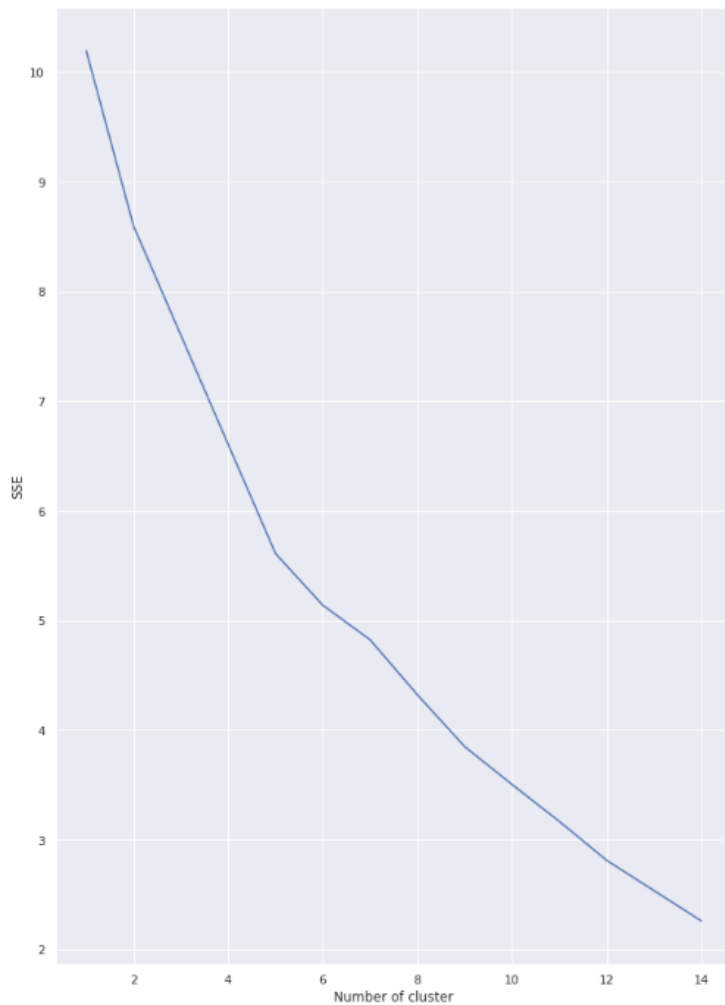
Inferential Analysis:

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.

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

The following image shows the result:

|	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Ajni	Venue Category_Platform	Venue Category_Train Station	Venue Category_Business Service	[0.9333333333333333]
1	Bajaj Nagar	Venue Category_Plaza	Venue Category_Asian Restaurant	Venue Category_Falafel Restaurant	[0.501003109799038]
2	Byramji Town	Venue Category_Gym	Venue Category_Fast Food Restaurant	Venue Category_Department Store	[0.7068074161641663]

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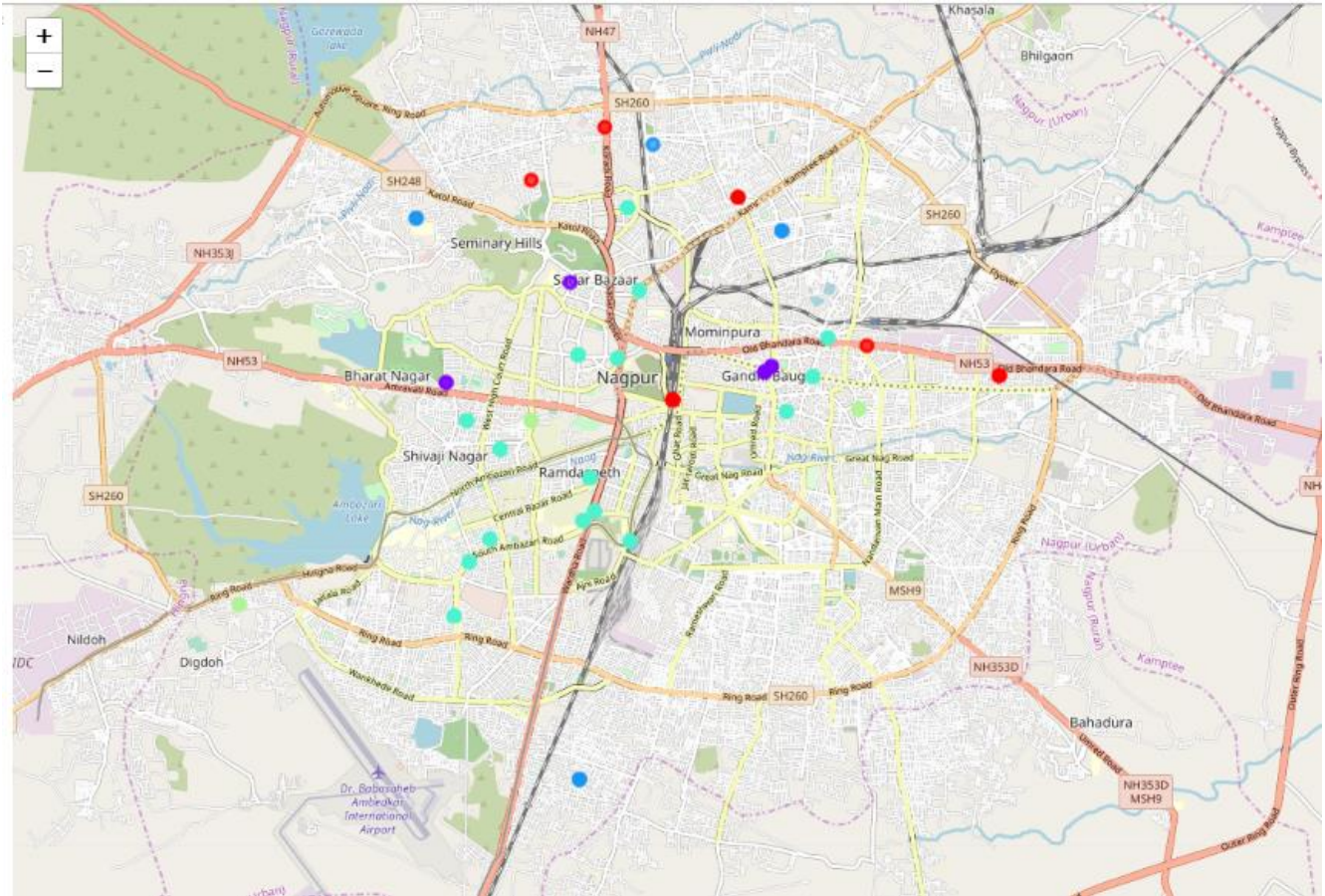
Another observation that we can make is that choosing number of clustering could produce very diverse results. Some may be over fitted or some may be under fitted. Hence analysis of number of clusters must be done. Refer the elbow graph in the Methodology section.

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.