The Battle of Neighborhoods

Recommendation system in a machine learning model, developed to recommend neighborhood to open a new ‘Multiplex’ in New Delhi, India

Swati Jindal | Applied Data Science Capstone by IBM through Coursera | April, 2020

Table of Contents

1. INTRODUCTION / BUSINESS PROBLEM

A) BACKGROUND

New Delhi, the capital of India, is spread across in 1,484 sq. km. and is the largest commercial center in northern India. Against the national per capita income of Rs.1,34,432, Delhi's per capita income stands at Rs.3,89,143 (annual for FY2019-20). The annual GDP growth rate of Delhi is also at 7.42% against national average of 5%. In fact, 85% of Delhi's economic activity is generated in the service sector which is also the driving factor for growth. More specifically, banking and insurance, real estate, trade, tourism and communications are driving the progress in the state. Further, the cumulative FDI inflows to Delhi during April 2000–June 2019 amounted to US$ 89.68 billion.

With such large work force driven by service sector and the limited area of New Delhi in sq. km., recreation activities play a dominant role during weekends considering the higher per capita income of the residents of New Delhi.

It is pertinent to mention that Indian cinema has an annual output of around 2000 feature films per year in Hindi as well as regional languages. Further, maximum foreign language feature films including English films have large viewership in India, especially in metropolitan cities, including New Delhi.

With such attractive economic prospects as well as large viewership of cinemas, property developers have huge opportunities to explore construction of profitable multiplexes. However, arriving at the best location for building multiplexes is one of the most challenging decisions for deriving its prospects of being a success.

B) DESCRIPTION OF PROBLEM

The objective of the Capstone Project is to analyse and select best cluster and location in New Delhi, India to open a new 'Multiplex'. The data science methodology and machine learning techniques such as clustering and statistical plotting library of seaborn have been used to provide best solutions at arriving at the best cluster along with illustrative example to derive location in a neighborhood.

In addition to addressing the query of arriving at the best cluster for a Multiplex in current scenario, the Project also builds a recommendation system on the basis of various parameters, like type of venues, neighborhood etc. to provide insights into development of ‘Multiplex’ in New Delhi, India.

C) TARGET AUDIENCE

The target audience in the instant case are investors and property developers keen to invest or construct new Multiplex in New Delhi, India. With the present economic growth rate, heavy dependence of New Delhi on service sector and broad spectrum of feature films being released in India, the development of Multiplex appears to be an attractive business proposal in New Delhi.

D) SUCCESS RATE

With application of data science methodology and use of machine learning techniques, detail insights can be provided to various stakeholders regarding clusters and neighborhoods where multiplexes can be developed.

2. DATA ANALYSIS - A DESCRIPTION OF THE DATA AND HOW IT WILL BE USED TO SOLVE THE PROBLEM

DATA REQUIREMENT AND COLLECTION

(i) List of neighborhoods in New Delhi, India have been collected for the instant Project. This is required to segregate neighborhoods into clusters and carry out further analysis.

(ii) Latitude and Longitude coordinates of identified neighborhoods have been collected for the instant Project. This has been used for plotting on maps as well as to arrive at the venue data of various neighborhoods.

(iii) Use of Foursquare API for fetching the nearest venue locations for defining clusters along with details of venues and their names. This has been used to make detailed analysis for arriving at the best locations for developing Multiplex.

(iv) Venue data, especially that related to ‘Multiplex’. This has helped in clustering neighborhoods for arriving at the results of the project and making final recommendation.

3. METHODOLOGY

A) WEB SCRAPING AND BUILDING DATAFRAME

- In this step, list of neighborhoods were scared from Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi)

- The same was done using Python requests and beautifulsoup packages to extract the list of neighborhood data. With these packages, the list of names of neighborhoods of New Delhi, India were populated into a dataframe.

B) GEOGRAPHICAL COORDINATES OF IDENTIFIED NEIGHBORHOODS

- In this step, Geocoder package has been used to extract geographical coordinates of identified neighborhoods of New Delhi, India. After gathering the data, the same has been populated in pandas dataframe.

- The identified neighborhoods are then visualized in a map using Folium package.

- The process has helped in checking the output of geographical coordinates data returned by Geocoder and its correct plotting on the map.

C) FOURSQUARE API FOR EXPLORING IDENTIFIED NEIGHBORHOODS

- Foursquare API has been used to fetch nearest venue locations so as to use them in forming a cluster. In the present case, Foursquare API has been used to get top 100 venues that are within the radius of 2000 meters.

- For the said purpose, a new Foursquare Developer Account has been opened for obtaining Client ID and Client secret key.

- Thereafter, API calls are made to Foursquare API by passing the geographical coordinates of identified neighborhoods in a Python loop.

- Foursquare returns venue data in JSON format from which venue name, venue category, venue latitude and venue longitude are extracted.

- Each neighborhood is then analyzed by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category.

- Data is accordingly prepared for clustering. Thereafter, the data is filtered for analyzing data related to ‘Multiplex’.

D) K-MEANS CLUSTERING

- In this step, clustering on data is performed using k-means clustering. In this method, k number of centroids are identified and then allocated to nearest cluster, while keeping centroids as small as possible.

- This method is useful in the instant case considering the fact that clustering would help in narrowing down the target neighborhood by eliminating clusters and then coming at the final result of the problem at hand.

- In the instant case, neighborhoods have been clustered into three (3) clusters based on frequency of occurrence of ‘Multiplex’. The clusters identify the level of concentration of multiplexes in each neighborhood and which is the best suited neighborhood with fewer number of multiplexes and providing opportunity to open a new multiplex.

- Thereafter, clusters are then visualized in a map using Folium package.

E) STATISCIAL ANALYSIS OF CLUSTERS

- Considering the first cluster has very sparse and wide neighborhood data related to multiplexes, it is considered appropriate to carry out statistical analysis on the other two clusters.

- The statistical plotting library of seaborn is then used on two clusters to analyze the results arrived in the analysis carried out while using k-means clustering.

- On identification of the cluster, as an illustrative case, one neighborhood is explored to arrive at the final decision of opening a multiplex.

4. RESULT OF THE ANALYSIS

A) RESULTS OF K-MEANS CLUSTERING

The results of k-means clustering shows that on categorizing the neighborhoods into three (3) clusters, frequency of occurrence of ‘Multiplex’ can be analyzed as below:

a) Cluster 0: Neighborhoods with least number of multiplexes
b) Cluster 1: Neighborhoods with high concentration of multiplexes
c) Cluster 2: Neighborhoods with moderate concentration of multiplexes

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

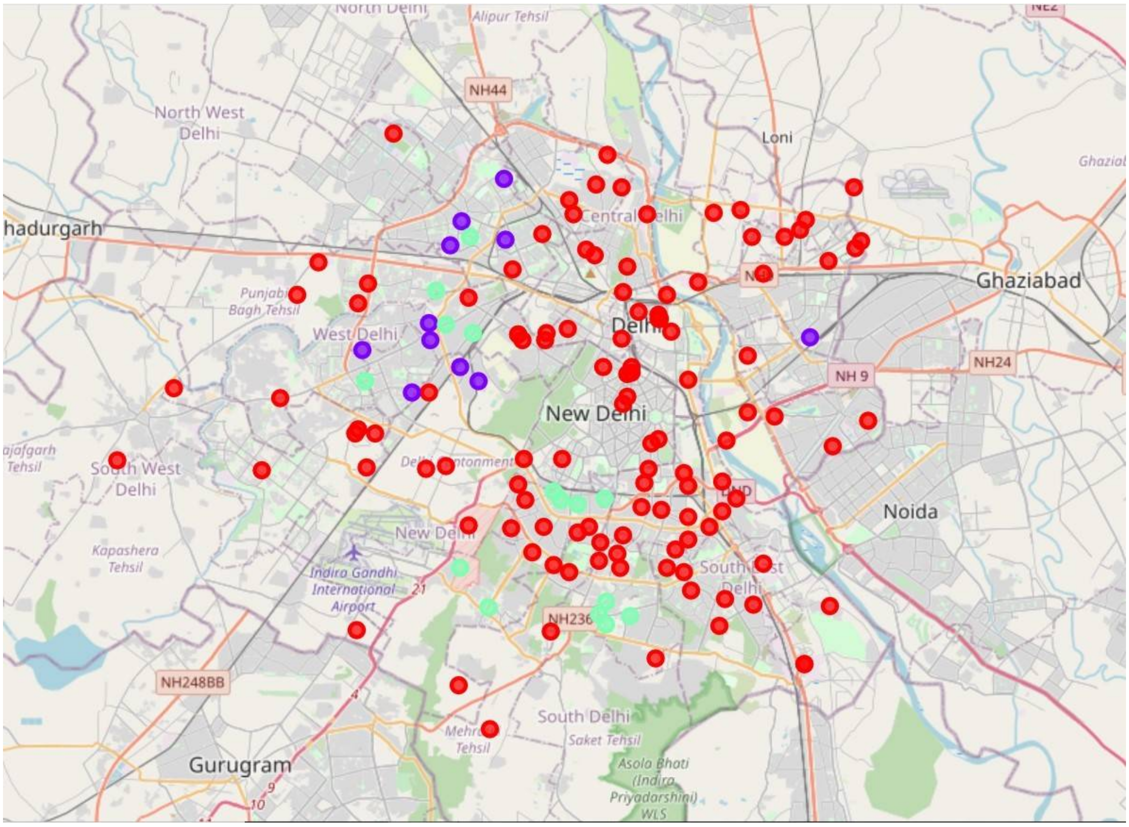

Fig.1: Map showing three clusters with concentration of multiplexes

B) STATISCIAL ANALYSIS OF CLUSTERS

- Considering the first cluster (Cluster 0) has very sparse & wide neighborhood data related to multiplexes and the result of k-means clustering shows centroids as very small, it is considered appropriate to drop Cluster 0 from further analysis and carry out statistical analysis on the other two clusters.

- The statistical plotting library of seaborn is then used on two clusters to analyze the results arrived in the analysis carried out while using k-means clustering. The same is depicted in the graphs as below:

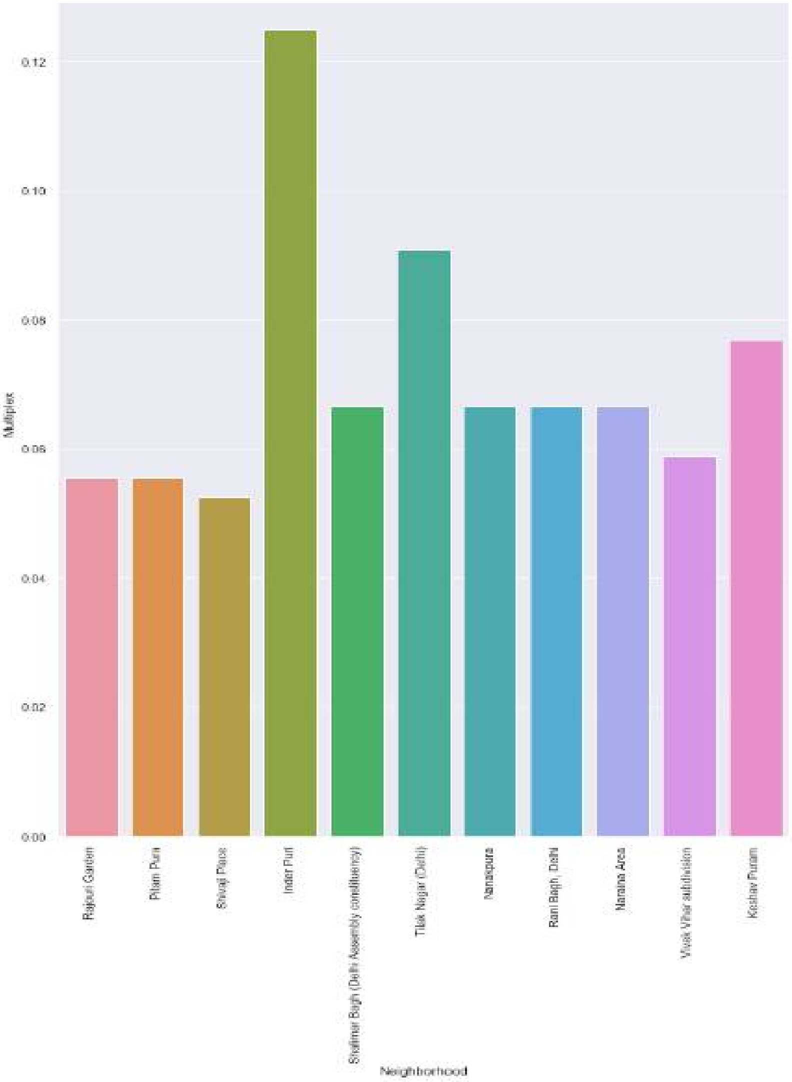

Fig.2: Bar chart for cluster 1 with high concentration of multiplexes

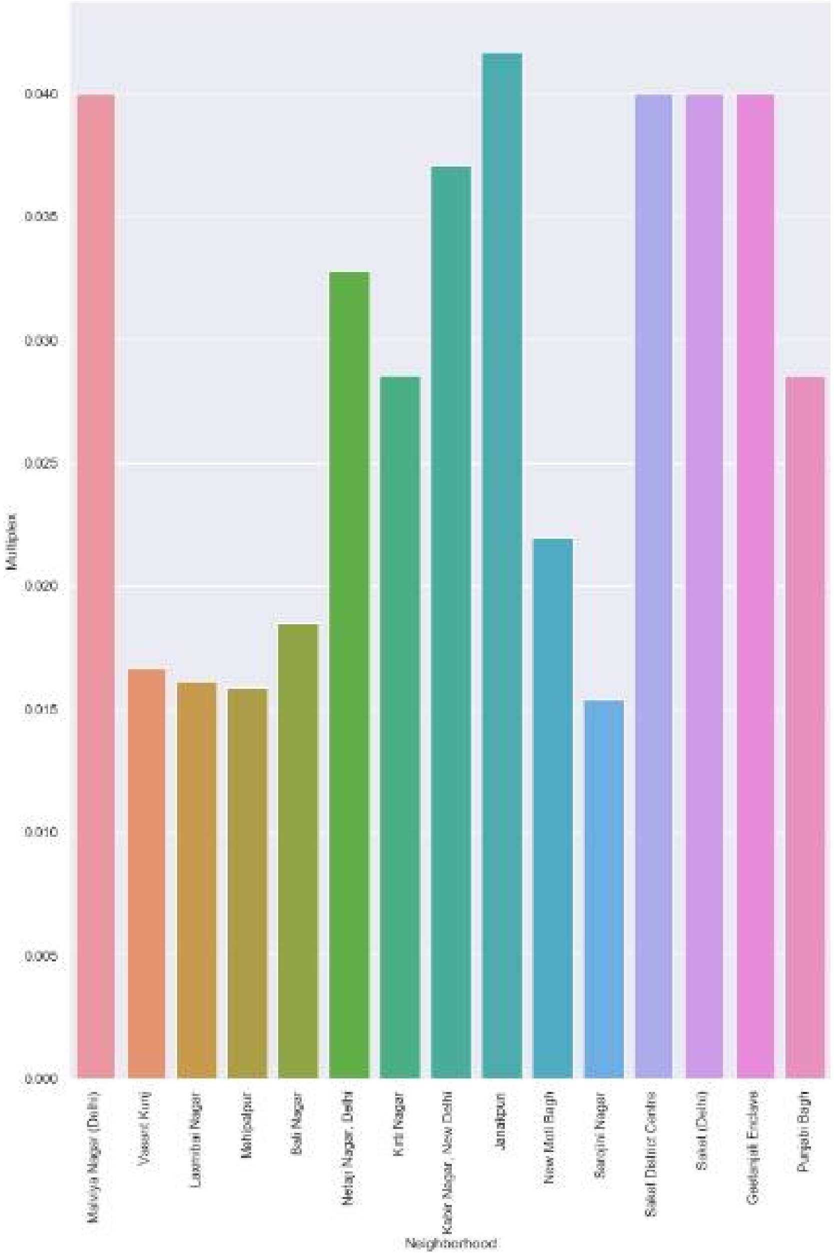

Fig.3: Bar chart for cluster 2 with moderate concentration of multiplexes

- The result of statistical analysis shows that almost all neighborhoods in Cluster 1 have similar and high concentration of multiplexes. Further, neighborhoods in Cluster 2 have moderate and varied concentration of multiplexes.

5. DISCUSSION

- Based on clustering and statistical analysis, the result shows that Cluster 1 has limited scope of developing a new multiplex considering high concentration of multiplexes.

- Accordingly, Cluster 2 is identified and shortlisted for further analysis. With this identification, it is found that 15 neighborhoods in Cluster 2 provide opportunity for opening multiplexes.

- As an illustrative case, one neighborhood is then explored to arrive at the final decision of opening a multiplex. In the instant case, the neighborhood chosen is 'Punjabi Bagh'.

- It can be seen from further analysis of the neighborhood 'Punjabi Bagh', that it has restaurants, café, pizza place, hotels, bars, metro stations etc. This clearly states that the location has lot of recreational activities other than just one multiplex in the vicinity.

- With moderate concentration of multiplex and high level of recreational activities in 'Punjabi Bagh' the location is ideal for developing profitable multiplexes.

6. CONCLUSION

- The Project recommends investors and property developers to use the findings made in the report to open new multiplexes across 15 neighborhoods in Cluster 2 that have limited competition.

- Further, investors and property developers can avoid Cluster 1 with high concentration of multiplexes which will result in unnecessary competition and low profit margins.

- The Project has been developed as a powerful data model with increase in accuracy as more and more data is loaded to the model.