Coursera Capstone

IBM Applied Data Science Capstone

Restaurant Location Recommendation Using Neighborhood Clustering – New Delhi

Dec 2020

Introduction

Delhi officially known as the National Capital Territory of Delhi (NCT), is a city and a union territory of India containing New Delhi, the capital of India. It is bordered by the state of Haryana on three sides and by Uttar Pradesh to the east. The NCT covers an area of 1,484 square kilometers (573 sq mi).According to the 2011 census, Delhi's city proper population was over 11 million, the second-highest in India after Mumbai, while the whole NCT's population was about 16.8 million. Delhi's urban area is now considered to extend beyond the NCT boundaries, and include the neighboring satellite cities of Ghaziabad, Faridabad, Gurgaon and Noida in an area called the National Capital Region (NCR) and had an estimated 2016 population of over 26 million people, making it the world's second-largest urban area according to the United Nations. Recent estimates of the metro economy of its urban area have ranked Delhi either the most or second-most productive metro area of India. Delhi is the second-wealthiest city in India after Mumbai and is home to 18 billionaires and 23,000 millionaires. Delhi ranks fifth among the Indian states and union territories in human development index. Delhi has the second-highest GDP per capita in India. Delhi is of great historical significance as an important commercial, transport, and cultural hub, as well as the political center of India.

Business Problem

In this project will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening an Indian Cuisine restaurant in Delhi, India. Finding a suitable location for restaurants in major cities like Delhi proves to be a daunting task. Various factors such as over-saturation or no demand ,for the type of restaurant that the customer wants to open, effect the success or failure of the restaurant. Hence, customers can bolster their decisions using the descriptive and predictive capabilities of data science.

We need to find locations (Neighborhood) that have a potentially unfulfilled demand for Indian Restaurant. Also, we need locations that have low competition and are not already crowded. We would also prefer location as close to popular city Neighborhood, assuming the first two conditions are met.

We will use our data science powers to generate a few most promising neighborhoods based on these criteria. Advantages of each area will then be clearly Expressed so that best possible final location can be chosen by stakeholders.

Data

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

- Delhi data containing the neighborhoods and boroughs.

- Latitude and longitude coordinates of those neighborhoods. This is required to plot the map and get the venue data.

- Venue data, particularly data related to restaurants. We are going to use this data to perform further analysis of the neighborhoods. Data Source and methods to extract them Delhi data containing the neighborhoods and boroughs will be obtained from the open data source:

https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi. After it, we will get the geographical coordinates of the neighborhoods (latitude and longitude) using Python Geocoder package. Finally, we will use Foursquare API to get the venue data for the neighborhoods defined at the previous step. Foursquare has one of the largest databases of 105+ million places and over 125,000 developers use this application. Foursquare API provides many categories of the venue data; we are particularly interested in the restaurant data to solve the business problem defined above. This project will require using of many data science skills, from web scrapping (open source dataset), working with API (Foursquare), data cleaning, data wrangling, to map visualization (Folium). In the next Methodology section, we will discuss and describe any exploratory data analysis that we did, any inferential statistical testing that we performed, and what machine learning techniques were used.

Results and Discussions

Our Analysis was done on over 129 neighborhoods, containing over 848 restaurants within 2km radius of every neighborhood. We segragated these neighborhoods on the basis of types and amounts of restaurants. Five clusters were obtained, each having a unique collection of restaurants. Since, we were focused on finding optimal neighborhoods for opening Indian restaurants, we selected cluster 2 and 3 which had the highest number of Indian restaurants. The above actions left us with the only those neighborhoods that had a shared characteristics of and that had a high demand for Indian restaurants.

Next, we plotted a heat map for analysing the density of restaurants in the remaining neighborhoods. This allowed us to select neighborhoods that had few or no Indian restaurants and were not overcrowded by

other kinds of restaurants. A total of 57 neighborhoods were left. After this, we found out the top three most popular neighborhoods(namely: Canaught Place, Hauz khas Village and Indira Gandhi International Airport), and the distance of every remaining neighborhoods from all three of them. Then, we extracted top 5 closest neighborhoods from each of three most popular neighborhoods mentioned above. Taking the union of the resulting three dataset we get 11 neighborhoods that satisfy all three conditons layed out in the business problem by the customer.

The neighborhoods recommendation obtained here are not completely accurate. This is due to the limitations in the dataset used in the project. Due to lack of cross referencing sources, we may have missed a few neighborhoods from our consideration. The foursquare API does not contain, or does not rely, a comprehensive dataset about the restaurants present in delhi. Surely, in a city like Delhi with a population of over 19 million, there are much more restaurants than 848.