Coursera Capstone Project

IBM Applied Data Science Capstone

Opening a New Shopping Mall in Allahabad, India

BY Avanendra Pratap Singh

July 2020

Introduction

For many shoppers visiting lavish Shopping mall is a great way to relax and enjoy themselves during weekends and holidays. A shopping centre is a collection of independent retail stores, services, and a parking area conceived, constructed, and maintained by a management firm as a unit. Shopping centers may also contain restaurants, banks, theatres, professional offices, service stations, and other establishments. For retailers shopping mall provides a great platform by hosting a large crowd of shoppers to market their services and goods. Shopping malls are also the one way destination for variety of shoppers. Property developers are also taking advantage of this trend to construct more such malls in order to cater their demand. As a result many new malls are coming up rapidly in the city especially in the economic heart of the city i.e. Civil Lines. Obviously opening a new shopping malls seems to be quite complicated than it seems. There are list of factors which influences it. One of the major reasons happens to be choosing the appropriate location for the shopping mall.

Objective of the Project

The objective of this Capstone Project is to analyze and determine the best locations in the city of Allahabad, India to open up a Shopping mall. I will be using data science methodology and machine learning techniques like data clustering. This Capstone project aims to provide answer to a question: If a property-developer is looking to open up a new shopping mall in the city of Allahabad, India, where would I recommend that they open it?

Target Audience for this project

This project is particularly useful for the property-developers and investors who are looking to open or invest in the shopping malls in the city of Allahabad, India. This project is timely as currently the whole nation is trying to recover from the serious blow of pandemic Covid-19 and many nations' economy has been crippled from the medical expenditures incurred treating the Corona positive Patients. Government of all nations is sincerely making efforts to revive the economy and in doing so is making efforts to encourage building of such things.

Data Used in the Project

To solve the following problem we need to gather the following data:

1. List of the neighborhoods in the city of Allahabad, India. This defines the scope of the project which will be confined to the city of Allahabad, India.
2. Latitude and Longitude coordinates of each neighborhood. This is required in order to plot the maps and choosing the appropriate venues.

3. Venue data especially related to the shopping malls. We will use this data to apply data clustering on the neighborhoods.

Gathering of data to be used

The Wikipedia page containing the list of neighborhoods in city is https://en.m.wikipedia.org/wiki/Category:Neighbourhoods_in_Allahabad. It contains a list of 42 elements. We will use web scrapping techniques to extract the data from the given page. We will be using python request and BeautifulSoup packages. Then we will get the geographical coordinates of each neighbor using python Geocoder which will help us collect latitudes and longitudes of the neighborhoods.

After that we will use FourSquare API to get the venue data for those neighborhoods. FourSqaure has the largest database of 105+ million and is used by 150,000 developers. FourSquare API will provide many categories of the venue data, we are particularly interested in the Shopping mall category. This project will use data science skills, like web scrapping (BeautifulSoup), API(FourSquare), data cleaning , data wrangling, Machine Learning(k-means clustering) and map visualization (Folium).

Procedure

Firstly we will need the list of the neighborhood in the city of Allahabad. Fortunately for us the needed list is available at the Wikipedia page

https://en.m.wikipedia.org/wiki/Category:Neighbourhoods_in_Allahabad. We will use web scrapping techniques like python request and BeautifulSoup packages to extract the list of the neighborhoods. Then we will use python Geocoder package to get latitude and longitude for each of the neighborhoods. Then we will populate the data into Pandas DataFrame and then visualize the map using Folium package.

Next we will use the FourSquare API to get the top 100 venues that are within the radius of 3000m. For this we will need to register in the FourSquare in order to obtain the secret key and ID. We then make API calls to the FourSquare passing the geographical coordinates of each neighborhood. FourSquare will return the venue data in the JSON Format of which we will extract the venue name, venue category, venue latitude and longitude. We can also check how many venues were returned for each neighborhood and how many unique categories can be curated from it. Then we will group neighborhoods and take the mean frequency of the occurrence of each venue category. We will filter the "Shopping mall" as the category for the neighborhood.

Lastly we will perform the K-means clustering on the venue data. We will divide the data into three clusters based on their frequency of occurring of "Shopping mall". This will help us identify which neighborhood has high and low number of concentration of Shopping Malls. Based on the occurrence of Shopping malls in the neighborhood we can decide which neighborhoods are suitable for opening a new Shopping Mall.