Coursera Capstone Project: The Battle of Neighbourhoods

By: Satyam Dutta

Indian Institute of Information Technology, Surat

May 27, 2020

INTRODUCTION

The importance of a shopping mall, as we all know, is mainly for people to get out of the house for a while and do something entertaining. Shopping malls can provide the best shopping experiences such as social gatherings, entertainment, performances, product launches, promotions and festivals. The events list at shopping malls goes on and on for any, particular, person to be entertained for a number of hours.

Any shopping mall can be a great place to hang out with friends, eat, shop, and more. We can go to all your favourite stores and I believe that parents enjoy it just as much as kids.

Shopping malls also tend to be a major tourist attraction. The malls can be more convenient and helpful, for a tourist, to have one central location to do all their shopping, rather than to have to drive many miles just to buy different types of products for their personal needs.

Shopping physically at a mall, compared to shopping online is very different. Going to a mall is always more convenient as for example, shopping for clothing can be a hassle online because we aren't able to try on the cloths and find something that fits for sure. Whereas, shopping at a mall allows us to try anything, we're interested in, on to make sure it fits before you invest in it.

As a result, there are many shopping malls in the Surat city and many more are being built. Opening shopping malls also allow the property developers to earn consistent rental income. But opening a new mall requires serious considerations and it is a lot more complicated than it seems; especially the locations of shopping mall is one of the most important decisions that will determine whether the mall will be a success or failure.

Business Problem

The objective of this project is to analyse and select the best locations in Surat, Gujarat to open a new shopping mall. Using the various aspects of Data Science like visualisation and Machine Learning techniques like clustering, this project aims to provide answer to one of the prime questions, i.e., 'What should be the recommended place to open a new shopping mall in a developed city like Surat?'

Target Audience

This project will provide useful insight to the business developers and investors, who are looking forward to open or invest into the new shopping malls in the capital city of Gujarat, i.e., Surat.

Sources of Data & Tools required

1. *List of neighbourhoods in Surat:*

The Wikipedia page 'https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Surat' contains a list of neighbourhoods in Surat, with a total of 76 neighbourhoods. Web scraping techniques will be used to extract this data from the source page.

2. *GPS coordinates of the neighbourhoods:*

Geographical coordinates (latitude, longitude) of the neighbourhoods will be obtained, using Python Geocoder package, which will help us to plot the map and obtain venue data.

3. *Foursquare API:*

After the above steps, Foursquare API will be used to extract data of the neighbourhoods. The data obtained will be used to cluster the neighbourhoods. A machine learning model (k-means clustering) can be used to do the same and recommend the best place to construct new malls.

Methodology

Firstly, we obtain the list of neighbourhoods in Surat. The same is available at the Wikipedia page 'https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Surat'. Then we perform web scraping using BeautifulSoup4 and other Python modules to extract the required data from rest of the webpage and convert it into a dataframe. Further, we will add geographical coordinates of the neighbourhoods using the Python Geocoder package in order to be able to use Foursquare API to gain detailed information about the neighbourhoods we need. After gathering all the neighbourhoods' and their location data, we will populate the data into a pandas dataframe and visualize it in a map using Folium (a python package).

Next, we use Foursquare API to obtain top 100 venues that are within the radius of 3000 metres. Assuming that we already have the credentials required to use the API, we will construct a function to make an API call, passing the geographical coordinates for all the neighbourhoods using Python loop. The function returns the venue data in JSON format and we extract only the venue name, category, latitude and longitude from it.

We will then analyse the data by checking the number of venues returned for each neighbourhood, number of unique categories and also the mean and frequency of occurrence of each venue category. We then filter 'Shopping Mall' as venue category as we need to perform analysis based on shopping malls.

Lastly, we will perform data clustering using a machine learning model, K-means clustering. This algorithm identifies K number of centroids, then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. We will cluster the dataset into 3 clusters, based on their frequency of occurrence for 'Shopping Mall'. The observations and obtained results will help us to know which neighbourhoods have what concentration of shopping malls. Based on the result, we will be able to come up with an answer for the question put up in the Business Problem that which neighbourhoods are the suitable locations to open a new shopping mall.

Result

The results of the K-means clustering tell us that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for 'Shopping Mall' and conclude the following:

- Cluster 0: Neighbourhoods with moderate number of shopping malls.
- Cluster 1: Neighbourhoods with low number of shopping malls.
- Cluster 2: Neighbourhoods with large number of shopping malls.

Discussion

We observe that most of the shopping malls are concentrated in the central area of Surat, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to totally no shopping malls in the neighbourhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 2 are likely to be suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, this also shows that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still having very few shopping malls. Therefore, this project recommends business developers and investors to capitalize on these findings to open new shopping malls in neighbourhoods in cluster 1 with little to no competition and to avoid neighbourhoods in cluster 2 which already have high concentration of shopping malls and are suffering from intense competition.

Conclusion

In this project, we went through every step that a data analyst usually undergoes, starting from identifying business problem, collecting and extracting all the relevant and needed data, transforming and visualising raw data into informative datasets and maps respectively, using various tools and finally applying a machine learning model to get the answer of our business problem, hence, providing recommendations to the concerned stakeholders. The findings of this project will help the concerned stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded locations to construct or invest into a new shopping mall.