Coursera Capstone

IBM Applied Data Science Capstone

Opening a new multiplex in Kolkata, India

By: Ritesh Manna
Date:February 2020

Background

A multiplex is a movie theater complex with multiple screens within a single complex. They are usually housed in a specially designed building.

For many people visiting a multiplex is a great way to relax and enjoy themselves during the time of holidays and weekends. It is one of the few destinations where one can watch newly released movies. For retailers, large crowds in multiplexes provides a good distribution channel to market their products and services.

This project is particularly useful for property developers who are looking to open or invest in new multiplexes around the city of Kolkata. Location is one of the most important decisions that will determine whether the multiplex will be a success or a failure.

Business Problem

The aim of this project is to analyse and select the best locations in the city of Kolkata to open a new multiplex. Using data science methodologies and Machine Learning techniques like clustering, this projects tries to answer the question - In the city of Kolkata, India if a property developers is looking for a location to open a new multiplex what would you recommend?

Data

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

- List of neighbourhoods in Kolkata, India.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to multiplexes. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them:

This url (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Kolkata) contains a list of neighbourhoods in Kolkata, with a total of 199 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use the Foursquare API to get the venue data for those neighbourhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Multiplex category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping, working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

The list of neighbourhoods in the city of Kolkata is available on the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Kolkata). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that will allow us to convert addresses into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas dataFrame and then visualize the neighbourhoods in a map using the Folium package. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 1 km. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Multiplex" data, we will filter the "Multiplex" as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Multiplex". The results will allow us to identify which neighbourhoods have higher concentration of multiplexes while which neighbourhoods have fewer number of multiplexes. Based on the occurrence of multiplexes in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new multiplexes.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Multiplex":

- Cluster 0: Neighbourhoods with a low number of multiplexes.
- Cluster 1: Neighbourhoods with a high number of multiplexes
- Cluster 2: Neighbourhoods with a moderate count of multiplexes.

Discussion

As observations noted from the map in the Results section most of the multiplexes are in cluster 1 while cluster 2 consists of moderate number of multiplexes. On the other hand, cluster 0 has a very low number to no multiplexes in the neighbourhoods. This represents a great opportunity and high potential areas to open new multiplexes as there is very little to no competition from existing malls. Meanwhile, multiplexes in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of multiplexes. Therefore, this project recommends property developers to capitalize on these findings to open new multiplexes in neighbourhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new multiplexes in neighbourhoods in cluster 2 with moderate competition. Lastly, property developers are suggested to avoid neighbourhoods in cluster 1 which already have a high concentration of multiplexes.

.Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new multiplex. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new multiplex.