COURSERA CAPSTONE

IBM Applied Data Science Capstone

Opening a new Movie Theatre in the city of Hyderabad

By: Roshan

January 2020

Introduction

Movies are a currently a great source of entertainment. It’s one of the best ways to come out of the modern stressful life. Considering the city of Hyderabad, there are many movie loving people and is also a great business. People love to watch all sorts of movies, from regional to worldwide from comedy to science fiction. Thus movie theatres are among the best businesses. There are many established theatres running successfully in the city. As the population in the city is on a rise, new theatres will have a good scope of development and can attract people. So one can find new theatres making use of this opportunity to earn money by meeting the demand. But a new one can’t simply be set up. Various factors may influence the setting up new theatres and one of the main factor is business competition. Lesser the competition more the success. The aim of this project is to find such a suitable location by considering the location of all the other theatres and finding a place where there are not many.

Business Problem

The objective of this capstone project to is to analyse and select the best location in the city of Hyderabad to open a new movie Theatre. Using data science methodology and machine learning techniques, the project aims at answering the question, what is the best place for a movie theatre businessman in the city of Hyderabad to open a new theatre, in terms of less business competition?

Target audience of this project

This project can be very useful for movie distributors or theatrical exhibitors who are looking to invest in or open a new movie theatre in the city of Hyderabad. As there are already a large number of theatres in the city it can be difficult to decide the right place to build a new one. There may be many other factors which can influence the building of a new one like population, incentives, etc… but one of the main factor is business competition, which directly effects the success of the theatre. Thus this project can help the movie distributors to filter out the areas having a low occupancy of theatres, and they can further think based on the listed out regions if necessary.

Data Required

Data required to the solve the problem is:

- Neighbourhood data of the city of Hyderabad, this includes all the place names. This data can be brought from the Wikipedia page here.
- Latitude and longitude data of the list of places in the city. The python package Geocoder can be used for this.
- Last and the most important is the venue data of each place. This can be brought by using the Foursquare API.

Ways to extract the data:

- Firstly, for the neighbourhood data, the source is a Wikipedia page hence the data needs to be scraped, and the python beautifulsoup package and the requests package will be used for that purpose.
- The python Geocoder package can be of great use to get the latitude and longitude data of different places. Another python package, Folium is used for map illustrations.
- Venue details can be brought by making an API call to the foursquare data. For this an account is required with a client id and client secret, and the url for the call includes the same.
- The python package sklearn is used to implement machine learning algorithms in this case the KMean Clustering algorithm.

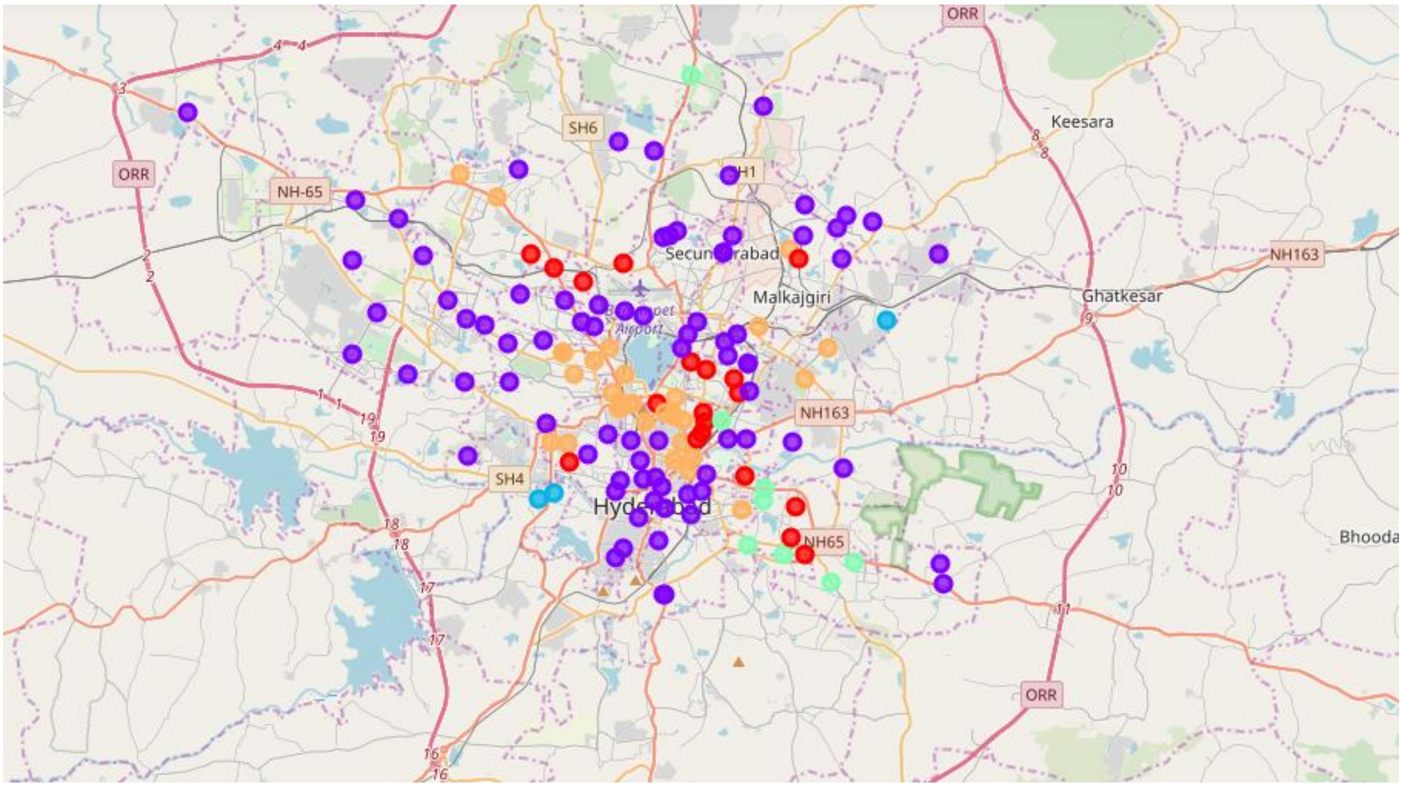
Methodology

- The main idea to solve the problem is to consider only the places where there are very less or no theatres. This ensures there is no competition to the new theatre.
- Further we can also check if the particular place has any other neighbourhood with a certain number of theatres. We can neglect such places too. Thus we can select places where there are no theatres both within and around the place, thus ensuring very less business competition.
- This being the main idea, the next step is data collection. The data required is neighbourhood information, which is got from Wikipedia, the venues information of the particular place which is got from the Foursquare API.
- The neighbourhood information from Wikipedia is got by web scraping using python beautifulsoup module. Then we also need the latitude and longitude information of each place for which geopy module is used.

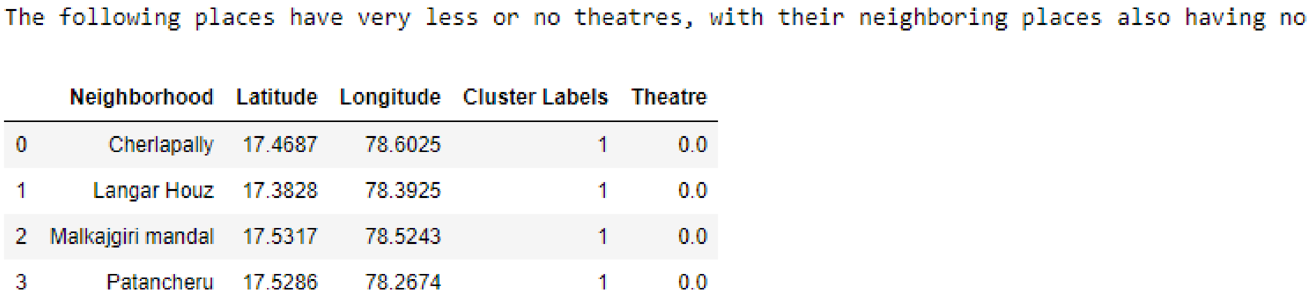
- Together with scrapping and the latitude longitude a hyd_neighbourhood dataframe is formed.
- These neighbourhoods are plotted on the map using the python Folium module to visualize the location.
- Next we need to collect the venue details of each place which can be got by using the Foursquare API. We need to mention the credentials to generate the URL.
- Thus a venues dataframe is formed with the details of the venues, their latitudes longitudes and the venue categories in each neighbourhood.
- As we also want to compare the neighbourhoods with other neighbourhoods close by, we create another neighbours dataframe which holds a list of neighbouring places for each neighbourhood.
- To make this dataframe we need to set a threshold, i.e. a fixed distance between to places to consider them as neighbours. If the distance is lesser than this threshold, only then will they be added to the list.
- Now out of all the venue categories, we need to filter out the theatres categories, and create the theatres dataframe, we should also rename the categories alike.
- We can also plot them on the map to visualize the theatre location.
- Now we have to generate the hot encoding of the venues dataframe to be used in the KMeans Clustering algorithm. But we need to consider only the theatres column in the encoding for the algorithm.
- Now applying the KMeans and make 5 clusters. And visualize by plotting on a map.
- Consider the cluster which has very less or no theatres. Of these theatres we also need to compare with the neighbourhoods dataframe to check if the neighbouring places belong to a cluster with less or high theatres. If less, then the places can be considered else no.
- Thus we generate places with no or very less number or no theatres, even in their surroundings, and these can be considered for a new theatre construction.

Results

The results from the KMeans divides the places into 5 categories based on the number of theatres. Cluster 1 having the least number of theatres.



Here the purple colour is the cluster 1. Now after checking with their neighbouring places we finally filter out the places with least number of theatres. After checking the neighbours the resulting places are as follows:



```
The following places have very less or no theatres, with their neighboring places also having no
```

	Neighborhood	Latitude	Longitude	Cluster Labels	Theatre
0	Cherlapally	17.4687	78.6025	1	0.0
1	Langar Houz	17.3828	78.3925	1	0.0
2	Malkajgiri mandal	17.5317	78.5243	1	0.0
3	Patancheru	17.5286	78.2674	1	0.0

Note: For this outcome threshold is fixed as 4Km to consider as neighbour.

Discussion

- Most of the theatres are located in the centre part of the city.
- The threshold can be increased or reduced based on the requirement.
- This project only shows places where there is very less business competition. While applying in real life we need to take in the population factor too, which is very important.
- Therefore, while considering the neighbouring places the threshold can be changed as per this requirement too.

- We can also work with other level of clusters, not just cluster 1, if we can manage a certain level of competition and get results accordingly.

Conclusion

Thus using the neighbourhood data from Wikipedia and the venue data from the Foursquare API for the city of Hyderabad, we could solve a business problem of finding a right place to open a new movie theatre, where we can be assured of less business competition.

However, population is another important factor which is not considered in this project to get the results. This only tells the places with less competition, population can be added and further analysis can be made.

THANK YOU!