Capstone Project – IBM

The Battle of Neighbourhoods

Opening of New Shopping complex in San Diego, California

Created By- Varun Gupta

June 2020

Introduction –

The IBM data science professional certificate has taken me through an amazing journey and towards the end creating our own project has helped me a lot. The problem statement is that a businessman wants to open a shopping mall in San Diego, California but he doesn't want to open it in a place where there is already a though competition and tons of shopping complex and nor he wants to open it in a place where there is no future scope for this kind of activity. Shopping complex are one stop destination for every kind of stuff whether it is clothing, footwear, theatre, or food corner. Shopping mall caters all these needs. This is one of those projects that need a large capital to build and no investor would want this to fail therefore it is quite important to predict whether this is the best place or not. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem –

The naïve approach to the problem should be opening the complex in area where there are already many successful malls, but this would defy the purpose of the problem. It is not necessary that a place having shopping areas is the only fit place to target by the businessman. The main aim is to determine which areas are commercially better than the rest of the neighbourhoods since this ensures that these areas are for business and then filtering among them which areas having low/moderate/high number of complex.

Target Audience –

This analysis is quite useful to the property developers and investors who are looking for an opportunity to invest in San Diego. This project is important because this city of California have been growing lately and many investors are seeking an opportunity to invest and flourish in the market. Del Mar Heights is a place where there was almost nothing at some time but after the establishment of the city mall, the place has grown tons and has noticed a hike in the property price.

Data –

The main components required are the following data –

- The neighbourhoods' data which is largely extracted from Wikipedia sources.
- The co-ordinates which means the latitudes and longitudes of that place.
- The venues detail about each neighbourhood in our dataset.

Sources and methods to extract the data –

• The Wikipedia page (https://en.wikipedia.org/wiki/List_of_communities_and_neighborhoods_of_San_Diego) contains a list of neighbourhoods in San Diego, with a total of 120 different area. The technique used to extract data is with help of the python package Beautiful Soup.

• The latitude and longitude of a place are found with the help of the geocoder package in python. With this we were able to extract co-ordinates for every neighbourhood in dataset to feed in the foursquare API.

• After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category to help us to solve the business problem put forward.

This project is all in one in which we ourselves define the problem statement and collect the data, clean those and then after a bit of analysis we apply machine learning techniques that are efficient in figuring out patterns in our data.

Methodology –

• The main element is the data collection. The neighbourhood's data is freely available on the Wikipedia page mentioned before. We construct a soup object from the web scraper package beautiful soup package and extract the data. This is the list of locations and now we use the geocoder package in-order to find out the latitude and longitude of these places. This data can be visualized using the Folium library with markers on the locations.

• Now we will use the foursquare API to obtain the top venues near a place in our dataset. 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.

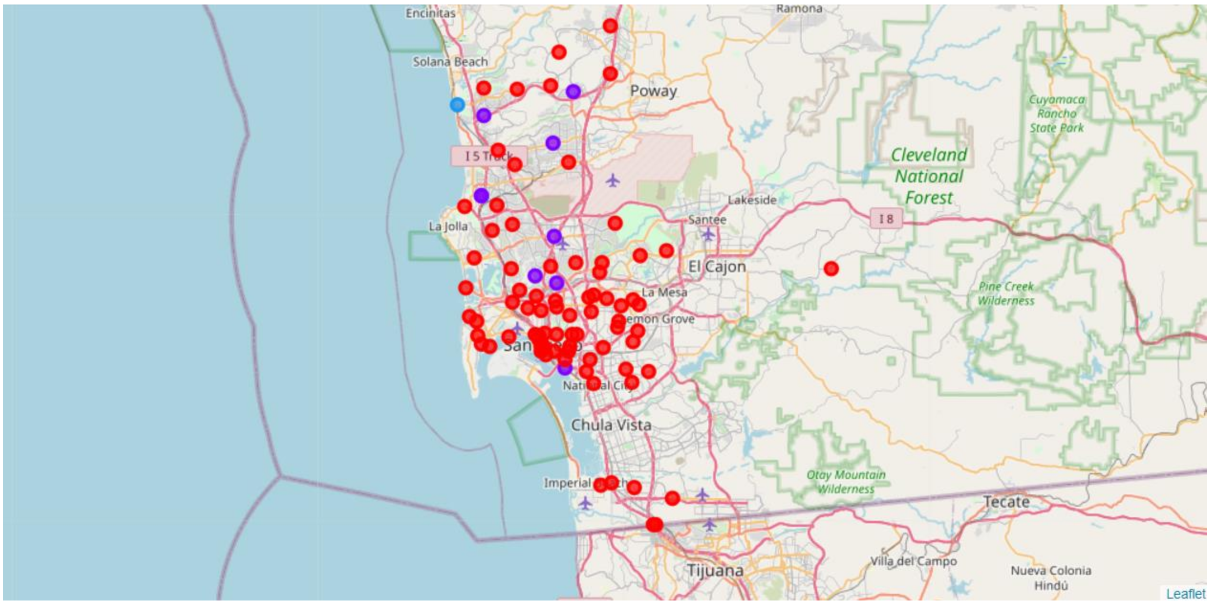
• The final stage is the unsupervised learning in which we first identify the cluster where there is high commercial activities and then we find the cluster where the shopping malls are situated and which places can be new hotspots for shopping complex. We will use the K-means clustering algorithm and test it out on various values of K.

Results –

The result from the k-means clustering first recognizes the places where there are any commercial activities. The second cluster recognizes the density of the shopping complex.

- Cluster 0: Neighbourhoods with moderate number of shopping malls
- Cluster 1: Neighbourhoods with low number to no existence of shopping malls
- Cluster 2: Neighbourhoods with high concentration of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in blue colour.



Analysis and Conclusion –

Shopping Label 0 - indicates that the area has very less proportion of shopping complex and has other features.

Shopping Label 1 - indicates that the area has moderate amount of shopping complex and that they are flourishing

Shopping Label 2 - indicates that the area has good shopping complex and it is known for them. This tells us the areas which are commercial and where you might find a place for your business. The labels have a deep meaning which depends upon the consumer. So, the consumer might be afraid of competition or he may find some cheap land in cluster 0 of shopping that is where there are low number of complex. This decision depends upon the mindset you have and the assets. An apt choice would be built in an area where there is some shopping complex which are there to tell us that the area recognizes them and that they are flourishing. The third kind is for people who are more confident or risk takers because they already know that the cluster 2 is famous for shopping and if we are able to build something extraordinary then we might end up stealing away a lot of consumers which have been built by the other complex. This area is sensitive to competition and ups and down.

Future Aspects –

One thing that I wanted to incorporate was that if somehow I could get the land pricing of the areas then that would also have been a contributing factor in determining the area where there could have been a possibility of successful shopping complex.