Final Report

Capstone Project - The Battle of Neighbourhoods (Part-2)

Introduction:

Opening a restaurant is all about location, location, location. However, not every Restaurant is suitable for every location, and vice versa. It comes down to a combination of restaurant style, target audience, your competitors. If you can define your restaurant type and identify your target demographic and its most populated areas, you'll be well on your way to choosing a restaurant location that sets your business up for success. For my assignment I have taken a business case of opening an Indian restaurant in Saudi Arabia. And for this my biggest problem or challenge is to find a best suitable location where I can have highest population of east asians, more working-class people or more earnings and more affordable prices for rent or for land or building purchase and finally less competitors.

Anyone who wants to get into the restaurant business and wants help in finding the best location using Data Science and Machine Learning algorithms will be interested in this project report.

Data:

For our restaurant problem, we will focus on the Dammam of Saudi Arabia and work on getting the data from all the Saudi Arabia. To solve our problem of finding a best location to start an Indian restaurant in Saudi Arabia, we need datasets based on various parameters such as:

1. Population of target audience in all the Dammam of Saudi Arabia based on their ethnicity
2. Earnings data of the working class living in the target location.
3. We need the data about the required business floorspace and rateable value statistics of each Neighborhood.
4. Considering the competitors factor, we also need the data of existing Licensed Restaurants in each Neighborhood.

All the above required information is available at **Foursquare and geopy data.** Centres of candidate areas will be generated algorithmically and approximate addresses of centres of those areas will be obtained using Google Maps API reverse geocoding.

Methodology:

First we need to get the geo-coordinates of the borough and the geo-coordinates of the Neighborhood of the borough from the web.

To read data from these URLs, I have used the **requests**, **urllib** and **BeautifulSoup** libraries of python.

After I have the geo-coordinates information of the neighbourhoods, I need the other data such as the venues or places of the neighbourhoods, the venue categories, working hours and so on. All this data is called Location data, and to get this data I need a reliable and efficient location data providers and hence I am using Foursquare as the data provider. I have used the Foursquare API to explore the Neighborhood in Dammam city. I have also used the **Explore** function to get the most common venue categories in each Neighborhood and then use this feature to group the neighbourhoods into clusters. To cluster the neighbourhoods I am using **K-means Clustering** algorithm.

Geopy module and **Nominatim** library is used to convert a given address into the latitude and longitude values.

To visualize the neighbourhoods, the library **Folium** is used, to display the maps of London, with the boroughs super imposed on it and to display the map of borough with the neighbourhoods superimposed on it.

A python function **getNearbyVenues()** is created , to give the venue details like venue name, venue latitude, venue longitude, venue category along with neighbourhood name, latitude and longitude for each neighbourhood.

After the venue data for each neighbourhood of the Saudi Arabia is generated , **One-Hot encoding** is applied on the venue category data, so that the analysis of the data will be easy in grouping the neighbourhoods based on the frequency of occurrence of each venue category.

Once the neighbourhoods are grouped based on the frequency of occurrence of the venue category, the top 10 venues of each neighbourhood are displayed as a dataframe.

After all the above data exploration and analysis and top 10 venues of each neighbourhood are identified, the K-means Clustering algorithm is applied to the resultant dataframe to segment the data into 5 Clusters and all these 5 clusters are visualised in a map using the Folium library and finally the 5 clusters are examined to determine the discriminating venue categories that distinguish each cluster.

Results:

In the Segmenting and Clustering section, the neighbourhoods of Dammam borough are explored, and the top 10 venues of each neighbourhood are listed. The neighbourhoods are clustered into 5 clusters using K-means algorithm and their most common neighbourhoods are identified. After applying the K-means algorithm the 2 neighbourhoods Macca and Medina are identified as best locations to open or start an Indian restaurant.

Discussion:

My observation after doing this analysis is the model we used could have given better results, if we had huge data to train and test our model. In spite of that this model gives us a better insight for our problem and also help us to gain better results. From the clustering results our problem finds a better solution of identifying the best location for the Indian restaurant. We could explore all the neighbourhoods of the borough and could list the most common venues based on their frequency of occurrence. From these results I can strongly recommend the Macca, Medina and few other neighbourhoods as a preferred location for our restaurant, as these areas have the restaurant venue as the most common venue.

Conclusion:

There is always room for improvement and hence the above solution I have provided can also be improved and the machine learning models can be trained and tested for best results depending upon the data we have.