FINAL REPORT

Part 2: Capstone Project – Battle of Neighborhoods

Introduction:

Opening a restaurant is all about 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 a Pizza restaurant in Bronx, New York. And for this my biggest problem or challenge is to find a best suitable location where I can have highest population who visit pizza stores more often, and less competitors.

Anyone who wants to get into the Pizza 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 **Neighborhoods of the Bronx** and work on getting the data from all the Neighborhoods. There are around 52 Bronx Neighborhoods with a population of around 1,471,160. To solve our problem of finding a best location to start a Pizza restaurant in the Bronx borough, we need to datasets based on various parameters such as:

- Population of target
- Most frequency visited places
- Competitors – Existing restaurants

All the above required information is available at **New York Dataset (IBM Box)**, which is a free and open data-sharing portal where anyone can access data relating to the city. The data is available in JSON format, which we can download and can use as-is for solving our problem.

Methodology:

To work on the solution, I have used Pandas library **Json_normalize** to read the data in JSON format and convert into pandas dataframe. Extensive data exploration analysis is done, where lot of data is cleaned and presented in a suitable format.

First we need to get the geo-coordinates of the borough and the geo-coordinates of the neighborhoods of the borough from the web. I have used the **New York Dataset** mentioned in the data section to get this data.

After I have the geo-coordinates information of the borough and its neighborhoods, I need the other data such as the venues or places of the neighborhoods, the venue categories, and so on. All this data is called Location data, and to get this data I need 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 neighborhoods in New York 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 neighborhoods into clusters. To cluster the neighborhoods 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 neighborhoods, the library **Folium** is used, to display the map of Bronx borough, with the neighborhoods super imposed on it.

A python function **getNearbyVenues()** is created, to give the venue details like venue name, venue latitude, venue longitude, venue category along with neighborhood name, latitude and longitude for each neighborhood. After the venue data for each neighborhood of the Bronx borough 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 neighborhoods based on the frequency of occurrence of each venue category. Once the neighborhoods are grouped based on the frequency of occurrence of the venue category, the top 3 venues of each neighborhood are displayed as a dataframe.

After all the above data exploration and analysis and top 3 venues of each neighborhood 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 visualized 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, the neighborhoods of Bronx borough are explored, and the top 3 venues of each neighborhood are listed. The neighborhoods are clustered into 5 clusters using K-means algorithm and their most common neighborhoods are identified. After applying the K-means algorithm the 16 neighborhoods listed: Norwood, Morris Park, North Riverdale, Castle Hill, Kingsbridge, Pelham Parkway, University Heights, Fordham, East Tremont, High Bridge, Melrose, Mott, Haven, Hunts Point, Eden Wald, Allerton, and Kingsbridge Heights are identified as best locations to open or start a Pizza restaurant out of 52 neighborhoods.

Discussion:

From my observation of the clustering results our problem finds a better solution of identifying the best location for the Pizza restaurant. We could explore all the neighborhoods of the borough and could list the most common venues based on their frequency of occurrence. From these results I can strongly recommend the Norwood, Morris Park, and few other neighborhoods 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 using other efficient Machine Learning algorithms.