**Capstone Project – Battle of Neighborhoods**

**Problem Description**

A chain of international coffee shops CuppaCut (fictitious company) based in London wants to open its first outlet in downtown Toronto. As part of their market research, they want to identify strategic locations for it store which would ensure footfall, good word of mouth publicity and revenue. Some of the considerations for the choice would be proximity to commercial buildings, competitors around the area( direct and indirect), accessibility by public transit.

**Possible Solution**

A practical, cost-effective solution is to leveraging location intelligence to identify groups of neighborhoods that are similar in characteristics. For example, a neighborhood with concentration of Italian restaurants can be identified. Once such groups are obtained it is easy to identify the suitable location of CuppaCut. Location information can be obtained from a location data provider such as Foursquare. Foursquare provides easy to use API’s to fetch location data for a neighborhood such as nearby venues, trending places, user reviews, etc. After the data is explored and cleaned K-Means clustering, an unsupervised machine learning algorithm, is applied to segment neighborhoods based on concentration of venues . The technique will give well separated segments which will help the coffee business pick the best location.

**Data Description and Preparation**

Information regarding postal code, boroughs and neighborhoods for Toronto was extracted from <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> using HTML parsing library Beautiful Soup. The text data was cleaned and fit into a dataframe with 3 columns – PostalCode, Borough and Neighborhood. A smaller dataset focusing on Toronto was filtered from this data which contained the string ‘Toronto’ in the Borough field. Then the latitude and longitude information for each PostalCode was obtained using pgeocode library and added to the dataframe.

Text

Description automatically generated

*Fig. A snapshot of the initial data frame*

Using a custom function which invokes FourSqaure API , a new dataframe containing information (name and category) for all venues within 500m of each neighborhood was generated and added to the original data frame. The venue categories were encoded and the frequency of venues for each neighborhood was calculated. Using this data a data frame containing neighborhood and top 10 most common venues was built.

Graphical user interface

Description automatically generated

*Fig. A snapshot of neighborhoods and nearby venues*

**Modelling**

The machine learning technique used for this problem is K-Means clustering which is an unsupervised algorithm which divides the data into K non-overlapping groups based on the proximity of each observation to the centroids of each group. The first step is to select a value for K that yields non overlapping clusters. Using elbow method which is based on the principle of minimizing the withing group distance or error , 4 was chosen as the value of K. Then using kmeans method of Scikit Learn library a model is fit on the input data frame. Cluster labels for each observation were then extracted from the model and appended to each observation. Having this information each cluster was then analyzed to study neighborhoods and the common venues surrounding them.

Chart, line chart

Description automatically generated

*Fig. Elbow method to find the optimal value of k*

**Analysis**

Out of the 3 clusters , cluster 0 has 3 neighborhoods, cluster 1 has 1 neighborhood and cluster 2 has the remaining neighborhoods. After closely analyzing cluster 2 it is evident that this cluster has neighborhoods with a high concentration of coffee shops , restaurants , bakeries, and other beverage shops which would mean high competition to Cuppacut if it were to choose one of the neighborhoods in this cluster. Whereas clusters 0 and 1 have neighborhoods with lesser number of food business which means lower competition of the coffee chain. But other factors including proximity to public transit, commercial spaces need to be taken into while selecting one of the neighborhoods from these clusters.

Map

Description automatically generated

*Fig. Snapshot of map overlaid with neighborhood clusters*

**Summary & Conclusion**

To help Cuppacut with its business decision first, basic location data including boroughs and neighborhoods for Toronto was extracted. Then more detailed data regarding venues for each neighborhood was extracted and the counts of venue types for each neighborhoods were calculated. This data was then grouped using clustering to find similar neighborhoods in terms of venue types. On further analysis it was found that ‘High Park’ , ‘Roselawn’, ‘Lawrence Park’ and ‘Danforth East’ are the best target neighborhoods for the coffee business. However to make a sound holistic decision these recommendations need to be complemented with further research.