**IBM Data Science Certificate:**

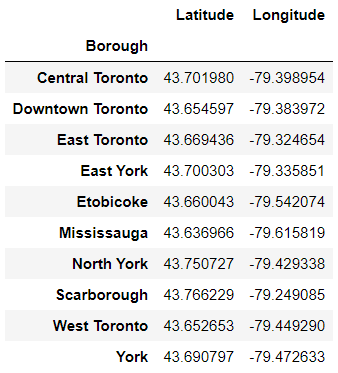
**Capstone Project-The Battle of Neighborhoods**

Introduction:

Location is vital to the success of a business. An entrepreneurship’s popularity will vary greatly on basis of the location in which it establishes its operation. The objective of this project was to create a localized model that tries to predict what type of venue would be popular based on its geographical coordinates, information that could be useful for entrepreneurs or local governments.

Data:

For this project two sources of data were used: geographical coordinates for the different neighborhoods in the test city Toronto and the geographical coordinates of the currently most popular venues in the city’s boroughs.

 For the neighborhood data, an archived Wikipedia entry for Canada's postal codes (available in https://en.wikipedia.org/w/index.php?title=List\_of\_postal\_codes\_of\_Canada:\_M&oldid=945633050.) was used in conjunction with data from the Geocoder Python package(the data was available in its entirety at http://cocl.us/Geospatial\_data, and this source was prefered in order to reduce the number of data fetching in the code)

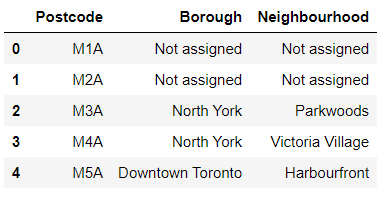
For the venue data, Foursquared API was used to source 200 venues in each borough with data available in 24/03/2020

The neighborhood data was used to build a table containing a reference coordinate for each borough.

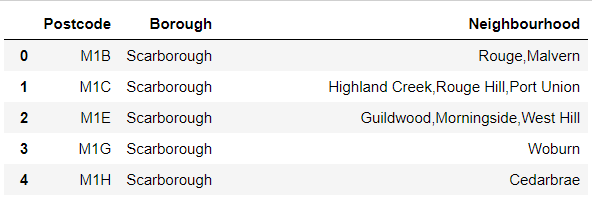
These locations were used to run queries in the Foursquared API to look for the most popular venues in the vicinity for each borough

The venue data was then processed and fed to a prediction model.

Methodology:

 First, the data for the postal codes for the different neighborhoods of Toronto was loaded into a pandas dataframe.

The data was then preprocessed:

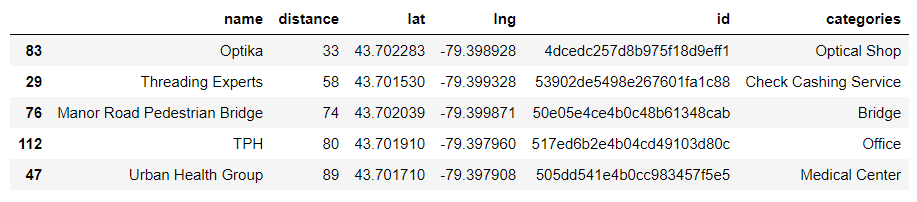
1. removed any entry without a borough assigned;
2. set any entry without a neighborhood assigned to have it set to its borough;
3. grouped neighborhoods by postcode.

After obtaining this first table the data for the coordinates for each postal code were loaded into a table, which had its column names edited to allow for merging the data of the two tables.

By grouping the different neighborhoods by boroughs and averaging the values for the respective neighboorhoods' latitude and longitude the coordinates to use for each of the 10 boroughs was determined.

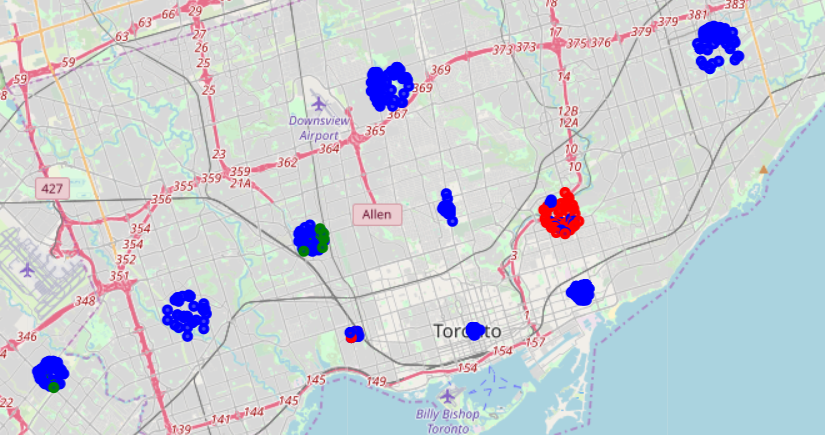
Next, a query for each borough was conducted. This query consisted initially of 200 venues searched using the borough’s coordinates. The data was then loaded into a pandas dataframe and preprocessed:

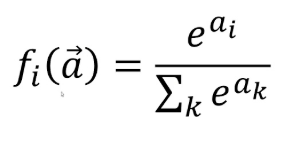
1. entries were sorted by ascending order based in the distance to the reference coordinates;
2. any venue with a distance over 1000 from the coordinates was dropped to avoid registering venues in different neighborhoods;
3. columns with excess information were dropped to avoid polluting the table;
4. the data for the categories column was refined for better understanding and entries without a defined category were dropped.



After obtaining a dataset for each of the boroughs a preliminary evaluation of the categories was performed. From it was observed the presence of several categories, with little recurrence, complicating data analysis. To solve this issue the categories were grouped using a simple text search into ‘Restaurant/Bar’, ‘Store/Shop’ and ‘Services’.

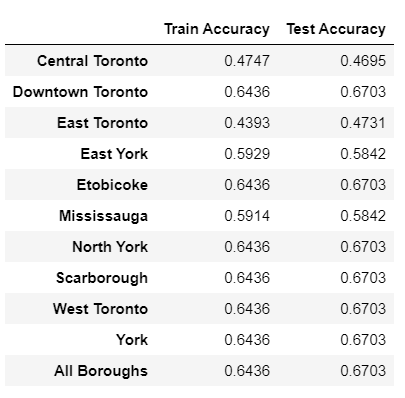
Following the grouping for the categories in each table and the creation of a new column to register them, the data was plotted using the folium package. Restaurants were dotted in green, stores in red and services in blue.



 Lastly, using the datasets for each borough, a Multinomial Logistic Regression with a softmax function was used to build a model to predict to which group a popular venue would belong based in its coordinates. This function and algorithm were chosen due to the target data being multi categorical in nature.

A model was trained and tested using the datasets for each of the boroughs and for the entirety of Toronto (by appending the data of all boroughs to a single table). The accuracy of the models was evaluated for the training and testing datasets of each borough.

Results:



The accuracy of the models was evaluated for the training and testing datasets of each borough and for the combined data.

Training and test accuracy were plotted to a table. From it was observed the coincidence in values for several of the models, pertaining to boroughs with all entries belonging to the services group and to the model trained for all data.

Discussion:

Two main observations can be made for the results of this project:

* Firstly, the methodology used to group the categories for the venues presented crude results. Due to the variety of categories, the approach used allowed for little separation in the data, with most of the entries being allocated in the services group.
* Secondly, the accuracy for the models was considerably low, especially for boroughs with higher variety in group data.

Conclusion:

The project was moderately successful in developing its desired model. Improvements can be made to the process of classifying the venues, either by using more criterion, using machine learning algorithms or directly sourcing the data( the need for several calls to Foursquared API to obtain more detailed data for each venue could pose issues).

Due to the low amount of information available to the model (only the 2 geographical location values are used as independent variables) and the plurality of categorical values the target data can assume(the better the categorization of the venues, the more diverse the possible values), low accuracy can be expected for the prediction models. As such, the predictions from these should be used in conjunction with other data sources to fuel any decision making.