Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Table of contents

1. Introduction: Business Problem

Background

A relatively new company, called ABC Lifestyle Food, recently successfully opened a store in Manhattan, New York, now wants to expand its second store to a different market. The company's store offers healthy (non-alcoholic) drinks and light food to people, both local and out of town visitors. The company sales products attract people with different ethnicities, different ages and different education backgrounds. The common trait is that people are conscious about healthy food or people want to try healthy food or drink.

Its food or drink offering are a small store format with a few choices, similar to Jumba Juice. One of its “ingredients” of success is that store needs to be in a sufficient traffic area. So the first thing in the ABC Lifestyle Food company's mind is to open the second store in a similar market, and they have decided to open in Toronto, Canada.

The first thing the company wants to do is to decide on which neighborhood to open. This project precisely answers the first step of this problem, to find neighborhoods where they have potentially most chance of being successful!

Approach

Because ABC Lifestyle Food store attracts people with different demographics, there is not a set rule to determine which neighborhood to pick. They elected to consult with a data science service company to help them pick a few neighborhoods similar to the neighborhood their first store is in. They will then choose a location among those similar neighborhoods.

The data science consulting determined that they can use Machine Learning technique to find such neighborhoods. The data science consulting will develop a K-means model to group

Manhattan's neighborhoods and identify the kind of neighborhood the current store is in. Then they will apply the K-means model to Toronto, Canada to identify similar neighborhoods.

2. Data

1) The Neighborhood of current store in New York. Based on the company store address, it has been determined that the neighborhood name is "Lincoln Square"
2) New York neighborhood list data. This data has borough sand neighborhoods in New York and their coordinates.
3) Toronto Wiki Page, which has zip codes and their borough and neighborhood names for Toronto.
4) Folium data, which is a map data to help visualize neighborhoods.
5) Foursquare data. This contains most common venues of given neighborhood. This data is used for Manhattan and Toronto.

3. Methodology

Step 1, tie data together:

- Folium data uses latitude and longitude to draw map to visualize neighborhoods.
- Latitude and longitude are tied to local venues by Foursquare data.
- Neighborhoods are clustered by their venues.

Now we have data for every neighborhood all venues listed within its 500 meter radius. This is for both Manhattan and Toronto. This data has over 3K rows for Manhattan and over 4K rows for Toronto. See an example print below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

Step 2, the venue data is then aggregated by venue category, and converted to a dataframe with neighborhood as a row, venue category as column, and with % of venues of the category in each cell. See below for a sample print using Toronto.

	Neighbourhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum
0	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0

Step 3, build K-means model using Manhattan data.

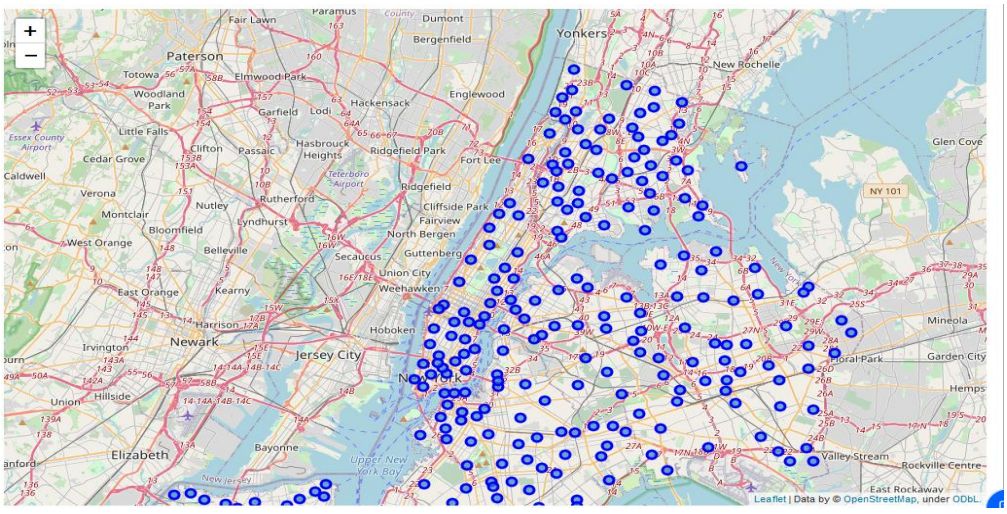
Step 4, apply the model to Toronto data to find the neighborhoods with the same cluster number as the neighborhood of “Lingcoln Square” in. This is the list of neighborhoods for ABC Lifestyle Foods to consider their new store location.

4. Analysis

Step 1, download New York neighborhood list from online. It is then processed to a dataframe as a sample print below.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Step 2, to help visualize NY neighborhoods, the geopy is utilized to map neighborhoods. See map below.



A similar visual is also provided for Toronto, Canada.

Step 3, Manhattan neighborhoods was extracted from NY data for modeling use.

Step 4, Foursquare data was extracted to attach venues to the Manhattan neighborhood data. (please see a sample print in section 3, step 1).

Step 5, using one-hot, convert neighborhood-venue data to a neighborhood-venue-category data for modeling. (please see a sample print in section 3, step 2). Manhattan neighborhood-venue-category data has following dimensions.

```
print(manhattan_onehot.shape)
```

```
(3201, 324)
```

Step 6, download Toronto postal code-neighborhood data from its wiki page and delete boroughs without name. See below of a sample print.

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Step 7, a geo coordinate data is downloaded ties coordinates to the neighborhood via postal code. And multiple entries in the neighborhood column are then converted to multiple rows while post code column was dropped.

Step 8, a series of steps, like step 2 through step 5 for New York were performed to visualize Toronto neighborhood and prepare a Toronto neighborhood-venue-category data. See a sample print below.

	Neighbourhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asia Restaurar
0	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

The Toronto data for applying the model has following data shape:

```
(4214, 272)
```

Step 9, both Manhattan and Toronto data are subsetted to with only common features, so model developed on Manhattan can be applied to Toronto. Following print shows that both have 210 common columns.

```
manhattan_onehot_2 = manhattan_onehot[cats_com]
df_tor_onehot_2 =df_tor_onehot[cats_com]
manhattan_onehot_2.shape, df_tor_onehot_2.shape
```

```
]:  ((3201, 210), (4214, 210))
```

Step 10, both data are then aggregated to calculate means for each neighborhood each category. See sample print and their dataframe shape below.

	Neighbourhood	Mobile Phone Shop	Social Club	Gym / Fitness Center	Trail	Hostel	Gaming Cafe	Cajun / Creole Restaurant	Taiwanese Restaurant	Noodle House	Bus Station	Chocolate Shop	Bookstore	Convenience Store	College Arts Building
0	Adelaide	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.02	0.0	0.0
1	Agincourt North	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.00	0.0	0.0
2	Albion Gardens	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.00	0.0	0.0
3	Bathurst Quay	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.00	0.0	0.0
4	Beaumond Heights	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.00	0.0	0.0

```
(40, 210)
(204, 210)
```

Manhattan data has 40 rows, i.e. neighborhoods, while Toronto has 204 rows.

Step 11, build a K-means clustering model using Manhattan data; find the cluster number for “Lincoln Square” neighborhood.

```
Target_Cluster_num = manhattan_merged[manhattan_merged['Neighborhood']=='Lincoln Square']['Cluster Labels'].iloc[0]
print('Lincoln Square has cluster number {}'.format(Target_Cluster_num))
```

```
Lincoln Square has cluster number 2
```

Step 12, apply the K-means model to Toronto data.

```
toronto_grouped_clustering = toronto_grouped.drop('Neighbourhood', 1)
toronto_cluster = kmeans.predict(toronto_grouped_clustering)
```

Step 13, find neighborhoods with the same cluster number. There are 50 neighborhoods with the same cluster number as “Lincoln Square” in Manhattan. See blow for code and a sample list.

```
Target_neighbourhoods = toronto_merged.loc[toronto_merged['Cluster Labels']==float(Target_Cluster_num)]
Target_neighbourhoods.shape
```

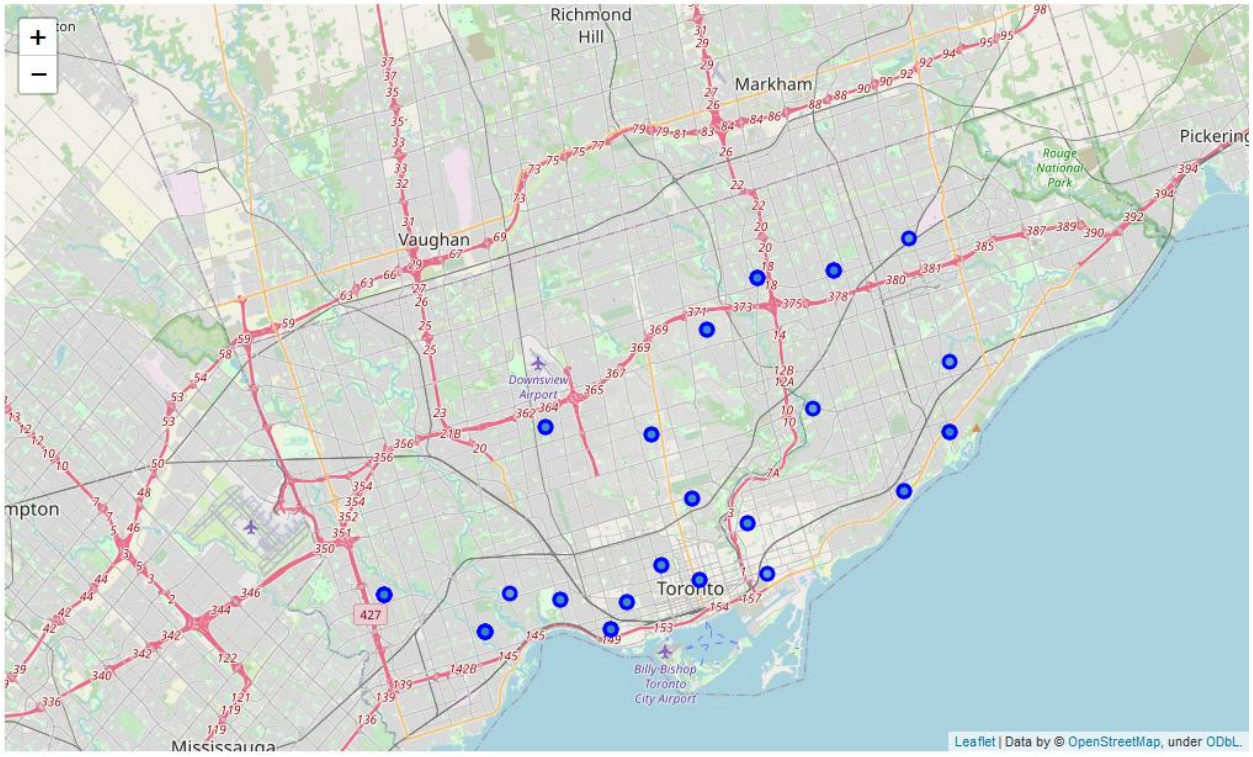
```
]: (50, 19)
```

```
tgt_tor_bor = Target_neighbourhoods[['Neighbourhood','Latitude', 'Longitude']].reset_index(drop=True)
tgt_tor_bor
```

```
]:
```

	Neighbourhood	Latitude	Longitude
0	Victoria Village	43.725882	-79.315572
1	Lawrence Manor	43.718518	-79.464763
2	Lawrence Heights	43.718518	-79.464763
3	Garden District	43.657162	-79.378937
4	Ryerson	43.657162	-79.378937
5	West Deane Park	43.650943	-79.554724
6	Princess Gardens	43.650943	-79.554724
7	Martin Grove	43.650943	-79.554724

Step 14, finally, “put” those neighborhoods on the map for better visualization. Recall that some neighborhoods share the same coordinates (same postal codes), so the number of neighborhoods displayed on the map is smaller.



5. Results and Discussion

The result is a list of neighborhoods in Toronto. Those neighborhoods are similar to "Lincoln Square" based on venue profile. The ABC Lifestyle Foods can confidently use this list to find a suitable available store front for their 2nd store.

To make this model better, additional features can be added to the modeling data. Those features are such as public transportation, neighborhood average income, estimated average visitor and such. Given the limited time, this project did not explore additional data source.

6. Conclusion

The consulting company successfully used machine learning technique helped the ABC Lifestyle Foods find a list of suitable neighborhoods in Toronto for their first expansion. This technique is useful for many companies thinking about their next expansion. This technique is different from traditional method, which normally subjectively matches the neighborhood profile. The machine learning method employed in this project uses essentially a "scoring" method to match the neighborhood profile, which is superior to the traditional method.