Applied Data Science - The Battle of Neighborhoods in Toronto

Leveraging Foursquare location data to explore and cluster neighborhoods

Valerio Marra
IBM Data Science Professional Certificate

The business problem

Definition

Maintaining **neighborhood compatibility** when designing a new real estate development.

It’s a daunting task since 200+ venue categories might distinguish a city and its neighborhoods.

Foursquare

Foursquare location data provides information such as **venue categories**.

Other information include, overall rating, # of tips, agree counts, users and users' friends, popular/trending spots.

Target Audience

The target audience for this data science problem is **real estate developers** interested in finding a *Toronto, Ontario, Canada* neighborhood compatible with their design.

Challenges deep-dive

Challenge 1

Maintain neighborhood compatibility

The new real estate development design needs to be implemented in a neighborhood with the **same character**, or look and feel.

Challenge 2

Analyze big data

Determining the neighborhoods' distinguishing venue categories of the city of Toronto involves analyzing **200+ categories** for **2000+ venues**.

Challenge 3

Define the right methodology

Many **machine learning** algorithms are available.

Their use has to inform business decisions with results that can be properly interpreted.

Solution

Scraping, transforming, clustering, distinguishing big data with...

...**K-means** as the machine learning algorithm of choice together with a **learned decision tree** as an additional tool for drawing the correct conclusions.

Implementation

Data

Postal codes are needed to extract latitude and longitude coordinates of the neighborhoods of interest. *Source: Wikipedia pages.*

Latitude and longitude coordinate are used to retrieve **Foursquare location data** containing venue information.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Data

Foursquare is a location data provider.

By constructing a specific URL a request can be sent to the **Foursquare API** to extract, from Foursquare location data, the unique categories of the venues making up a neighborhood.

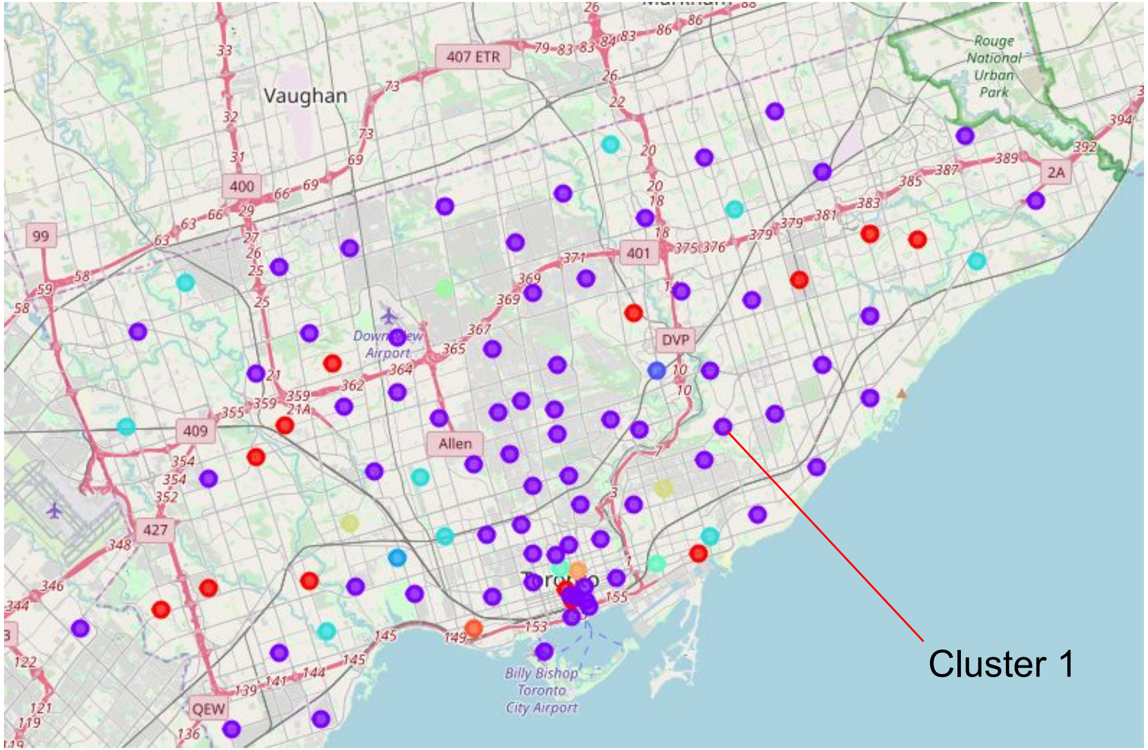
This allows for the **determination of the mean of the frequency of occurrence of each category**. The used clustering technique, k-means, depends on most common venue data. *Source: Foursquare Places API.*

```
----Adelaide,      King, Richmond----
                  venue  freq
0           Coffee Shop  0.07
1                  Café  0.05
2   American Restaurant  0.04
3            Steakhouse  0.04
4                   Gym  0.03
```

Methodology

Used the **elbow method** to find the optimum number of clusters.

Methodology



Toronto clustered in **10** neighborhoods by **K-means**.

Methodology

The top **10** venues for the first **5** neighborhoods.

Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Rouge, Malvern	43.806686	-79.194353	1	Fast Food Restaurant	Dumpling Restaurant	Diner	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	Hardware Store
Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	1	Bar	Yoga Studio	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Eastern European Restaurant	Field
Guildwood, Morningside, West Hill	43.763573	-79.188711	4	Mexican Restaurant	Pizza Place	Medical Center	Electronics Store	Breakfast Spot	Rental Car Location	Drugstore	Discount Store	Dog Run	Doner Restaurant
Woburn	43.770992	-79.216917	0	Coffee Shop	Korean Restaurant	Mexican Restaurant	Yoga Studio	Discount Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant
Cedarbrae	43.773136	-79.239476	0	Athletics & Sports	Hakka Restaurant	Bakery	Thai Restaurant	Caribbean Restaurant	Bank	Fried Chicken Joint	Donut Shop	Dog Run	Doner Restaurant

Methodology

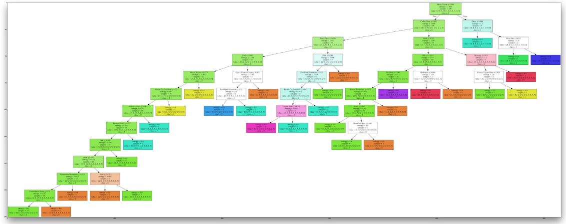
Examining each cluster by checking the **centroid** values.

Cluster Labels	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store
0	0.000769	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000	0.006154	0.000000	0.00000	0.000000	0.001538	0.000769	0.000000
1	0.001679	0.000165	0.000165	0.005866	0.001035	0.001035	0.00207	0.00207	0.00207	0.014297	0.000442	0.00058	0.000145	0.001526	0.000000	0.001053
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.041667	0.00000	0.000000	0.000000	0.000000	0.041667
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000	0.030072	0.000000	0.00000	0.000000	0.000000	0.005682	0.005682
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
8	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000	0.010000	0.000000	0.00000	0.000000	0.010000	0.000000	0.000000
9	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.00000	0.00000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000

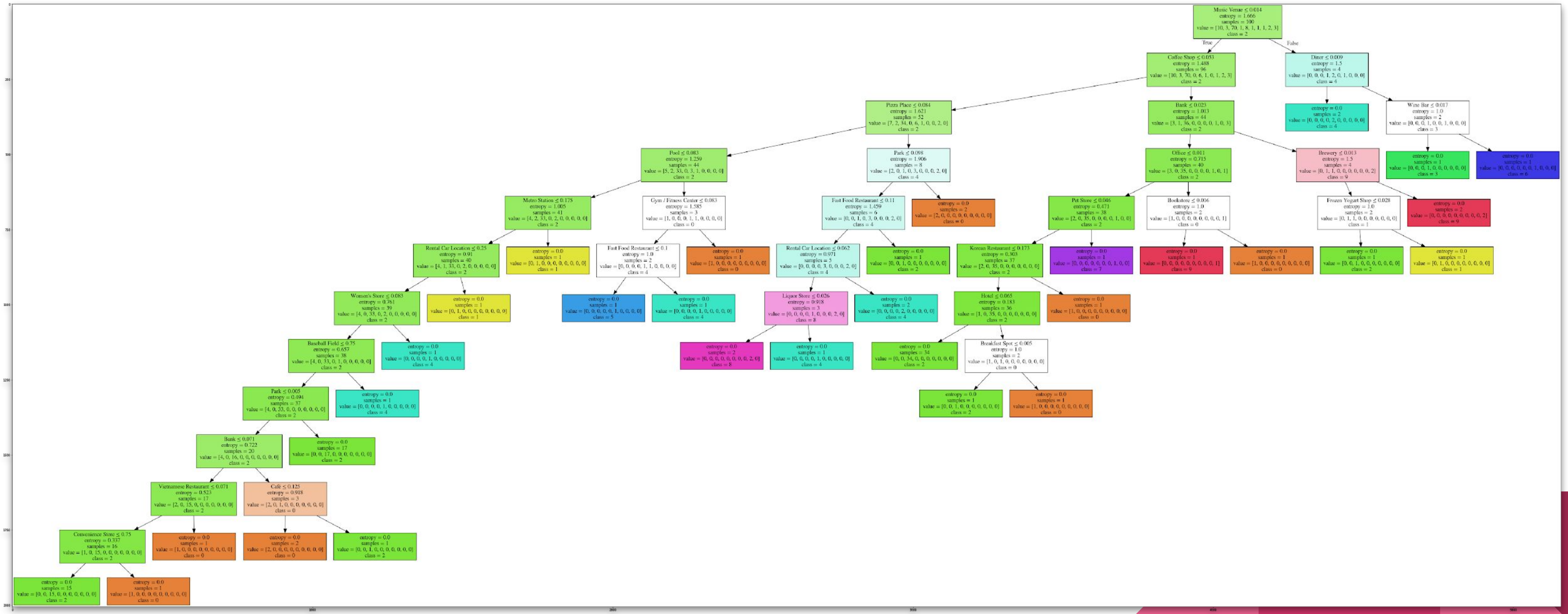
Methodology

Checking the centroid values by averaging the features in each cluster to determine the *character* of each cluster does not prove effective with such a high number (250+) of categories.

Using a popular machine learning algorithm such as **decision tree** to better capture and label the “essence” of clusters is necessary. *The input to the decision tree are the results from k-means.*



Methodology



Results

Health Food Hub

Feature Matrix representing the real estate developer's design, which in this particular case is a **health food hub**.

```
X.loc[X.tail(1).index[0],'Park'] = 1.
```

```
X.loc[X.tail(1).index[0],'Playground'] = 2.
```

```
X.loc[X.tail(1).index[0],'Theater'] = 1.
```

```
X.loc[X.tail(1).index[0],'Wine Bar'] = 2.
```

```
X.loc[X.tail(1).index[0],'Yoga Studio'] = 3.
```

```
X.loc[X.tail(1).index[0],'Café'] = 3.
```

```
X.loc[X.tail(1).index[0],'Cheese Shop'] = 1.
```

```
X.loc[X.tail(1).index[0],'Chocolate Shop'] = 1.
```

```
X.loc[X.tail(1).index[0],'Creperie'] = 1.
```

```
X.loc[X.tail(1).index[0],'Frozen Yogurt Shop'] = 1.
```

```
X.loc[X.tail(1).index[0],'Gastropub'] = 2.
```

```
X.loc[X.tail(1).index[0],'Health Food Store'] = 3.
```

```
X.loc[X.tail(1).index[0],'Mediterranean Restaurant'] = 1.
```

```
X.loc[X.tail(1).index[0],'Movie Theater'] = 1.
```

```
X.loc[X.tail(1).index[0],'Organic Grocery'] = 1.
```

Cluster 1

The **decision tree** trained on clustering results from **k-means** predicts the real estate developer's design as belonging to...

Observations

Observations

Clustering (k-means): the *elbow method* has been used to find the optimum number of clusters ensuring that data have been properly handled and interpreted.

Decision tree: The distribution of neighborhoods across the cluster makes the likelihood of using a *biased* train/test split on the decision tree very high. Following this line of reasoning, the training set used all neighborhoods available and the accuracy of the decision tree has been evaluated *visually* only.

Cluster Labels	
0	13
1	69
2	1
3	1
4	9
5	2
6	1
7	2
8	1
9	1

Recommendations

Recommendations

The location data available allows for searching **Foursquare** users well acquainted (for example in terms of submitted tips, agree counts, and number of friends) with the neighborhoods of interest, i.e. those belonging to cluster 1. The real estate developer might consider engaging such users with a **survey** to better inform his decision and improve his design.

The number of features considered in this analysis, i.e. the number of venue categories, is greater than 250. In this case the interpretation of each cluster proves difficult. The use of a **learned decision tree** is recommended, when clustering is adopted, as an additional tool for drawing the correct conclusions. Such a approach has been followed in the presented analysis.

Conclusions

Conclusions

Given the abundance of location data available to inform an important business problem such as the development of a new real estate design, the choice of finding a solution with the support of **data science** provided the real estate developers with unique insights.

Cluster 1, predicted as the best choice for the real estate developer's design, comprises 69 neighborhoods. Such a high number of neighborhoods will provide the real estate developer with **plenty of opportunities** to find the needed square footage and surrounding infrastructures.

Thank you!