Predicting Suitable Neighborhoods to Open a Coffee Shop

Wendel Zhao

11/18/2020

Table of Contents

1. Introduction

Coffee shop is one of the most common and popular shops in any city. There are usually many coffee shops in a city, however, the demand of coffee shops is still large and can be a golden mine. It is also neighborhood dependent. Select a suitable neighborhood to open a coffee shop is the first and most fundamental step for either a chain enterprise or a personal start-up. Choose the right location and you could end up with a thriving business that can be passed down to your children. Choose wrong and you could lose everything financially.

The objective of this project is to screen out the suitable neighborhoods in Toronto to open coffee shops based on the neighborhoods' characteristics.

2. Data acquisition and cleaning

The target city is Toronto. Two major data sources are utilized. First one is the neighborhood information with location, and the second one is the venues in each neighborhood.

The neighborhood information is available in Wikipedia, as shown in the following picture.

List of postal codes of Canada: M

From Wikipedia, the free encyclopedia

This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website,[1] via its applications for such smartphones as the iPhone and BlackBerry,[2] and sells hard-copy directories and CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries.

Toronto - 103 FSAs [edit]

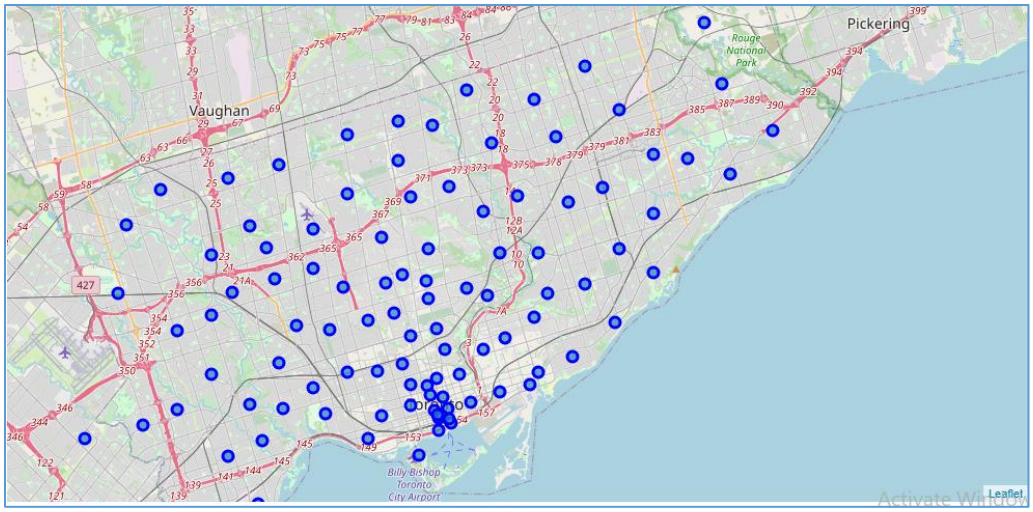
Note: There are no rural FSAs in Toronto, hence no postal codes should start with M0. However, the postal code M0R 8T0 is assigned to an Amazon warehouse in Mississauga, suggesting that Canada Post may have reserved the M0 FSA for high volume addresses.

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village

After scraping and processing, the cleaned data frame as shown below is ready to be utilized for further actions.

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

The neighborhoods can be visualized to provide a general overview:



The venues or shops in the neighborhoods of Toronto is fetched using Foursquare API. The commercial ecosystem is largely reflected by the shops in the neighborhoods, especially mature neighborhoods/cities.

The following picture shows a snapshot of the combined data frame with neighborhood and venues. It has a total number of 2166 venues/shops in the data frame, and this data set will be utilized for the analysis.

	PostalCode	PostalZone Latitude	PostalZone Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M3A	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	M3A	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	M4A	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	M4A	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	M4A	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

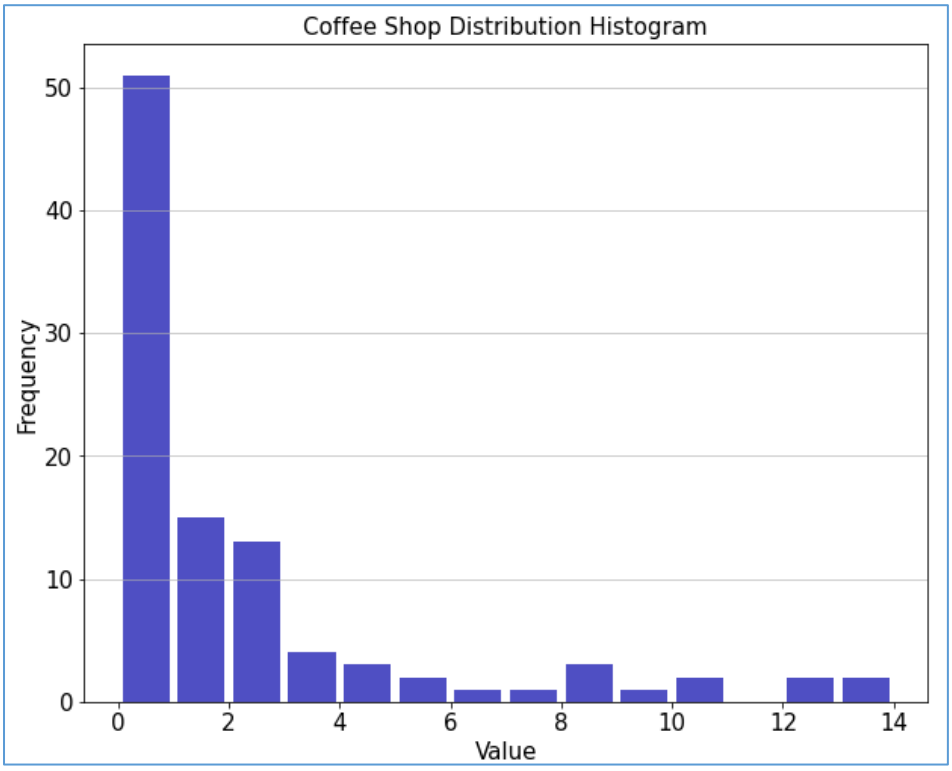
```
toronto_venues.shape
```

```
(2166, 7)
```

3. Methodology and Results

3.1 Evaluation of Coffee Shops

Since the target is coffee shop, the existing coffee shop is firstly investigated. The following picture shows the histogram of existing coffee shops.



Surprisingly, as the histogram indicates, there are large number of neighborhoods which don't have coffee shops.

The following snapshot shows few examples of the neighborhoods that do not have coffee shop. There is no specific abnormal characteristics from these neighborhoods, which indicates that there are chances to open coffee shops from these neighborhoods.

	PostalCode	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
1	M1C	Construction & Landscaping	Bar	Yoga Studio	Dumpling Restaurant	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant
43	M4P	Food & Drink Shop	Gym / Fitness Center	Park	Breakfast Spot	Sandwich Place	Department Store	Dog Run	Dance Studio	Hotel	Donut Shop
46	M4T	Trail	Playground	Yoga Studio	Discount Store	Distribution Center	Dive Bar	Dog Run	Doner Restaurant	Donut Shop	Drugstore
48	M4W	Park	Trail	Playground	Tennis Court	Donut Shop	Diner	Discount Store	Distribution Center	Dive Bar	Dog Run
99	M9W	Garden Center	Bar	Rental Car Location	Drugstore	Ethiopian Restaurant	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Diner

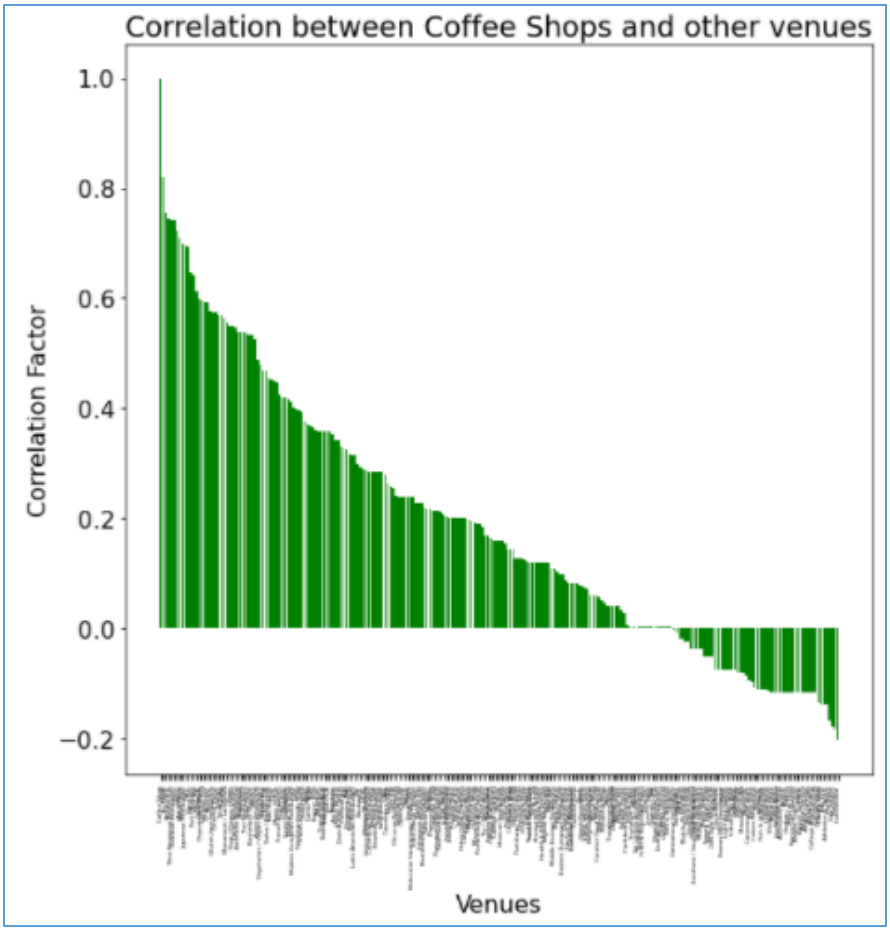
Hence, the dataset is split into two sets:

- The neighborhoods with coffee shops (used for model building)
- The neighborhoods without coffee shops (used for candidate selection)

The first subset is used to build the predictive model. The second subset is used as the candidates for coffee shop opportunities screening utilizing the predictive model.

3.2 Predictive Modeling

There is a total number of 277 unique venues in the dataset. The correlation matrix is evaluated to investigate the correlations between the coffee shop and the other venues, as indicated by the following chart. There are a total of 244 categories showing either positive or negative correlation with coffee shops. For the purpose of simplification and negative correlation with some venues, all these venues are taken into account.



In order to have a higher confidence prediction, 6 machine learning regression models are evaluated, as illustrated from the following picture.

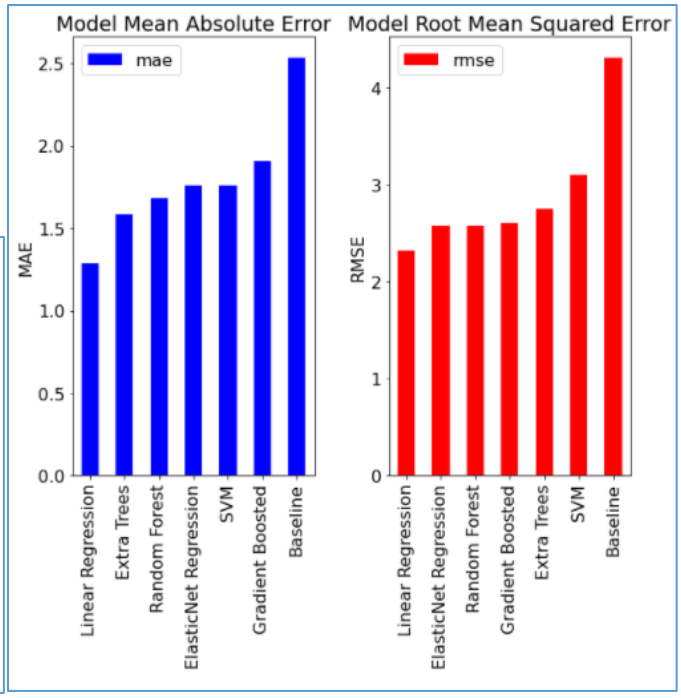
```
model1 = LinearRegression()
model2 = ElasticNet(alpha=1.0, l1_ratio=0.5)
model3 = RandomForestRegressor(n_estimators=50)
model4 = ExtraTreesRegressor(n_estimators=50)
model5 = SVR(kernel='rbf', degree=3, C=1.0, gamma='auto')
model6 = GradientBoostingRegressor(n_estimators=20)
```

The data set is split into training and testing data sets with the proportion of 70/30.

3.3 Results

MAE and RMSE utilized to compare the model accuracy, as shown by the following two pictures. They indicate that that Linear Regression has the best accuracy. It has around 50% improvement.

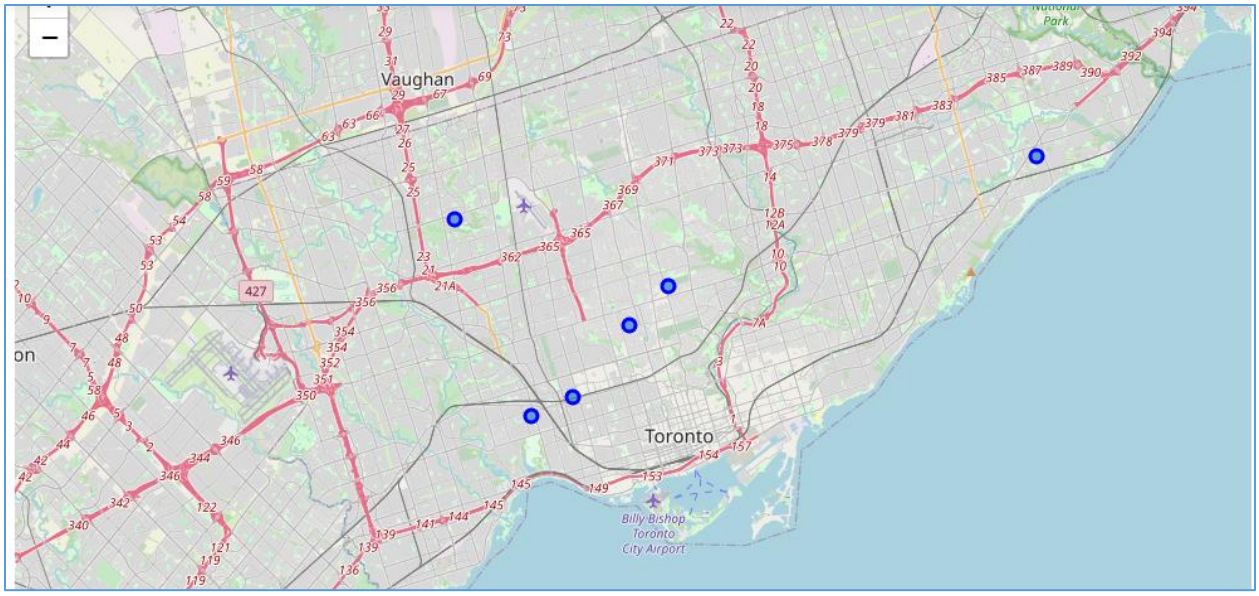
	mae	rmse
Linear Regression	1.28699	2.3239
ElasticNet Regression	1.75553	2.56829
Random Forest	1.68267	2.56917
Extra Trees	1.58133	2.74418
SVM	1.75892	3.09648
Gradient Boosted	1.90635	2.60797
Baseline	2.53333	4.30504



The Best neighborhoods (top 6) are listed in the table below. Based on the commercial ecosystem, each of these neighborhoods is predicted to be suitable to have at least two coffee shops. The analysis provides a good insight for store location. Indeed, the best place to open a coffee shop should also taking into account time-lapse data of shops, rental, etc. The multiple models yield similar prediction, which adds more confidence for these neighborhood.

	PostalCode	Linear Regression	ElasticNet Regression	Random Forest	Extra Trees	SVM	Gradient Boosted	Borough	Neighbourhood	Latitude	Longitude
0	M6P	3.897320	2.641193	2.10	2.36	2.390410	3.555138	West Toronto	High Park, The Junction South	43.661608	-79.464763
1	M1E	2.576944	2.274831	1.70	1.82	1.788216	1.859084	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
2	M6H	2.160494	2.497412	1.86	2.02	1.964447	2.959219	West Toronto	Dufferin, Dovercourt Village	43.669005	-79.442259
3	M4P	2.138635	3.188268	4.38	1.66	1.852964	5.928621	Central Toronto	Davisville North	43.712751	-79.390197
4	M5P	2.118647	2.237996	1.62	1.84	1.698490	1.859084	Central Toronto	Forest Hill North & West, Forest Hill Road Park	43.696948	-79.411307
5	M3L	2.025981	3.188268	4.20	1.58	1.759907	6.124065	North York	Downsview	43.739015	-79.506944

The top 6 candidate neighborhood are visualized as shown in the map below.



4. Conclusions and Way Forward

The project predicts the best neighborhoods to open a coffee shop in Toronto. Foursquare API is utilized to obtain all the venues in the neighborhoods. Multiple ML models (6 models) utilized to evaluate the predictions. Top 6 locations indicates good potential to open the coffee shop.

As a way forward for real world location selection, it is important to perform more detailed location analysis to block or street level. And more data are worth to be incorporated. The data should include but not limit to time-lapse data of the shops, rental information, policy/government information, etc.