Prediction of Ideal Neighborhoods to Open New Coffee Shops

# Introductiuon

Toronto is the biggest city and financial center of Canada with a population of 6,197,000 (as of year 2020). The city is divided into 103 postal code neighborhoods, which are highly heterogeneous in function and population. The residence in each neighborhood also highly diverse in income, education, ethnicity origin et al.. Thus the demand of coffee shops is neighborhood specific.

An ideal neighborhood to open new coffee shops is determined by the law of supply and demand. If the predict number of total coffee shops in a neighborhood is larger than the existing number of coffee shops, there is demand over supply. Thus, the neighborhoods would support a new coffee shop.

# Data Acquisition and Cleaning

The Toronto city postal code and neighborhood were scrapped from Wikipedia (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>) . I parse the table of postal code to a panda dataframe with BeautifulSoup.

The latitude and longtitude data of each postal code were obtained from geocoder. However, due to some technical difficulties with geocoder server, I use the alternative approach, a csv table provided by capstone project. The postal code and coordinate tables were merged into a single dataframe.

I explored the neighborhood with FourSquare API with the latitude and longtitude of each neighborhood. The venues were limited by radius of 500. All venues names and category were added into a new dataframe, which then transform into one-hot format. This new one-hot table has 2123 venue entries and 269 categories.

I tallied the number of venues in each categories in each neighborhood. Since some categories have very small number of venues, I only keep the top 20 categories (at least 22 venues in each categories). This new table was used for training linear regression model to predict the number of coffee shop in each neighborhood.

# Methodology

BeautifulSoup was used for scraping the neighborhood information from Wikipedia. The longitude and latitude of each neighborhood were retrived with Geopy.geocoder. The neighborhoods were superimposed on the map of Toronto with folium. The venues of selected neighborhoods were retrived from FourSquare API.

The table of venues and neighborhood were covert to One-hot dataframe with pandas.get\_dummies. Only the top 20 venues of all Tronoto were kept for further analysis. A table were created for predicting each venue categories were created. Due to the limitation of sample size (neighborhood number), I used K-fold validation for splitting the train/test set due to small sample size into four. The four pieces of sample were train on 75% of sample with Multi-linear regression (scikit-learn). The 25% test set were used to obtain metrics: Coefficient of determination,(R2) Mean squared error and coefficeincy. Mean values of four values of each metric in each categories were store in a table.

# Results

Graphical user interface, application

Description automatically generated

Figure 1. Distibution of 103 postal code neighbor hood in Toronto

Toronto is a the largest city in Canada with 103 postal code with at least 2123 venues exhibited in FourSquare API. Within each postal code, there are multiple venuse, but varied greatly from neighborhood to neighborhood, from only 4 venues in Agicourt to over 100 in University of Toronto, which likely the result of population density. Each neighborhood also have different preference of venues, for example, Agicount favor Skating Rink and Lounge while Beford Park favor Italian Restaurant and Coffee Shop, which likely affect by the function of the neighborhood.

Graphical user interface, application, Word

Description automatically generated

Figure 2. Top 20 Venue categories in Toronto

In the Toronto city, there are different number of venue categories. Coffee shop is the most favorable venue since resident drink lots of coffee. Also lots of restaurant and eateries in the city. The relative big number coffee shop allow us to do prediction more accurately.

Table 1. Metrics of Multi-Linear Regression for predicting different venue categories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **name** | **r2** | **mse** | **coef** |
| **0** | Coffee Shop | 0.631204 | 3.228614 | 3.228614 |
| **18** | Seafood Restaurant | 0.419422 | 0.266531 | 0.266531 |
| **1** | Café | 0.269827 | 1.332593 | 1.332593 |
| **9** | Bakery | 0.107140 | 0.467728 | 0.467728 |
| **8** | Sandwich Place | -0.058658 | 0.439785 | 0.439785 |
| **12** | Fast Food Restaurant | -0.160709 | 0.358247 | 0.358247 |
| **2** | Restaurant | -0.260870 | 0.968311 | 0.968311 |
| **6** | Hotel | -0.349577 | 1.136891 | 1.136891 |
| **5** | Italian Restaurant | -0.454025 | 1.226595 | 1.226595 |
| **7** | Japanese Restaurant | -0.485812 | 0.673359 | 0.673359 |
| **19** | Pharmacy | -0.545532 | 0.288691 | 0.288691 |
| **3** | Park | -0.614441 | 0.708853 | 0.708853 |
| **14** | Sushi Restaurant | -0.712241 | 0.762295 | 0.762295 |
| **4** | Pizza Place | -0.745298 | 1.023304 | 1.023304 |
| **15** | Bank | -1.067827 | 0.551912 | 0.551912 |
| **11** | Gym | -1.141358 | 0.827415 | 0.827415 |
| **17** | Breakfast Spot | -1.347859 | 0.603516 | 0.603516 |
| **16** | Bar | -1.605440 | 0.656598 | 0.656598 |
| **13** | Grocery Store | -1.807140 | 0.822999 | 0.822999 |
| **10** | Clothing Store | -12.434890 | 2.082187 | 2.082187 |

Since only 103 neighborhoods in Toronto, I use kfold (k=4) to split the train/test set for Multi-Linear regression. The average R2 of predicting Coffee shop is 0.63 with mean square root of 3.23, which is the target of prediction among all venue categories. Therefore, I used all samples to train the multi-linear regression model to predict the number of coffee shop in each neighborhoods. The R2 score: 0.8719796647644302. coefficiency: [-0.05698575 0.36527995 0.49928682 0.06049385 0.6038981 0.61981172 0.43625695 0.82013551 0.44346885 0.34628185 -0.05717205 -0.55038192 -0.10135574 0.18406391 0.71333691 0.15898316 0.49716479 0.61669379 -0.61514679] and the intercept: -0.1697859631662182.

Table 2. Prediction of Number of Coffee Shop in each neighborhood

| **Neighborhood** | **Coffee Shop** | **predict** | **diff** |
| --- | --- | --- | --- |
| **89** | University of Toronto, Harbord | 0 | 4.252432 | 4.252432 |
| **18** | Davisville | 2 | 4.685824 | 2.685824 |
| **19** | Davisville North | 0 | 2.269935 | 2.269935 |
| **88** | Toronto Dominion Centre, Design Exchange | 11 | 12.575825 | 1.575825 |
| **27** | Downsview West | 0 | 1.459938 | 1.459938 |
| **39** | Guildwood, Morningside, West Hill | 0 | 1.405996 | 1.405996 |
| **38** | Golden Mile, Clairlea, Oakridge | 0 | 1.216439 | 1.216439 |
| **37** | Glencairn | 0 | 1.209227 | 1.209227 |
| **46** | India Bazaar, The Beaches West | 0 | 1.205436 | 1.205436 |
| **77** | St. James Town | 5 | 6.193834 | 1.193834 |
| **36** | Garden District, Ryerson | 9 | 10.176947 | 1.176947 |
| **50** | Kingsview Village, St. Phillips, Martin Grove ... | 0 | 1.149636 | 1.149636 |
| **41** | High Park, The Junction South | 0 | 1.135422 | 1.135422 |
| **94** | Wexford, Maryvale | 0 | 1.093818 | 1.093818 |
| **10** | Cedarbrae | 0 | 0.987020 | 0.987020 |
| **78** | St. James Town, Cabbagetown | 4 | 4.974818 | 0.974818 |
| **65** | Parkdale, Roncesvalles | 1 | 1.952705 | 0.952705 |
| **4** | Bedford Park, Lawrence Manor East | 2 | 2.933427 | 0.933427 |
| **3** | Bayview Village | 0 | 0.922822 | 0.922822 |
| **70** | Rosedale | 0 | 0.828788 | 0.828788 |
| **9** | Caledonia-Fairbanks | 0 | 0.828788 | 0.828788 |
| **87** | Thorncliffe Park | 1 | 1.784826 | 0.784826 |
| **60** | North Park, Maple Leaf Park, Upwood Park | 0 | 0.772970 | 0.772970 |
| **28** | Dufferin, Dovercourt Village | 0 | 0.700124 | 0.700124 |
| **54** | Little Portugal, Trinity | 2 | 2.682671 | 0.682671 |
| **2** | Bathurst Manor, Wilson Heights, Downsview North | 2 | 2.571001 | 0.571001 |
| **80** | Studio District | 3 | 3.541661 | 0.541661 |
| **66** | Parkview Hill, Woodbine Gardens | 0 | 0.489571 | 0.489571 |
| **57** | Mimico NW, The Queensway West, South of Bloor,... | 0 | 0.384909 | 0.384909 |
| **56** | Milliken, Agincourt North, Steeles East, L'Amo... | 0 | 0.329501 | 0.329501 |
| **52** | Lawrence Park | 0 | 0.329501 | 0.329501 |
| **25** | Downsview East | 0 | 0.329501 | 0.329501 |
| **100** | York Mills West | 0 | 0.329501 | 0.329501 |
| **67** | Parkwoods | 0 | 0.329501 | 0.329501 |
| **99** | Woodbine Heights | 0 | 0.329501 | 0.329501 |
| **84** | The Danforth East | 0 | 0.329501 | 0.329501 |
| **86** | The Kingsway, Montgomery Road, Old Mill North | 0 | 0.329501 | 0.329501 |
| **93** | Weston | 0 | 0.329501 | 0.329501 |
| **101** | York Mills, Silver Hills | 0 | 0.329501 | 0.329501 |
| **74** | Runnymede, The Junction North | 0 | 0.327379 | 0.327379 |
| **0** | Agincourt | 0 | 0.327379 | 0.327379 |
| **91** | West Deane Park, Princess Gardens, Martin Grov... | 0 | 0.273683 | 0.273683 |
| **13** | Church and Wellesley | 6 | 6.231539 | 0.231539 |
| **12** | Christie | 1 | 1.221586 | 0.221586 |
| **30** | Enclave of M4L | 0 | 0.204893 | 0.204893 |
| **15** | Clarks Corners, Tam O'Shanter, Sullivan | 0 | 0.167673 | 0.167673 |
| **73** | Runnymede, Swansea | 3 | 3.154021 | 0.154021 |
| **21** | Don Mills North | 0 | 0.152313 | 0.152313 |
| **22** | Don Mills South | 2 | 2.115390 | 0.115390 |
| **34** | First Canadian Place, Underground city | 10 | 10.101927 | 0.101927 |
| **7** | Brockton, Parkdale Village, Exhibition Place | 2 | 2.066689 | 0.066689 |
| **82** | The Annex, North Midtown, Yorkville | 2 | 2.064297 | 0.064297 |
| **53** | Leaside | 3 | 3.022176 | 0.022176 |
| **35** | Forest Hill North & West | 0 | 0.014278 | 0.014278 |
| **72** | Rouge Hill, Port Union, Highland Creek | 0 | -0.010803 | -0.010803 |
| **14** | Clairville, Humberwood, Woodbine Downs, West H... | 0 | -0.010803 | -0.010803 |
| **47** | Islington Avenue | 0 | -0.109292 | -0.109292 |
| **17** | Commerce Court, Victoria Hotel | 13 | 12.880054 | -0.119946 |
| **51** | Lawrence Manor, Lawrence Heights | 1 | 0.869060 | -0.130940 |
| **43** | Humber Summit | 0 | -0.166464 | -0.166464 |
| **16** | Cliffside, Cliffcrest, Scarborough Village West | 0 | -0.169786 | -0.169786 |
| **75** | Scarborough Village | 0 | -0.169786 | -0.169786 |
| **83** | The Beaches | 0 | -0.169786 | -0.169786 |
| **24** | Downsview Central | 0 | -0.169786 | -0.169786 |
| **97** | Willowdale, Newtonbrook | 0 | -0.169786 | -0.169786 |
| **23** | Dorset Park, Wexford Heights, Scarborough Town... | 0 | -0.169786 | -0.169786 |
| **71** | Roselawn | 0 | -0.169786 | -0.169786 |
| **63** | Old Mill South, King's Mill Park, Sunnylea, Hu... | 0 | -0.169786 | -0.169786 |
| **45** | Humewood-Cedarvale | 0 | -0.169786 | -0.169786 |
| **44** | Humberlea, Emery | 0 | -0.169786 | -0.169786 |
| **6** | Birch Cliff, Cliffside West | 0 | -0.226772 | -0.226772 |
| **58** | Moore Park, Summerhill East | 0 | -0.226958 | -0.226958 |
| **92** | Westmount | 1 | 0.771337 | -0.228663 |
| **1** | Alderwood, Long Branch | 1 | 0.714165 | -0.285835 |
| **20** | Del Ray, Mount Dennis, Keelsdale and Silverthorn | 1 | 0.650350 | -0.349650 |
| **95** | Willowdale South | 3 | 2.508441 | -0.491559 |
| **33** | Fairview, Henry Farm, Oriole | 5 | 4.473838 | -0.526162 |
| **59** | New Toronto, Mimico South, Humber Bay Shores | 0 | -0.580230 | -0.580230 |
| **76** | South Steeles, Silverstone, Humbergate, Jamest... | 0 | -0.657397 | -0.657397 |
| **55** | Malvern, Rouge | 0 | -0.720168 | -0.720168 |
| **42** | Hillcrest Village | 0 | -0.720168 | -0.720168 |
| **79** | Steeles West, L'Amoreaux West | 1 | 0.205434 | -0.794566 |
| **81** | Summerhill West, Rathnelly, South Hill, Forest... | 2 | 1.153389 | -0.846611 |
| **5** | Berczy Park | 5 | 4.134735 | -0.865265 |
| **31** | Enclave of M5E | 12 | 11.041406 | -0.958594 |
| **62** | Northwood Park, York University | 1 | -0.010803 | -1.010803 |
| **8** | CN Tower, King and Spadina, Railway Lands, Har... | 1 | -0.010803 | -1.010803 |
| **90** | Victoria Village | 1 | -0.109292 | -1.109292 |
| **29** | Enclave of L4W | 3 | 1.832801 | -1.167199 |
| **48** | Kennedy Park, Ionview, East Birchmount Park | 1 | -0.169786 | -1.169786 |
| **85** | The Danforth West, Riverdale | 4 | 2.789066 | -1.210934 |
| **61** | North Toronto West | 2 | 0.779977 | -1.220023 |
| **26** | Downsview Northwest | 1 | -0.271142 | -1.271142 |
| **32** | Eringate, Bloordale Gardens, Old Burnhamthorpe... | 1 | -0.282138 | -1.282138 |
| **49** | Kensington Market, Chinatown, Grange Park | 3 | 1.640335 | -1.359665 |
| **69** | Richmond, Adelaide, King | 9 | 7.477445 | -1.522555 |
| **96** | Willowdale West | 1 | -0.825795 | -1.825795 |
| **98** | Woburn | 2 | -0.169786 | -2.169786 |

# Discussion

Base on the difference of predicted coffee shop number and the existing coffee shop number, we can hypothesize the neighborhood have demands for more coffee shops. From the table, University of Tornoto, Harbord have shortage of 4.25 but no coffee shop existed, which likely due to the limitation of opening venue in college campus to meet the need of students. The need coffee shop in Davisville and Davisville North also stunning, with deficit of more than 2. In contrast, Richmond, Willowdale and Woburn already have more coffee shop than the model predictions, where are not good choice to open new coffee shops.

# Conclusion

A multi-linear model is appropriate for predict the number of coffee shop in neighborhood. This report suggests a new coffee shop is preferably open in University of Tornoto, Harbord, Davisville and Davisville North. A new coffee shop is not favorable in Richmond, Willowdale and Woburn.