Identifying optimal business creation in a city: A Capstone Project for Applied Data Science

by John R. Crooker, Ph. D.

June 30, 2020

Introduction

Entrepreneurs frequently make decisions regarding the launch of a new business without the availability of market demand information. The potential for superior returns exists for the entrepreneur if this new market contains customers and clients with pent up demand for the services and offerings of the business. On the other hand, the entrepreneur could lose her entire stake if demand for business offerings fails to materialize.

Hiring a consultant to conduct an extensive survey and evaluation of the potential business expansion is costly. Further, the entrepreneur generally must identify a fairly precise physical location for the new business as well. If the entrepreneur is open to a wide geographic location, costs frequently grow geometrically.

The goal of this analysis is to demonstrate an empirical technique to assess the viability of new business in a market. For concreteness, we will consider new business creation in two cities: (1) New York and (2) Toronto. The algorithms can be applied to any city. The machine learning (ML) technique will make a recommendation regarding the type of business to open and the location. The techniques are described below in the Methodology section.

The ML techniques presented in this analysis uses the existing market structure to assess the diversity and dispersion across a city. The algorithms mathematically search for gaps by business types in geographic locations across a city. These gaps in business type and location indicate areas in which pent up demand is likely to exist. Another advantage of using these ML techniques are that they are relatively costless to employ relative to the formal market demand study and analysis.

Data

To analyze the opportunities for optimal business creation in a city, the ML techniques require neighborhood and postal code data. As will be discussed more fully in the Methodology section below, the algorithm will use FourSquare (foursquare.com) to download Latitude and Longitude information. Additionally, the algorithm will download the nearest 100 existing business within 500 meters of the Latitude and Longitude by postal code. To utilize the FourSquare algorithms to retrieve city data, see the Battle of Neighborhoods Jupyter Notebook link.
For the application in this analysis, we consider business expansion in two North American cities. They are New York and Toronto. For each city, we utilize existing data bases on neighborhoods. In the case of Toronto, we mine the neighborhood data from Wikipedia link.

The table below is a frequency table that identifies the count of Broad Category venues in New York City.

	Frequency	Percent
Value		
Dining	137	44.193548
Bars and Adult Entertainment	46	14.838710
Sports and Fitness	27	8.709677
Grocery and Shopping	26	8.387097
Utility	18	5.806452
Electronics	10	3.225806
Children	8	2.580645
Banks and Professional Services	7	2.258065
Travel and Hotels	7	2.258065
Clothing and Jewelry	6	1.935484
Automobile	5	1.612903
Medical	5	1.612903
Books	4	1.290323
Pets	4	1.290323

The frequency table above reveals that the Broad label 'Dining' covers 44.19% of all defined 'Venue Category' listings for New York City. The Broad label 'Books' covers just 1.29% of 'Venue Category' items. We next consider the frequency of these Broad Category listings applied to Venues in our New York and Toronto data sets.

New York - Frequency of Broad Venue Categories

	Value	Frequency	Percent
0	Dining	5048	53.446268
1	Medical	188	1.990471
2	Automobile	71	0.751720
3	Utility	202	2.138698
4	Grocery and Shopping	680	7.199576
5	Other	1011	10.704076
6	Bars and Adult Entertainment	908	9.613552
7	Banks and Professional Services	151	1.598729
8	Sports and Fitness	623	6.596083
9	Travel and Hotels	115	1.217575
10	Pets	53	0.561143
11	Electronics	129	1.365802
12	Children	46	0.487030
13	Clothing and Jewelry	172	1.821069
14	Books	48	0.508205

Toronto - Frequency of Broad Venue Categories

	Value	Frequency	Percent
0	Other	261	12.270804
1	Dining	1123	52.797367
2	Sports and Fitness	109	5.124589
3	Bars and Adult Entertainment	244	11.471556
4	Grocery and Shopping	125	5.876822
5	Clothing and Jewelry	64	3.008933
6	Electronics	18	0.846262
7	Banks and Professional Services	27	1.269394
8	Utility	32	1.504466
9	Medical	23	1.081335
10	Pets	10	0.470146
11	Books	20	0.940291
12	Travel and Hotels	46	2.162670
13	Children	10	0.470146
14	Automobile	15	0.705219

Density of Broad Categories by Neighborhood

The table below indicates the fraction of New York neighborhoods that contain each Broad type of Venue classification.

	Percent
Category	
Dining	95.289855
Other	87.318841
Grocery and Shopping	77.536232
Bars and Adult Entertainment	67.753623
Sports and Fitness	60.507246
Medical	44.927536
Utility	38.768116
Banks and Professional Services	36.594203
Clothing and Jewelry	34.782609
Electronics	30.434783
Travel and Hotels	23.913043
Automobile	22.101449
Pets	16.666667
Books	13.043478
Children	12.318841

The table below indicates the fraction of Toronto neighborhoods that contain each Broad type of Venue classification.

	Percent
Category	
Dining	81.914894
Other	78.723404
Bars and Adult Entertainment	55.319149
Grocery and Shopping	51.063830
Sports and Fitness	45.744681
Utility	27.659574
Banks and Professional Services	25.531915
Clothing and Jewelry	22.340426
Medical	22.340426
Travel and Hotels	21.276596
Electronics	15.957447
Books	15.957447
Automobile	15.957447
Pets	10.638298
Children	8.510638

Methodology

We use the neighborhood data for New York and Toronto to identify the neighborhoods with the most *diverse* mix of broad category venues. We posit that this broadly diverse areas of the city attract the most consumers. In a feedback loop, consumers likely choose to locate in and around these areas. This suggests that new business openings in these areas will likely be considered by the largest swath of potential consumers.

While this is true, these areas with highly diverse venues are also likely the most competitive. As they are highly competitive, the profit margins from any new market entrant are likely to be constrained. Ideally, we seek to identify a broad category venue type that is *underprovided* in an otherwise highly diverse community.

To identify the diversity of broad category venues in a neighborhood, we calculate the neighborhoods Entropy index with respect to broad category venue types. That is, we define

where is a vector with each element measuring the proportion of venues in that

neighborhood that are in broad category . The neighborhood with the largest Entropy index is interpreted as the most diverse neighborhood and consequently, most attractive, *ceteris paribus*. The neighborhood with the smallest Entropy index is interpreted as the least diverse and least attractive, *ceteris paribus*.
For each neighborhood, we also calculate the attractiveness of opening a business in the neighborhood. This is done using a *cross-entropy* formulization. The cross-entropy formulization allows us to consider a *benchmark* for product offerings in a neighborhood. For New York, we found that 53.44% of all Venue broad category types were 'Dining'. Thus, our benchmark in a neighborhood is for up to 53.44% of all Venues to be in this 'Dining' type. If there are **fewer** than 53.44% of Venues in a neighborhood in this type, we consider it a potential opportunity to open a new 'Dining' venue in this neighborhood. If more than 53.44% of Venues in this neighborhood are in this 'Dining' category, we consider the neighborhood as oversaturated in 'Dining' establishments and would avoid opening 'Dining' category venues in the neighborhood.

Formally, the *Cross-Entropy* is defined as

where is used precisely as above and is the cities *benchmark* proportion of venues in each of the categories. For each of New York and Toronto, we order the neighborhoods from largest to smallest in terms of the Cross-Entropy. The most attractive venue categories for new business starts are the venue categories such that

is as large as possible. We identify these optimal choices for New York and Toronto in the Results section below.

Results

Using the techniques described above in the Methodology, we calculate the top 10 neighborhoods in New York and Toronto according to the *Cross-Entropy* index. The largest value for this index indicates a city neighborhood with the most underdeveloped locations for new venue entry. These neighborhoods are reported in the following two tables.

New York

	Neighborhood	Cross-Entropy
183	Ravenswood	-0.057501
103	Hamilton Heights	-0.063948
150	Bayside	-0.380734
13	Bedford Park	-0.485924
39	Edgewater Park	-0.594268
129	Astoria	-0.767261
104	Manhattanville	-0.802834
117	Greenwich Village	-0.862518
190	Brookville	-1.159461
266	Hunters Point	-1.530178

Ravenswood comes out slightly ahead of Hamilton Heights in New York as neighborhoods with some Broad category venue types that are underweighted relative to the typical pattern in New York. Looks review the existing current Broad category frequency for Ravenswood in New York.

	Frequency	Percent
Value		
Dining	23	82.142857
Bars and Adult Entertainment	3	10.714286
Grocery and Shopping	1	3.571429
Other	1	3.571429

We see that Ravenswood, New York is oversaturated in terms of 'Dining' venue categories. The benchmark number of 'Dining' venues in New York is 53.45% while Ravenswood has 82.14% of its venues in this category. The number of *Bars and Adult Entertainment* venues is somewhat overdeveloped at 10.71% versus the benchmark 9.61%. All other categories are either unrepresented or underdeveloped in Ravenswood. This suggests some opportunities for entrepreneurial expansion.

Toronto

	Neighborhood	Cross-Entropy
21	Central Bay Street	-0.301761
5	Malvern, Rouge	-1.197006
19	Woburn	-1.197006
17	Berczy Park	-1.209888
88	First Canadian Place, Underground city	-1.649461
90	Church and Wellesley	-1.716272
43	Commerce Court, Victoria Hotel	-1.996939
62	Westmount	-2.112128
63	Wexford, Maryvale	-2.184671
9	Glencairn	-2.184671

Central Bay Street jumps out as a neighborhood with expansion opportunities in Toronto. Reviewing the existing Broad category frequency for Central Bay Street in Toronto, we find the following breakdown.

	Frequency	Percent
Value		
Dining	48	73.846154
Bars and Adult Entertainment	4	6.153846
Grocery and Shopping	4	6.153846
Sports and Fitness	3	4.615385
Other	3	4.615385
Books	1	1.538462
Travel and Hotels	1	1.538462
Utility	1	1.538462

We see that 'Dining' venues in Central Bay Street, Toronto seem overweighed at 73.85% versus the benchmark 52.80%. However, nearly every other category is underweighted versus the Toronto benchmark. Thus, we see substantial expansion opportunities in Central Bay Street, Toronto according to our ML techniques.

Conclusion

The goal of this analysis was to develop a ML technique to recognize business expansion opportunities in major cities. As we must keep in mind that regional variations likely exist across populations, we should not expect business opportunities in city A transfer directly to city B.

Our techniques are likely well-suited to identify these regional variations in consumer tastes and preferences. This is because the introduced technique utilizes the overall business structure of the city to establish a baseline for venues in a neighborhood.

With these regional variations accounted for in identifying opportunities, the algorithm quantifies the degree of opportunity by neighborhood. This allows us to quickly identify the neighborhoods with the greatest degree of underrepresented venues while simultaneously ensuring that a substantial venue infrastructure exists to attract consumers to the neighborhood.

The advantage of these techniques is that they allow the entrepreneur to quickly focus in and identify the optimal business opportunities in a city while taking into account regional differences in tastes and preferences. This is likely particularly advantageous to an organization considering multiple cities including unfamiliar cities.

In the analysis above, we identified Ravenswood, New York as possessing substantial expansion opportunities. In Toronto, we identified Central Bay Street as providing opportunities to introduce venues with pent up demand.