Executive summary

Toronto is a world alpha city. It is large, sprawling, and diverse. This means two contradictory things for people wanting to open a restaurant business: First, the city is a prime target due to its population, wealth, and connectiveness; and second, there is a great degree of competition and saturation. There is conflicting anecdotal and statistical data on how likely it is for a restaurant business to succeed at conception, but one thing is for sure: failure rates are extremely high for small startups. There are many factors that affect the chance of success of a startup restaurant, but many of these factors collapse into one word: location. Here, we present a tool that advises new restaurants about the types of neighborhoods most suitable for opening the business. We cluster based on a number of parameters, some of them are obtained purely from data sources, and some are derived. The main contribution in this work is that we consider the type of world cuisine the restaurant intends to serve, and then derive parameters based on saturation, potential, as well as amenability to diversity. These are then used to derive three indicators for likelihood to succeed: a total metric, a linear regression score, and a cluster with similar neighborhoods.

1. Detailed discussion of the business problem

Restaurants are among the least likely businesses to thrive. On conception, three out of five restaurants can expect to not make it past the first year. The ratio rises to four out of five at four years [1]. One major problem with restaurants is that they are a low profit margin industry. They are also extremely sensitive to economic and political upheaval as well as natural disasters. For example, the financial crisis of 2008 squeezed restaurant businesses so tight that only the ones that already had very high margins could remain in business [2]. The COVID-19 pandemic is also decimating restaurant businesses around the world, and the impact is still ongoing.

According to CNBC, the number one factor that affects the success of a restaurant is location [3]. This is also known by common wisdom. But we have to ask ourselves, what makes a location more suitable for opening a restaurant? The following are possible factors:

- The location has a high population. Population can translate into clientele, but it does not have to ..
- Wealth of the location. This is probably more important than population. In fact, this could be the single most important factor in determining foot traffic for the restaurant
- Density of restaurants already in the location
- Density of restaurants similar to the one we intend to open

New ventures in the restaurant business desperately need pointers on where a “good” place to open the business would be. There is a need for this advisory especially to small startups with a limited budget. It is important to understand that most of the anecdotal evidence on failing

restaurants is actually based on impressions from smaller startups. Larger restaurants and instant multi-branch networks can expect a much higher success rate [4].

In this work, we develop a tool to help new restaurant owners decide the chances that their business will succeed in certain neighborhoods. We introduce a number of metrics that together characterize a neighborhood's ability to support the business. Some of the metrics are raw and obtained directly from Foursquare or other sources listed below, some are derived. The metrics will characterize the saturation of restaurants in the neighborhood, but will also have indicators of the diversity, willingness, and wealth of the clientele. To do this, we will fine tune our metrics to include information about restaurants by type of cuisine. The metrics are discussed in detail in the methodology section, but are listed below for reference:

- Population of the postcode
- Income per person in the postcode
- Proliferation of restaurant business in the location
- Proliferation of restaurants of particular cuisine in the location
- Diversity of restaurants in the location
- A metric that indicates spending power per person per restaurant

Target audience: The target are prospective owners of new restaurants, particularly small restaurants with a staff count of less than a couple dozen. This is a high risk, low margin, high closure business model that is extremely sensitive to location.

Summary of the problem: Given I will be opening a business of a specific cuisine type, what is the best group of locations that could support my business. The location has to have high potential to generate traffic, but should also show signs of low saturation. Once I have chosen a cluster of locations, what can you tell me about the relative advantage of each location within the cluster?

2. Data needed and data sources

Our master data frame will contain information about a group of postal codes in downtown Toronto. For identification purposes we need the postal code, name of neighborhoods, and borough. We also need latitude and longitude data to query Foursquare. We have to somehow obtain data on income and population to characterize the potential of the location. Finally, we have to obtain restaurant count by world cuisine, for which we will use Foursquare. Datasets and their sources are discussed in detail below.

Postal code, neighborhood, and borough
This has been obtained in week 3. Code for obtaining it is included in the notebook. Sources are listed below. In week 3, we also introduced a reduced frame containing only boroughs with the string "Toronto" in the name. We will also need a reduced frame here for reasons that will be explained below.

Source for neighborhood information:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Latitude and longitude data by postal code
This has also been done in week 3. Data can be obtained by interrogating geoawesomeness or by pulling from the file provided in week 3 (URL attached below). We choose the latter due to the lack of reliability of the tool.
https://cocl.us/Geospatial_data

Number of restaurants around center of postal code
This is the total number of restaurants drawn from a pool of world cuisines around the center of each of the postal codes in our main data frame. We will actually query Foursquare for the number of restaurants for each of the cuisines and will be storing them separately. This allows us to query Foursquare only once, then decide on the target restaurant cuisine offline. The list of target cuisines is ['Italian', 'Indian', 'Chinese', 'Japanese', 'Korean', 'French', 'Greek', 'Peruvian', 'Brazilian', 'Mexican', 'Spanish', 'German', 'Moroccan', 'Egyptian', 'Turkish', 'Persian', 'Thai']. This list can be changed, reduced, or expanded by the user. Once we obtain this information, we save it to a CSV file, allowing the rest of the algorithm to rerun without needing to query Foursquare. This allows us to rerun as many times as we need without being limited by Foursquare quotas.

Density of restaurants of particular cuisine around postal code center
This has to be distinguished from other cuisines when calculating one of the derived parameters. It will also be the label of our regression problem. This can be queried no different from other cuisines using Foursquare. In fact, as discussed above, we obtain and store raw data for all cuisines and store it in our main dataframe, allowing us to make and change the decision on the target cuisine without needing to contact Foursquare.

Population by postal code
As discussed above, population is a critical parameter in determining foot traffic. The tables below allow us to obtain this information by postal code. Since postal code (also known as FSA) will be the key in all our data frames, this will be very helpful.
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&SR=1&S=22&O=A&RPP=9999&PR=0
Data in CSV format:
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/File.cfm?T=1201&SR=1&RPP=9999&PR=0&CMA=0&CSD=0&S=22&O=A&Lang=Eng&OFT=CSV

GDP of postal code
https://open.canada.ca/data/en/dataset/1e8a8c6e-a7cc-4f08-a3cc-79c67427a102

https://www.canada.ca/content/dam/cra-arc/prog-policy/stats/individual-tax-stats-fsa/2015-tax-year/tbl1a-en.pdf
It is very hard to obtain income data by postal code in Canada. Census authorities do include very broad data about income in provinces. They also include data about income by gender, group, and other parameters. In terms of geographic distribution, the census office also has income data by census unit, this sometimes maps to FSA, but in other cases the mapping is not so easy. On the other hand CRA has tables on taxable income by FSA. This is as good as the data we need, if not more so. Reported (taxable) income is a very good indicator of the amount of spending power the population has in the location. There are two issues with this data:

- This is total income rather than per capita. This is not a big deal. A simple division can give us a per capita metric. Notice that total income by itself may be a valuable indicator, but it is not as good as per capital income due to lack of normalization by population
- The bigger problem is that data is in PDF format. This means we have to manually intervene to extract the data in a way that allows it to be included in the master frame
- This table provides data on population, however, the population data is misleading since it only indicates taxable population, we will use the population data from statcan instead

In the methodology section we will discuss how this data is preprocessed so that the metrics from the business problem section can be combined into a meaningful assessment of the location. Our aim will be to create “positive” metrics that are best when they increase and “negative” metrics that are the opposite.

Positive metrics will in general indicate potential for spending and traffic, negative metrics are indicators of competition or saturation. We will create six metrics, one of which will play the role of a parameter and a label under different scenarios. We will do three things with these metrics:

- Use K means clustering to cluster similar postcodes together. This will indicate which groups of postal codes are more likely to support our target cuisine. Using map visualization, we may be able to obtain some insight about whether the geographic distribution reflects a particular pattern
- We will calculate a “total metric” using the six parameters. This metric will range from 0 to 6. Zero will represent absolutely no potential for growth in the particular cuisine. A score of 6 represents the highest potential for success
- Linear regression will be used with the number of restaurants from the target cuisine being the label and all other metrics being features. We will use multiple linear regression and then calculate the error for every postal code. This error will then represent how far off the postal code is from what the model suggests. Positive errors could indicate oversaturation, while negative errors can be a sign of unused potential

References

[1] https://www.getorderly.com/blog/high-restaurant-failure-rate

[2] https://smallbusiness.chron.com/successful-profit-margins-restaurant-business-23578.html

[3] https://www.cnbc.com/2016/01/20/heres-the-real-reason-why-most-restaurants-fail.html

[4] https://www.forbes.com/sites/modeledbehavior/2017/01/29/no-most-restaurants-dont-fail-in-the-first-year/#21742fc04fcc