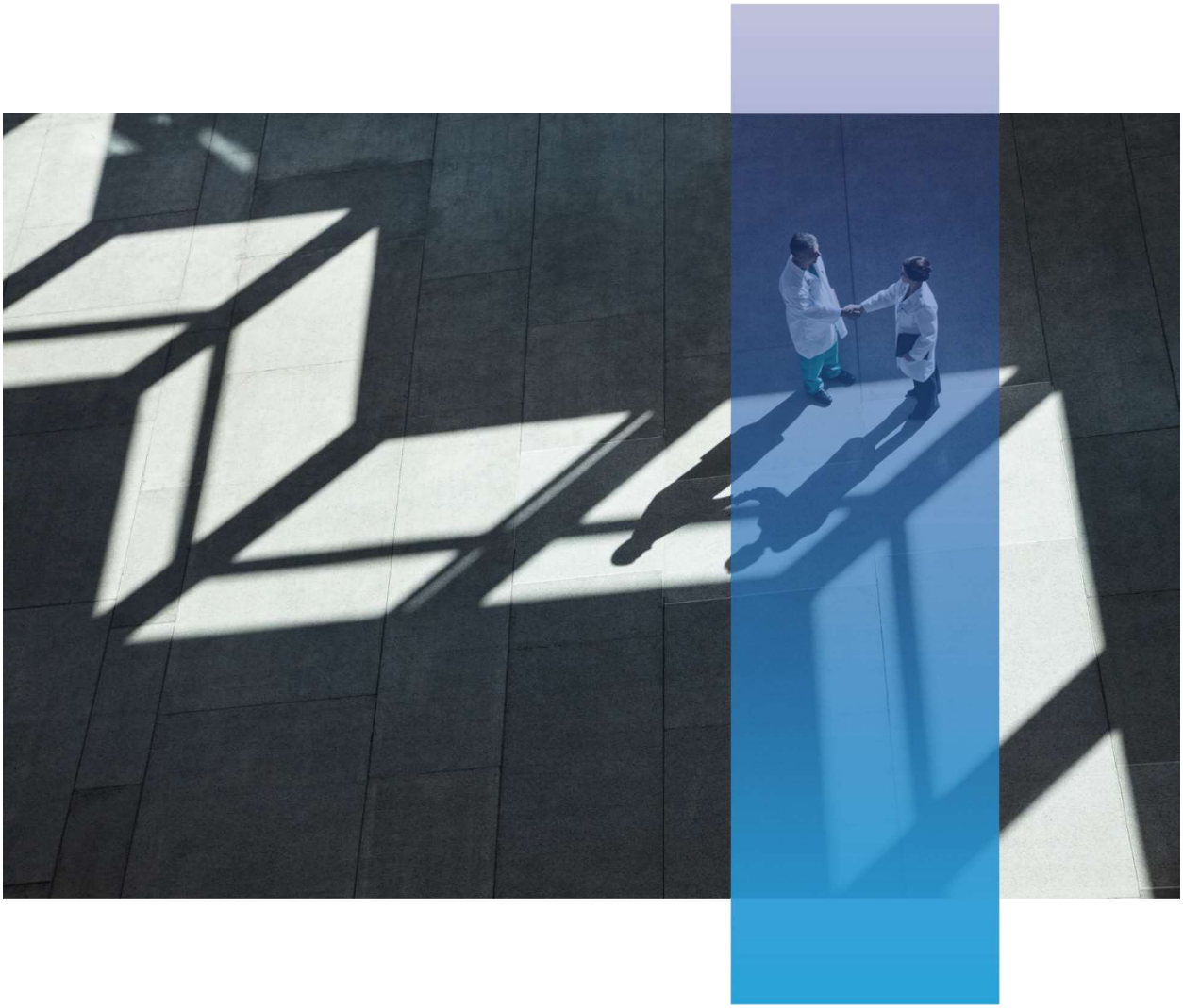
THE BEST NEIGHBOURHOOD FOR MY RESTAURANT

Capstone Project

August 2020

Nachiketa N. Kurhade



WHERE SHOULD I OPEN MY RESTAURANT?

Many entrepreneurs face this problem while opening their first restaurant or while opening a new branch of their restaurant. Hands down, setting up a new restaurant is costly indeed, and if it fails to return profits, the investment becomes a disaster. On the other hand, a restaurant in profit is a business success. So, if you want put your money on the right place, you better do your homework, you better have the formula for success!!
In this report we aim to determine the best neighborhood in the city of Toronto to open up a new restaurant which will require minimum initial cost and will result in maximum profit.

To solve this problem, we need to answer two questions:

1. Which neighborhoods are performing better in terms of restaurant business?
2. What are the factors that are making these neighborhoods perform better?

DATA

We will be breaking down this analysis in three steps

1. Collect data about neighborhoods in Toronto.
2. Define performance parameters and rank the neighborhoods by their performance.
3. Find the neighborhood characteristics that support your performance results.

For this analysis, the data was collected using Foursquare API, Open Data Portal – Toronto City, IBM Cognitive Class Labs, and Wikipedia.

- Primary neighborhood data (contains basic information about neighborhoods) was collected using the Wikipedia page that contains information about boroughs, neighborhoods, and their postal codes for the city of Toronto.
- Using data provided by IBM Cognitive Class Labs, we assigned co-ordinates to these neighborhoods, thus preparing our primary neighborhood database.
- Secondary neighborhood data (contains primary neighborhood database with venue data from each neighborhood) was collected using Foursquare API.
  Regular Foursquare API calls were made to collect information such as top 100 venues, their coordinates and venue category.
  Radius was set to 500 meters.
  Venue return Limit was set to 100
  A quick folium plot showed that some places had neighborhood centers are really close. Hence, we had to limit radius to 500 meters as increasing radius might have resulted in duplicate venues.
  Secondary data had too many venue categories, and many of these values were actually the same value written in a different way, so it was cleaned up to remove such duplicates.
- Tertiary database was created by separating restaurants from our venues data frame (Secondary data). As we don't have an actual turnover data of each of these restaurants, we needed to set up some other criteria to analyze how well they're performing.
  Using the data returned by Premium Foursquare API calls which return venue tips, we considered using two parameters, ‘Rating' and 'Check Ins', to determine their performance. As both of these parameters are vital and directly proportional to restaurant's performance, we will set up another parameter 'score' which is the product of both.
  Performance of a neighborhood was defined as the score of the respective neighborhood per restaurant. Therefore, the ‘performance’ parameter won't get biased like ‘score’ as a neighborhood ‘score’ can get more just because there are a greater number of restaurants in the first place. The definition of ‘performance’ corrects this problem and thus even if there are less number of restaurants in a particular neighborhood but if they are scoring exceptionally, the corresponding ‘performance’ will be better and the respective neighborhood will rank higher.
  This is how we answered our first question i.e. **Which neighborhoods are performing better in terms of restaurant business?**
  Ranking neighborhoods in ascending order of 'performance' should give us an idea about the neighborhoods whose restaurants are receiving most recommendations and are being frequently visited by people.

- Going through some reports on factors to consider while setting up a new restaurant, we realized there are more factors which may contribute to our 'performance'. This kind of data can-not be collected using Foursquare API. Therefore, extra database was collected from 'Open Data Portal – City of Toronto'. This extra data included information about

  1. Festivals and events
  2. Neighborhood profiles
  3. Parking lot facilities
  4. Places of interest and Attractions
  5. Public art work assets
  6. Culture
  7. Economics
  8. Recreation
  9. Safety
  10. Transportation

  Out of these I am excluding 'Festivals and events', 'Recreation' and 'Public art work assets' as given a close look they seem much less relevant compared to others.
  Also, we took liberty to collect the data from "Neighborhood profiles
  ", "Culture", "Economics", "Safety", and "Transportation" into a single dataset "overall" for easy handling.
  This data was used to check the common claims made by above mentioned reports by using machine learning algorithms like regression.
  Also, in Canada, the Neighborhoods as defined by postal data are different than the ones defined by the city administration, later being based on population census.
  This induces one complication, i.e., neighborhoods in this new data don't match to the ones we have been using till now. To solve this, we identified the new neighborhoods by their postal codes, since these boundaries overlap, the most frequent postal code had to be taken. This column served as key to connect this extra database with other databases we had been developing so far.
- Logically speaking, if we categorize venues from secondary database, we should be able to find some categories here which have significant impact on our 'performance'. But as our input set is categorical, we needed to convert it using one hot encoding first, then we performed the usual multiple linear regression analysis.
  By performing the analysis mentioned in last two bullets we were able to answer our second question i.e. **What are the factors that are making these neighborhoods perform better?**