IBM/Coursera´s Applied Data Science Capstone course

Assignment: Capstone Project - The Battle of Neighborhoods

Final Report

Project Title

Candidate neighborhoods for opening new restaurants - Methodology and case study.

Author: Paschoal Molinari

Date: 2021-02-24

Introduction

Welcome!

This is the report of the project “Neighborhoods candidate to the opening of new restaurants - Methodology and case study” presented as the final project of the course “Applied Data Science Capstone” given by IBM on the online platform Coursera.

The problem to be solved is finding neighborhoods that are good candidates for opening new restaurants. The audience for this Project is companies and entrepreneurs in the restaurant business involved in opening new venues.

A specific methodology for restaurants is developed and applied to an Italian restaurant case study in the city of Toronto, Canada.

The project is developed using Data Science tools such as the Python language, tools and libraries, and important geospatial data from Foursquare.

Business Problem

The business problem in question is finding neighborhoods in a specific city that are potentially good candidates for opening new restaurants. This problem affects many companies and entrepreneurs in the restaurant industry who need timely information for decision-making when choosing a new business location.

Usually finding a good neighborhood for a new business is a time and resource-consuming effort and has some blind spots of factual data about demand and competition. Foursquare geospatial data can be used to uncover these spots.

The solution to this problem aims to:

- Seize business opportunities with high demand and low competition.
- Increase return on investment (ROI).
- Maximize the chances of a successful restaurant opening.
- Add the value of proximity to potential customers of new restaurants.
- Reduce the risk, time, and cost of neighborhood selection.

The problem is complex and the methodology developed selects potential good candidates neighborhoods for more detailed analysis in the field and final business decision.

The methodology developed can be extended to other lines of business.

Data Description

For this Project, Foursquare's geospatial data are used.

The geospatial features are:

- Postal Code
- Neighborhood
- Neighborhood Latitude
- Neighborhood Longitude
- Venue
- Venue Latitude
- Venue Longitude
- Venue Category

The geospatial data obtained from Foursquare about the different venues (not just restaurants) are engineered, tabulated and new statistical features, which are part of the methodology, are elaborated.

As methodology´s case study we use geospatial data from the City of Toronto, as it is well known and has a wealth of details.

Specifically for the case study, the Postal Code of the City of Toronto are used from this link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The postal code is used to obtain the location (Latitude and Longitude) of Toronto´s neighborhoods. Also, some neighborhoods have more than one postal code, the postal code is used as the dataframes primary key avoiding eventual problems.

In the case study, additional features are generated:

1. Top 10 venues by neighborhood.
2. Total count of venues per neighborhood
3. Total count of restaurants by neighborhood
4. Total count of Italian restaurants by neighborhood.
5. Mean of restaurants by neighborhood.
6. Mean of Italian restaurants by neighborhood.

The means (items 5 and 6) are especially important for comparative analysis. The lower the mean the better is the chance of success for the new restaurant.

For example:

- Mean of restaurants (item 5) is calculated by dividing the count of restaurants (item 3) by the count of venues (item 2) for a given neighborhood.

- Mean of Italian restaurants (item 6) is calculated by dividing the count of Italian restaurants (item 4) by the count of venues (item 2) for a given neighborhood.

Finally, the machine learning clustering technique (K-means) is applied to the new features, specifically the means (items 5 and 6), generating more useful information about the potentially good neighborhood candidates.

The case study data are available in csv format files in the Project Github repository.

Methodology

Lets first understand some important concepts:

I. The higher the number of venues in a neighborhood probably the higher the number of people working or circulating in the neighborhoods. This means a higher need for venues like restaurants.

II. The lower the restaurant category mean to the total of all venues in the neighborhood probably the higher the need for a restaurant of this category.

III. The lower the restaurant of all categories mean to the total of all venues in the neighborhood probably the higher the need for restaurants in general.

IV. Also, the above means are useful **after** you add the new restaurant in the statistics.

From the above concepts, we decided to extract the means (II and III) to analyze and decide the potentially good candidates to be studied for open a new restaurant.

Since there are no specific weights to say which mean is more important than the other, the Machine Learning K-means is applied to cluster the data without the need for weights. This is good because weights are too subjective in this case.

With K-means clusters we have a 2-D vector instead of a single scalar value with subjective bias in its calculation. The smaller the vector probably the better the cluster.

Each mean value could eventually be classified as low, medium, or high value and the combination of it could be nine groups or clusters (3 possibilities x 3 possibilities). Then nine was the choice of K-means clusters and it worked very well.

The cluster closer to the origin (0,0) of the cartesian diagram is probably the best candidate to be further analyzed. It is the cluster with the lower normalized means vector.

The neighborhoods in this cluster are potentially good candidates for opening a new restaurant.

Finally, the clusters are sequenced according to the size of the vector (norm) of the center of each cluster.

The results are presented in maps and plots.

More considerations:

- In the case study, the small neighborhoods (10 venues or less) were excluded for simplicity. For small neighborhoods, the radius should be increased to something like 1500 meters.
- All restaurant categories have no distinction in this case study.
- The ranking of the restaurants (like Michellin stars, etc..) is not included.
- The average price of restaurant meals is not included.
- The population census of the neighborhoods is not included.
- The cost of open a new restaurant in a specific neighborhood is not included.
- The trending venues are not included.
- The venue impressions by Foursquare users are not included.

Some of the above considerations should be included in a more rigorous study depending on the goals or in the next steps of the decision-making process.

This is more a generic first-step approach to the problem and can focus the resources during the neighborhood selection process.

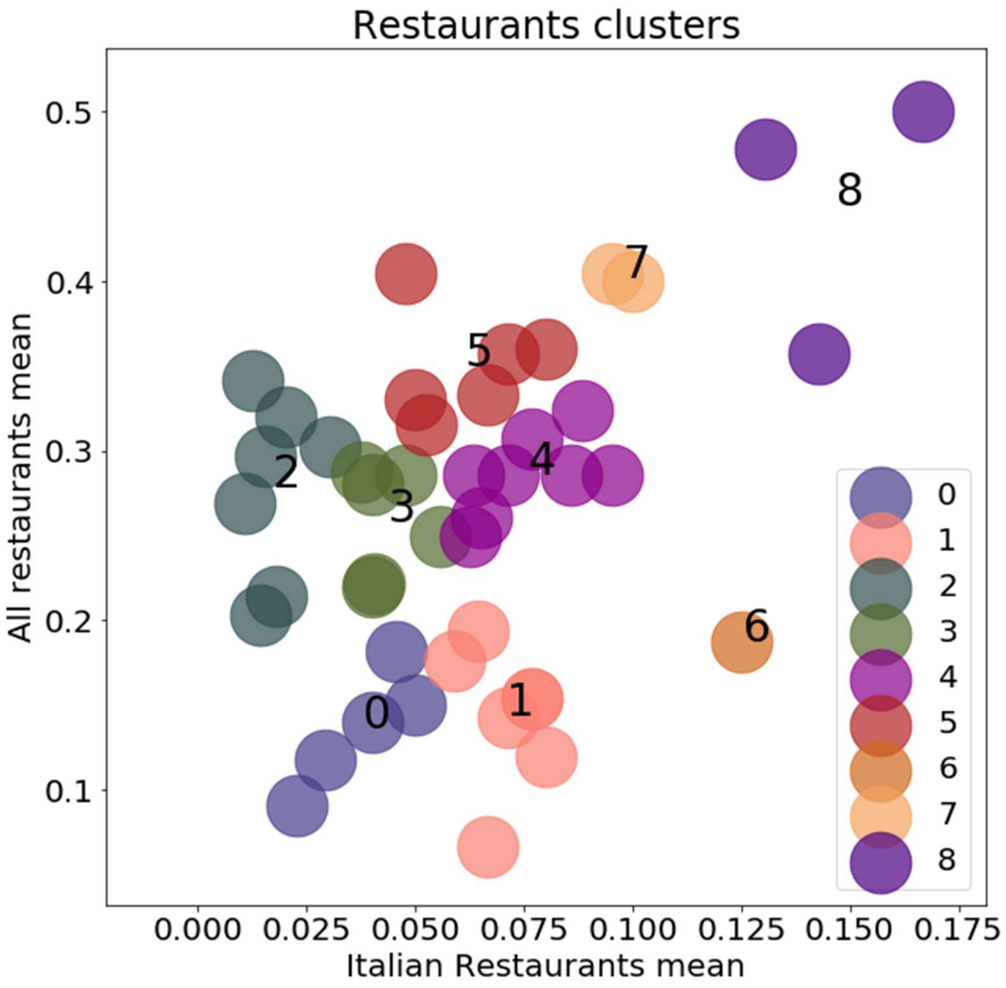
In the notebook, during the visualization of the geospatial map and means plots, it was used the same moderate non sequential color scheme. This way readers can analyze more comfortably the results. But for presentation may be an intense sequential color map could have more visual impact.

Results

The case study was very good and nine clusters were generated by the K-means machine learning.

The Plot of the clusters is below. The clusters are sparsely distributed.

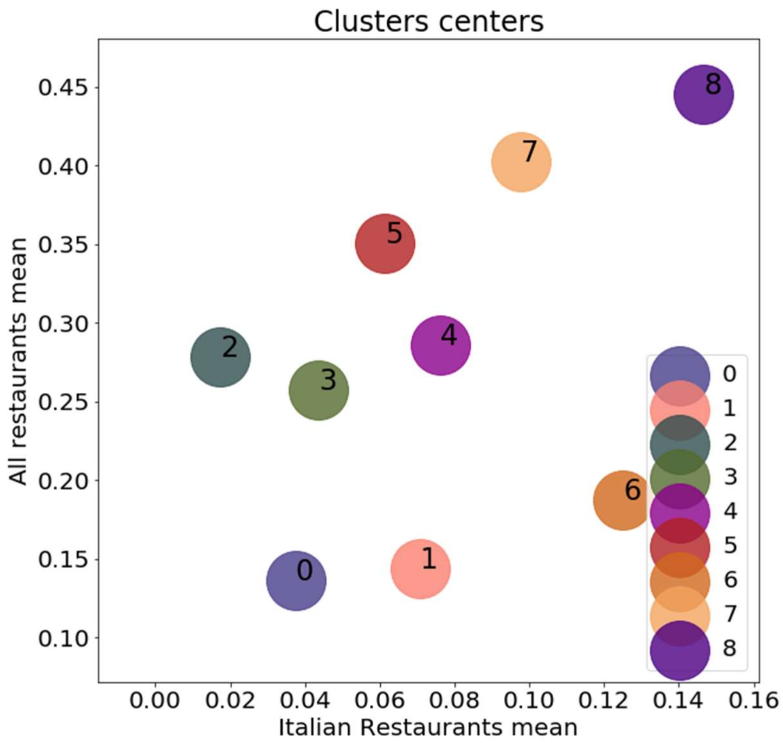
The clusters colors in the plots and the geospatial map are the same.



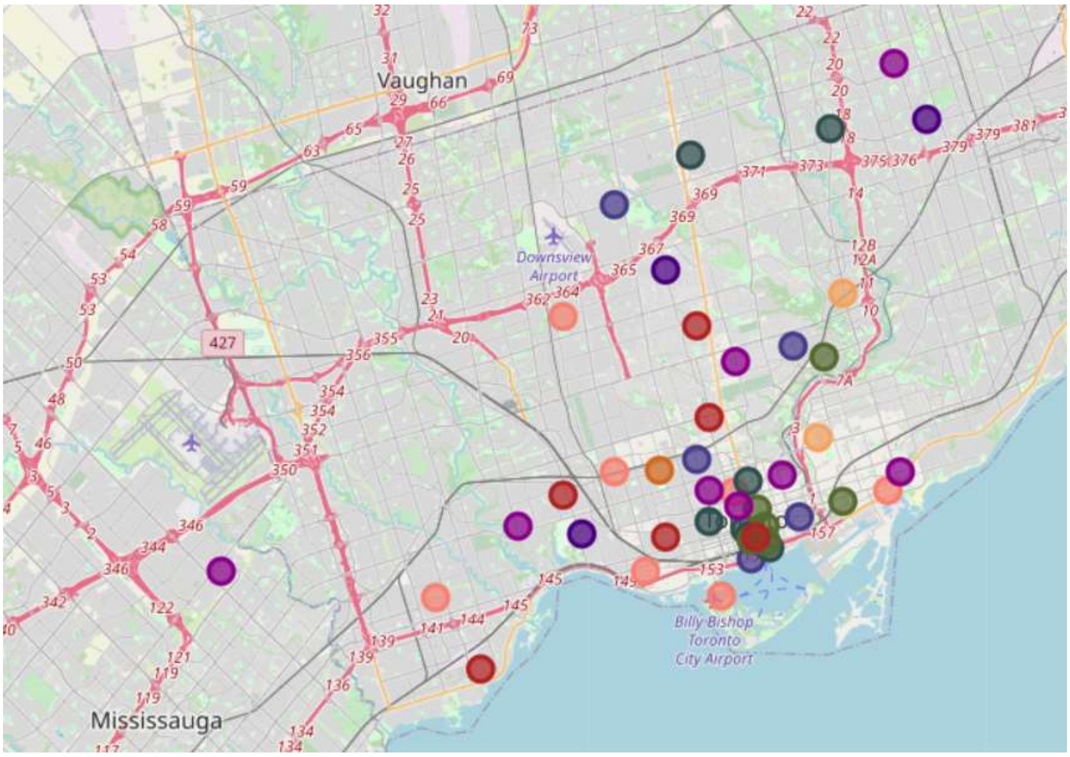
We can see that cluster 0 is the best because is the closest to the point (0,0). It has the smallest 2-D vector. Other good clusters are 1 and 2 followed by 3.

Care should be taken because there are some small data variations between different months/years of data acquisition from Foursquare. It may be enough to have a change in the cluster distributions. Full reproduction of this analysis is possible using the saved data (csv files) in the project repository.

Let's see, more clearly, the position of the center of the clusters.



In the notebook, it is possible to see the geospatial map too.



It can be seen that neighborhoods in the same cluster are not necessarily geographically close.

Discussion

After seeing the methodology and results some questions arise:

- Is cluster 0 the best cluster? Up to now, it is the best candidate but needs further analysis with new data.
- Are clusters 1 and 2 are better than cluster 3? Probably yes because it loses to 2 in Italian restaurant mean and 1 in all restaurant mean.
- And the other clusters? It should be analyzed after the first four clusters in a sequence like 4,5,6,7 and 8.

The notebook can be used to analyze small neighborhoods (10 or fewer venues). The suggestion is first to identify the small venues and then get more Foursquare data with 1500 meters radius and include a data census. This should give a better detail of the neighborhood's profile.

If the idea is to open an above price average restaurant then maybe the radius should be 1500 meters too.

Trending venues information could some interesting information to promotions, etc.

The Foursquare user´s venue impressions could be very helpful but need much more work. Here some NLP (Natural Language Processing) could be used.

Conclusion

The geospatial data from Foursquare is definitely of high quality and high value, and the machine learning K-Means for clustering is very fast and practical to discover important information in statistical features. The vectorization of the clusters' features´ average is an effective way to sequence the clusters and give them meaning.

The objective to find potentially good candidates neighborhoods to open a new restaurant was successfully achieved. This Project opens the opportunity for a deeper understanding and future projects in the area. The information based on factual updated geospatial data and machine learning is of high relevance to the business decision-making process.