Coursera Capstone Final Project

Predicting the best location with more potential revenue

Raphael Ferreira

July 6, 2019

1. Introduction

1.1 Background

A start-up based on Toronto wants to expand to New York. Its core business is delivery: customers wants to receive something from coffee shops, drugstores, pizza places, groceries, for example. Through the app customer demands some product, which it can indicate a specific place or not, and where it wants that product to be delivered. Most of the products are delivered by walk or by bike. So, spatial conditions are very important for such start-up, specially the category of venues that exist in each place, since it represents places where it could have more delivery transactions, and therefore more revenue.

1.2 Problem

In order to begin its operations in NY, this start-up, our client, wants to know the best location, which might have more potential to bring deliveries orders to them, and therefore more revenue. They already know its best location in Toronto, but they need to estimate where to start in NY. By the most common type of venue categories in its best location in Toronto, we need to predict the best location in NY, for this first move of our client.

1.3 Summary

This work is divided in two main parts:

- Unsupervised/spatial: where we get Foursquare Data and create clusters for Toronto and NY neighborhoods. In Toronto we find our target cluster: Cluster 3 - location with venues that are customers of our client

- Supervised: where we work with data wrangling (categories spatially sorted and transform data into numerical), visualization (threshold for classification) and prediction (LR and Decision Tree Models)

2. Data acquisition and cleaning

2.1 Data Sources

The dataset that we are going to use come from Foursquare, mainly the type of venues and its quantities for each neighborhood in Toronto and NY, so spatial data. The acquisition is made by API that gets the data from Foursquare. Through such data we first create the clusters for Toronto, recognize the most important cluster, then we have data frames that have the 10 most common types of venues for each neighborhood. The same procedure is done for NY. After that we transform such categorical data in numerical in order to predict which cluster in NY might have more revenue potential for our client.

2.2 Data Preparation

The data acquisition come as JSON. So, we need to get the venues information for each neighborhood in Toronto and NY, and it is done using latitude and longitude for each neighborhood. After that, it is created data frames that have venues names and their categories. Then, another data frame that contains venue categories for each neighborhood is created, however such data is categorical, and it is transformed in numerical with dummy variables, and the mean of their quantities is calculated. Such data is used in the unsupervised part for clustering Toronto and NY neighborhoods.

After that, each neighborhood is labeled with the cluster that it belongs to. So, it is created another data frame for each cluster, containing the 10 most common types of venues (categories) for each neighborhood in that cluster. This information is categorical, so another transformation for the clusters table is made, by grouping the categories in such cluster, so we can identify the more recurrent categories in that cluster. The final data frame is the number of categories (rows) by most common (columns), in order words, the quantity of categories is spatial ranked, since this was made during the clustering process.

In order to train the models, some data preparation is made, splitting the data into training set and test set. In prediction, we need to get just the categories from NY clusters that we also have in Toronto's cluster. And an important variable is chosen: a threshold for the Toronto cluster that is our client best location/cluster. We used descriptive statistics to choose such threshold, and then a "y" variable is created which classifies the kind of categories that are target for our client – from which more delivery transactions are created. This variable is fundamental in order to classify NY cluster venues, and reach the number of venues that are interesting to our client – potential clients of our client.

3. Unsupervised: clustering

This is the spatial aspect of the work. First, we created the Toronto's clusters. We used K-means in this task. 7 clusters were created. And the cluster 3 is our client best cluster.

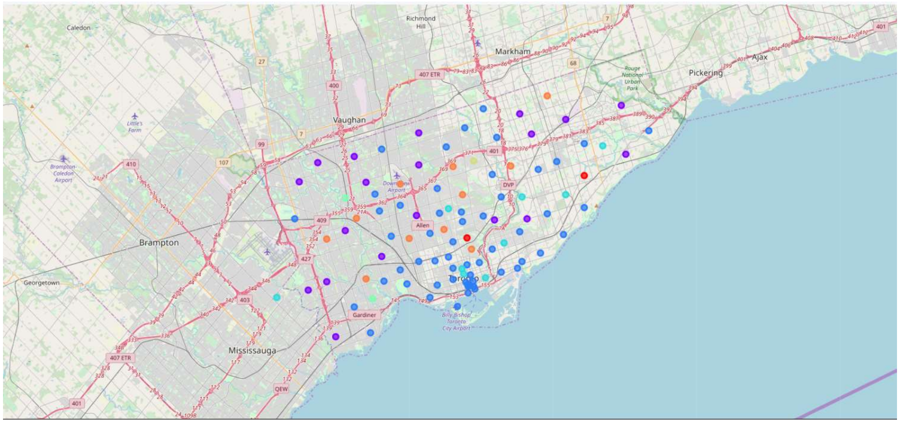

Figure 1. Toronto Clusters

The same procedure is done for NY, but we used 8 clusters instead of 7, since the results seemed better (clusters appeared to be more balanced in terms of quantities of neighborhood and venues).

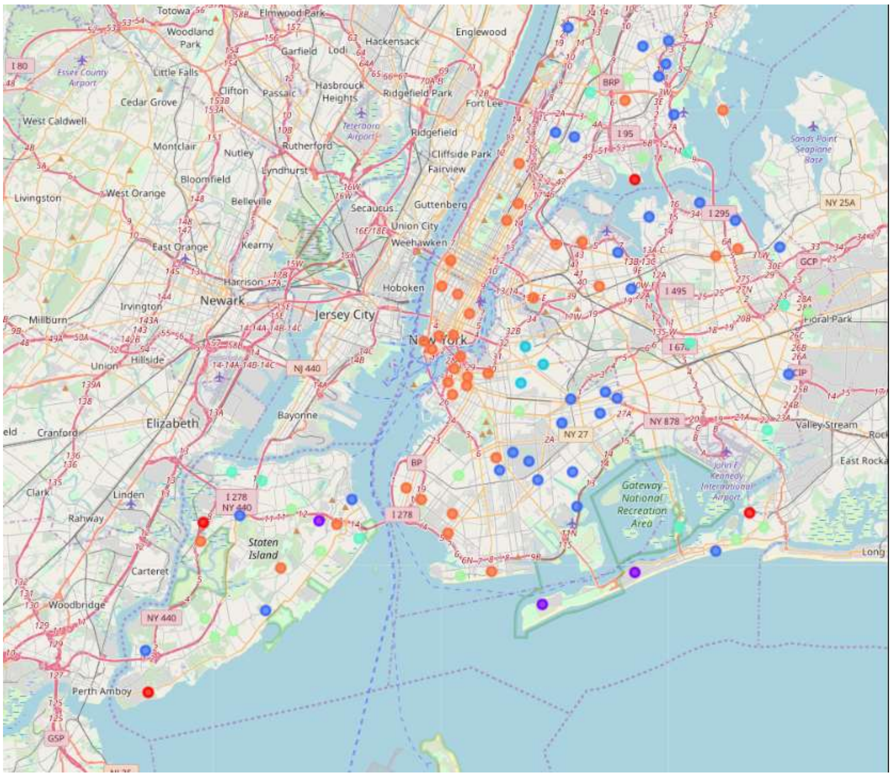

Figure 2. NY Clusters

In K-means, the number of clusters is heuristic. However, we tested different numbers for K, but the ones who brought more balanced clusters were the ones we chose.

4. Exploratory Analysis

The data that comes from the clusters gives some insights about the kind of venues in each cluster. Cluster 3 in Toronto is our client best cluster.

Table 1. Neighborhood and Venues in Toronto's cluster 3

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Highland Creek,Rouge Hill,Port Union	Construction & Landscaping	Bar	Yoga Studio	Dessert Shop	Event Space	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Drugstore	Dog Run
2	Guildwood,Morningside,West Hill	Electronics Store	Mexican Restaurant	Breakfast Spot	Pizza Place	Medical Center	Intersection	Rental Car Location	Spa	Tech Startup	Eastern European Restaurant
4	Cedarbrae	Caribbean Restaurant	Lounge	Bakery	Hakka Restaurant	Fried Chicken Joint	Thai Restaurant	Athletics & Sports	Bank	Diner	Dog Run
7	Clairlea,Golden Mile,Oakridge	Bus Line	Bakery	Park	Intersection	Fast Food Restaurant	Metro Station	Bus Station	Soccer Field	Creperie	Cuban Restaurant
8	Cliffcrest,Cliffside,Scarborough Village West	Motel	American Restaurant	Yoga Studio	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Drugstore	Dog Run	Discount Store	Diner
9	Birch Cliff,Cliffside West	College Stadium	General Entertainment	Skating Rink	Café	Deli / Bodega	Electronics Store	Eastern European Restaurant	Drugstore	Dog Run	Discount Store
10	Dorset Park,Scarborough Town Centre,Wexford He...	Indian Restaurant	Latin American Restaurant	Pet Store	Vietnamese Restaurant	Chinese Restaurant	Deli / Bodega	Electronics Store	Eastern European Restaurant	Drugstore	Dog Run
11	Maryvale,Wexford	Middle Eastern Restaurant	Shopping Mall	Sandwich Place	Bakery	Auto Garage	Breakfast Spot	Yoga Studio	Empanada Restaurant	Electronics Store	Eastern European Restaurant
12	Agincourt	Breakfast Spot	Chinese Restaurant	Sandwich Place	Lounge	Department Store	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Drugstore	Dog Run
17	Hillcrest Village	Golf Course	Mediterranean Restaurant	Pool	Dog Run	Yoga Studio	Dance Studio	Eastern European Restaurant	Drugstore	Discount Store	Diner
18	Fairview,Henry Farm,Oriole	Tea Room	Movie Theater	Smoothie Shop	Shopping Mall	Juice Bar	Burger Joint	Fast Food Restaurant	Bakery	Department Store	Candy Store
19	Bayview Village	Café	Chinese Restaurant	Bank	Japanese Restaurant	Department Store	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Drugstore	Dog Run
22	Willowdale South	Ramen Restaurant	Café	Pet Store	Indonesian Restaurant	Movie Theater	Plaza	Shopping Mall	Fast Food Restaurant	Steakhouse	Japanese Restaurant
26	Don Mills North	Café	Gym / Fitness Center	Baseball Field	Japanese Restaurant	Caribbean Restaurant	Department Store	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Drugstore
27	Flemingdon Park,Don Mills South	Gym	Dim Sum Restaurant	Italian Restaurant	Japanese Restaurant	Discount Store	Beer Store	Bike Shop	Asian Restaurant	Sporting Goods Shop	Clothing Store
32	Downsview Central	Food Truck	Korean Restaurant	Baseball Field	Home Service	Yoga Studio	Empanada Restaurant	Electronics Store	Eastern European Restaurant	Drugstore	Dog Run
36	Woodbine Heights	Pharmacy	Athletics & Sports	Cosmetics Shop	Curling Ice	Bus Stop	Skating Rink	Beer Store	Park	Furniture / Home Store	Dance Studio
37	The Beaches	Other Great Outdoors	Health Food Store	Trail	Pub	Yoga Studio	Dance Studio	Eastern European Restaurant	Drugstore	Dog Run	Discount Store
38	Leaside	Gym	Sushi Restaurant	Grocery Store	Fish & Chips Shop	Liquor Store	Coffee Shop	Clothing Store	Restaurant	Burger Joint	Smoothie Shop
39	Thorncliffe Park	Indian Restaurant	Yoga Studio	Housing Development	Pharmacy	Pizza Place	Discount Store	Sandwich Place	Burger Joint	Supermarket	Intersection
41	The Danforth West,Riverdale	Greek Restaurant	Ice Cream Shop	Italian Restaurant	Brewery	Fruit & Vegetable Store	Dessert Shop	Cosmetics Shop	Pizza Place	Pub	Yoga Studio
42	The Beaches West,India Bazaar	Gym	Sushi Restaurant	Ice Cream Shop	Fish & Chips Shop	Italian Restaurant	Fast Food Restaurant	Liquor Store	Movie Theater	Park	Pub
45	Davisville North	Gym	Breakfast Spot	Hotel	Food & Drink Shop	Park	Clothing Store	Sandwich Place	Grocery Store	American Restaurant	Art Gallery
47	Davisville	Dessert Shop	Italian Restaurant	Sushi Restaurant	Park	Coffee Shop	Pizza Place	Seafood Restaurant	Sandwich Place	Café	Thai Restaurant
51	Cabbagetown,St. James Town	Café	Pet Store	Taiwanese Restaurant	Gastropub	Indian Restaurant	Italian Restaurant	Japanese Restaurant	Jewelry Store	Diner	Pub
52	Church and Wellesley	Tea Room	Theme Restaurant	Breakfast Spot	Bookstore	Juice Bar	Diner	Salon / Barbershop	Restaurant	Ramen Restaurant	Dance Studio
54	Ryerson,Garden District	Café	Burger Joint	Ramen Restaurant	Tea Room	Burrito Place	Thai Restaurant	Theater	Plaza	Pizza Place	Movie Theater
55	St. James Town	Coffee Shop	Gastropub	Japanese Restaurant	BBQ Joint	Food Truck	Italian Restaurant	Middle Eastern Restaurant	Creperie	Cosmetics Shop	Church
56	Berczy Park	Farmers Market	Coffee Shop	Museum	Liquor Store	Seafood Restaurant	Steakhouse	Fish Market	Thai Restaurant	French Restaurant	Breakfast Spot
58	Adelaide,King,Richmond	Steakhouse	Coffee Shop	Pizza Place	Speakeasy	Bar	Hotel	Plaza	Asian Restaurant	Food Court	Seafood Restaurant
59	Harbourfront East,Toronto Islands,Union Station	Hotel	Plaza	Bubble Tea Shop	Sporting Goods Shop	Supermarket	Salad Place	Deli / Bodega	Bakery	Skating Rink	Café
60	Design Exchange,Toronto Dominion Centre	Coffee Shop	Café	Gym	Hotel	Beer Bar	Japanese Restaurant	Restaurant	Bakery	Pub	Deli / Bodega

Counting the numbers of categories in the clusters give us the kind of venues that bring more delivery transactions. For cluster 3, café, coffee shop and restaurants have more frequency.

Table 2. Toronto's cluster 3 category frequencies

	categ	clot3
33	Café	23.0
42	Coffee Shop	18.0
62	Eastern European Restaurant	16.0
138	Restaurant	14.0
60	Dog Run	14.0
15	Bakery	13.0
63	Electronics Store	13.0
61	Drugstore	13.0
169	Yoga Studio	12.0
104	Italian Restaurant	12.0
17	Bar	12.0
59	Discount Store	11.0
90	Gym	10.0
135	Pub	10.0
105	Japanese Restaurant	9.0

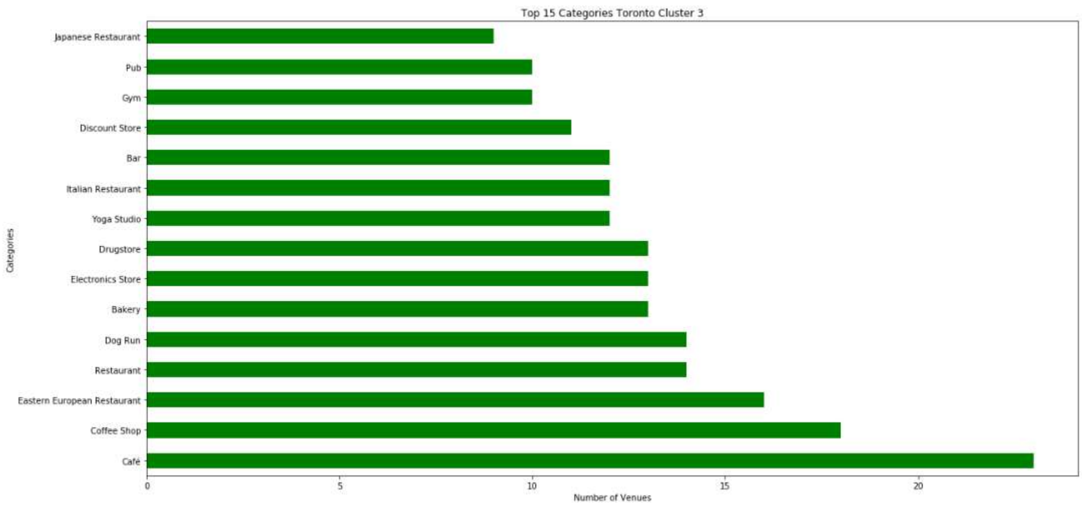

Figure 3. Toronto's Cluster 3 top 15th venue categories.

The data is very concentrated around 3. And this is an important information in order to decided the threshold to classify venue categories in cluster 3 that presents best potential to our client.

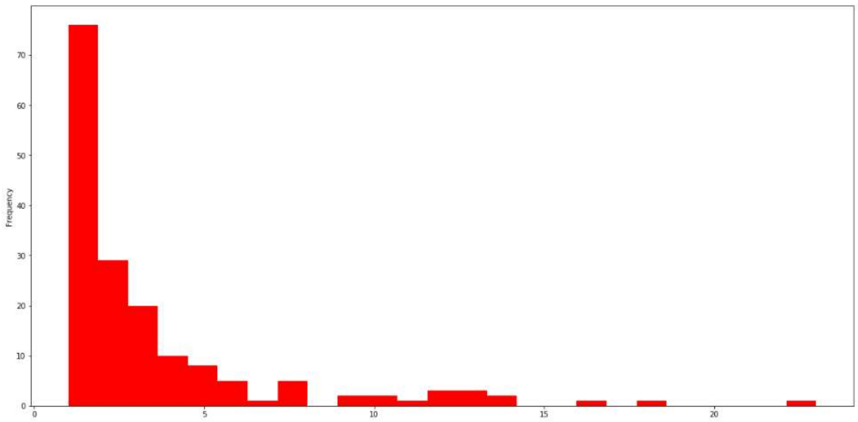

Figure 4. Toronto's Cluster 3 histogram.

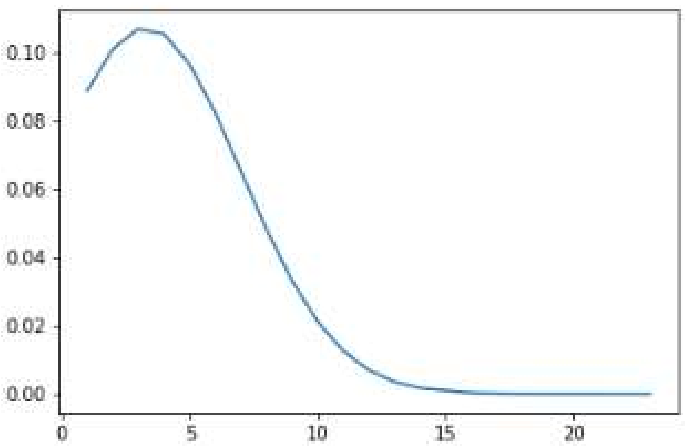

Figure 5. Toronto’s Cluster 3 pdf histogram.

By looking such graphs, 3 is the threshold that might suit better to our purpose.

Table 3. Toronto’s cluster 3 descriptive statistics

	clot3
count	170.000000
mean	3.294118
std	3.736351
min	1.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	23.000000

The data is also very sparse, with outliers that categorizes best the places where our client business have more success: related to food.

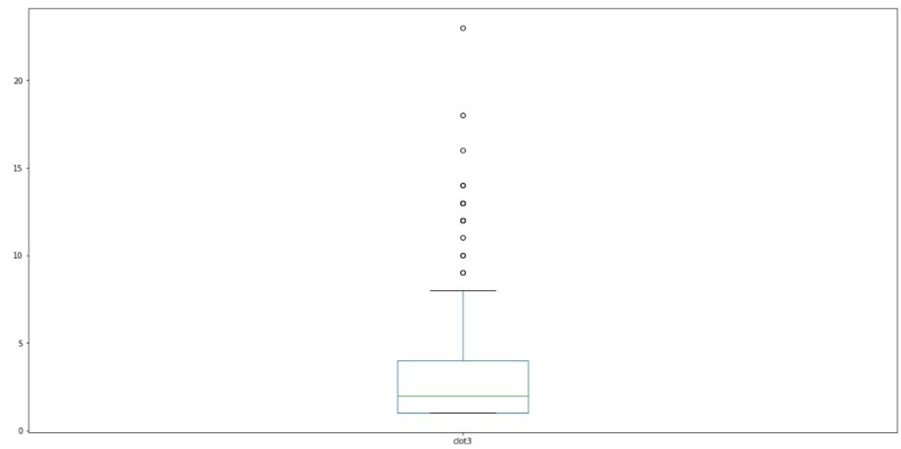

Figure 4. Toronto’s Cluster 3 boxplot.

From the 8 clusters in NY, only 3 have a considerable number of neighborhoods. So, our analysis will be concentrated in these 3: NY cluster 2, NY cluster 4 and NY cluster 8. A visualization of categories venues shows first insight that cluster 2 might have more potential.

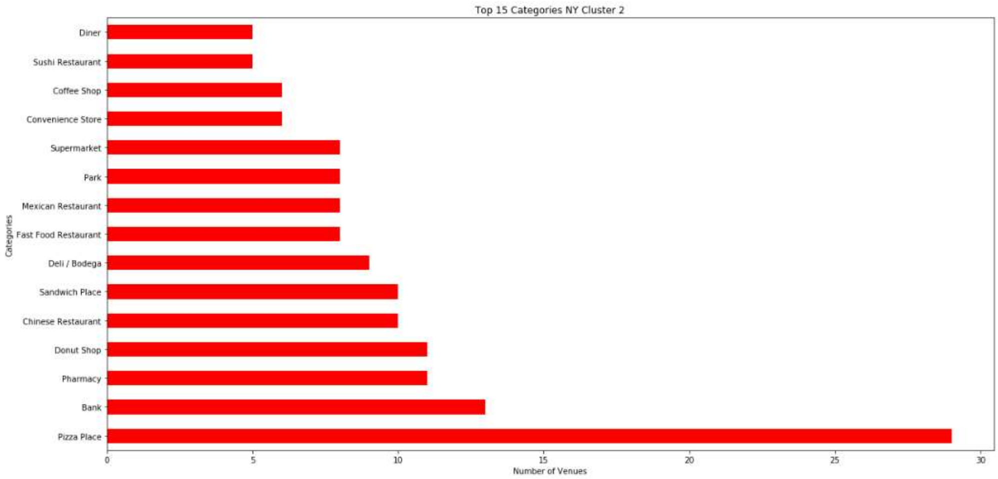

Figure 5. NY's Cluster 2 top 15^{th} venue categories.

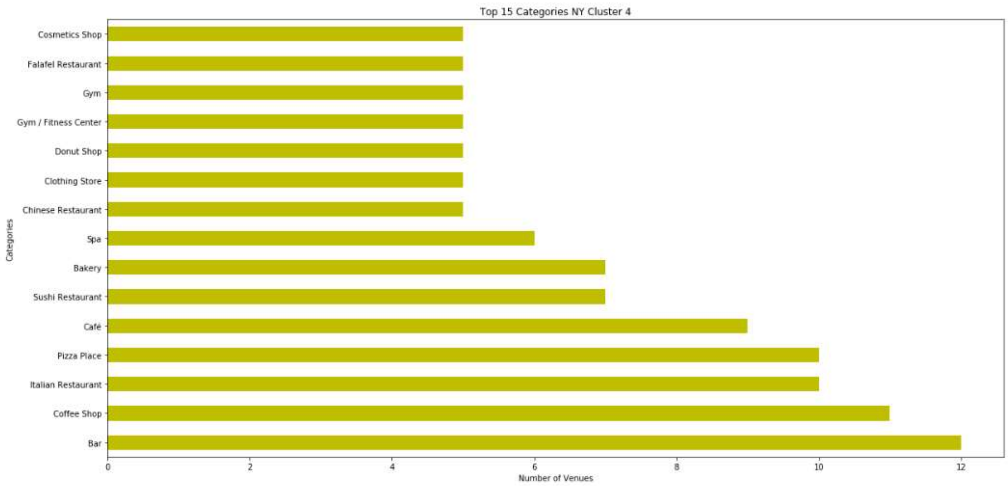

Figure 6. NY's Cluster 4 top 15^{th} venue categories.

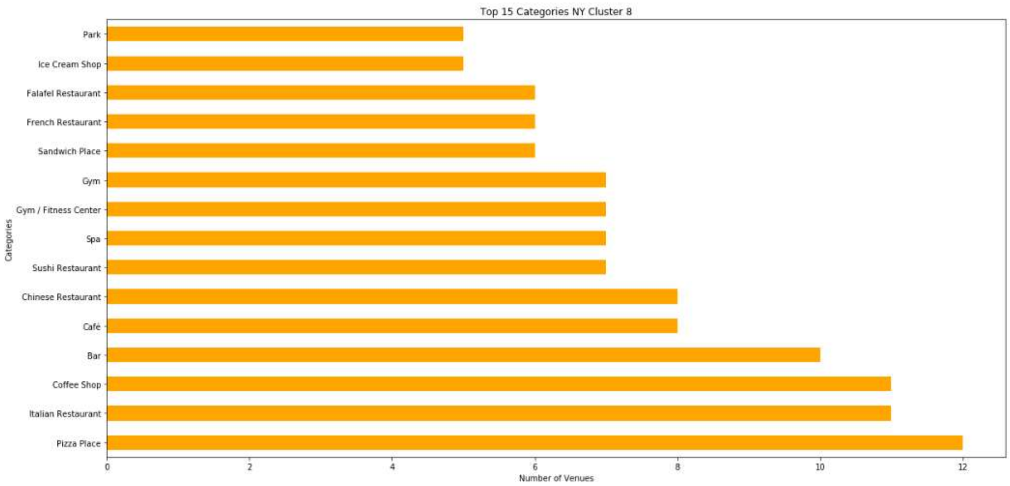

Figure 7. NY's Cluster 8 top 15^{th} venue categories.

However, the other 2 NY clusters have good frequencies for venue categories that are very interesting to our client, and a deeper analysis needs to take in account all venue categories arrangements, as well as the number of venues.

5. Methodology

5.1 Models

We decided to use two kind of models (machine learning): logistic regression and decision tree. Both are well-known estimation methods for classification problems. Particularly, decision tree is recognized to be good with categorical data, and logistic regression is very traditional when it comes to statistical analysis and modeling. But using two methods give us possibility to verify if the point to the direction, in other words if they present similar or not similar results.

We created the “y” variable for Toronto’s cluster 3 with threshold in 3: if a category venue has a frequency higher than 3, that it is a potential client to our client (a good proxy for delivery transactions: places with certain type of venues have more delivery transactions than others). This variable was appended to Toronto’s cluster 3 data frame in which contains venues categories frequencies by the 10 most commons (columns).

Therefore, we have a classification problem. Two type of models will help us in compare results from the training and testing part, as well the prediction part, when we classify the NY 3 clusters in order to verify which has more potential to our client.

Although we do not have too many data, the data set was split in training and test, for both models. After training and testing, we applied the models with NY’s cluster data.

5.2 Model Evaluation

Logistic regression presented better performance than the decision tree with the testing data set. Jaccard index (how many classifications were correct) for LR was 94.11% and for Decision tree was 78,31%. Logistic regression R-square was 0,698, a good result taking in account the kind and size of data that we are working with.

Table 4. Logistic Regression performance metrics

	precision	recall	f1-score	support
0.0	0.92	1.00	0.96	23
1.0	1.00	0.82	0.90	11
micro avg	0.94	0.94	0.94	34
macro avg	0.96	0.91	0.93	34
weighted avg	0.95	0.94	0.94	34

Table 5. Decision Tree performance metrics

	precision	recall	f1-score	support
0.0	0.72	1.00	0.84	28
1.0	1.00	0.52	0.69	23
micro avg	0.78	0.78	0.78	51
macro avg	0.86	0.76	0.76	51
weighted avg	0.85	0.78	0.77	51

F1-Score and the other metrics were better in Logistic Regression as well. And taking in account the confusion matrix of both models, we can realize that both classification models are good in classifying non potential clients instead of potential clients.

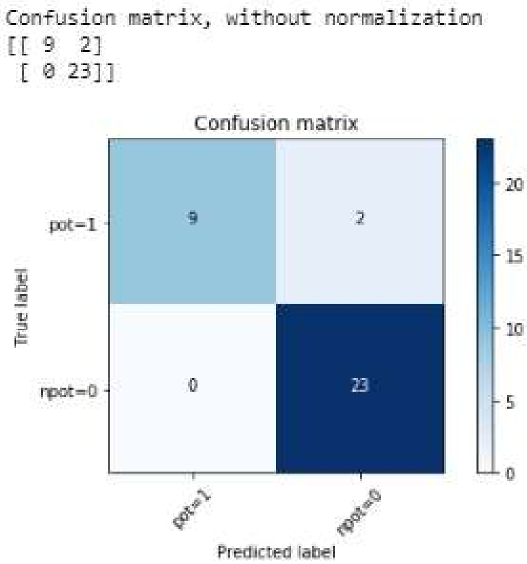

Figure 8. Logistic Regression Confusion Matrix.

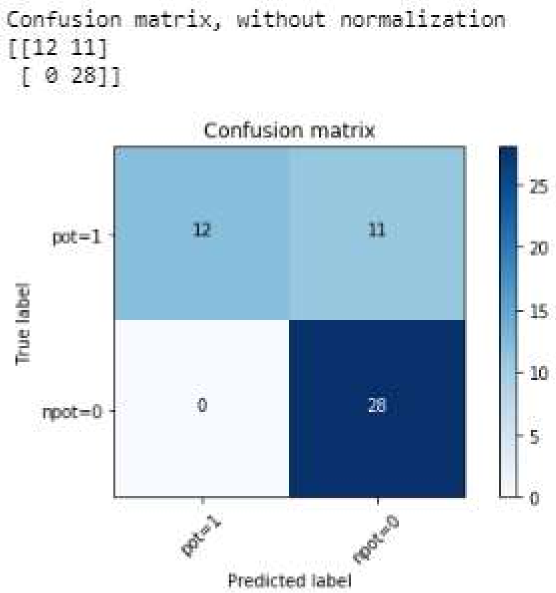

Figure 9. Decision Tree Confusion Matrix.

6. Results

After training and testing the models we used the data set for NY clusters: 2, 4 and 8. Then, the models classified, using each cluster data set, the categories venues that are interesting in terms of clients to our client – more potential to bring delivery transactions, and therefore revenue. The results were not so different between models and between NY clusters.

Table 5. Predictions for NY Clusters using Logistic Regression Model

	NY Cluster2	NY Cluster4	NY Cluster8
Proportion adherent	29.4	26.0	23.0
Cluster Size	340.0	340.0	350.0
Potential clients	100.0	88.0	82.0

Table 6. Predictions for NY Clusters using Decision Tree Model

	NY Cluster2	NY Cluster4	NY Cluster8
Proportion adherent	21.2	23.1	17.1
Cluster Size	340.0	340.0	350.0
Potential clients	72.0	79.0	82.0

Their proportion of adherent venue categories (classified as good for our client, y = 1) are similar, around 25%, just for cluster 8 in the Decision Tree model we see a number quite distant from the others. However, the size (number of venues) are needed in order to decide which NY Cluster is more interesting to our client business. Taking that in account, cluster 2 presents the best result in the Logistic Regression model. In the Decision Tree model, cluster 8 is the best one, but it is because it is bigger than others, since it has the lower classification rate.

7. Conclusion

In this work, we analyzed spatial data from Foursquare about neighborhood and venue categories in order to identify the best location for a start-up (delivery business) based on Toronto to begin its expansion in NY. We created clusters for Toronto and for NY (unsupervised), and from the best cluster for our client in Toronto we predict the best cluster in NY for our client (supervised).

From the results above, cluster 2 seems to be the best choice for our client begin its expansion in NY. This cluster had the best results in the best model. But cluster 4 is a good choice as well, since its results are not so far from those of cluster 2. Perhaps it could be a second phase of our client's expansions in NY.

8. Future Directions

We used just Foursquare data. A good improvement could be gathering other kind of data, like socioeconomic data for example. Such information can change the results that we found, bring more information. Other improvement could be testing SVM and Neural Network models in order to compare results. Moreover, the value of the threshold is something that could change results if it is changed, as well as the number of clusters for each city.