Effects of Restaurant Neighborhood and Cuisine on Rating

Pradeep Gopinathan

July 12, 2020

1. **Introduction**

1.1 Background

Urban areas are some of the hardest areas for new restaurants to be built in, as the costs of doing business are high, and the competition is fierce. However, the results of success can be quite lucrative. One important factor to consider is the population of the neighborhood, and what kinds of cuisine they do or do not want. Would-be restauranteurs need to consider if the neighborhood they are scouting as a location has an actual demand for what they are selling. Social media sites like Foursquare can help us understand this demand by showing current restaurants in the neighborhood, and what visitors have rated them as.

1.2 Problem

Our goal for this analysis is to attempt to build a predictive model that can see what rating a restaurant might get depending on the neighborhood it is located in and its type of cuisine. For this preliminary research, we are limiting ourselves to just downtown Toronto and the surrounding boroughs. Foursquare ratings are given as rankings between 1-10, so we will treat the problem as a classification one.

1.3 Interest

Our hope is to give restauranteurs a better understanding of what the population of each neighborhood is looking for in a restaurant, and where the food they hope to serve will be most welcomed.

2. Data

2.1 Data Source

For this preliminary investigation, we will limit ourselves to data accessed by the Foursquare API. First, we will collect a list of restaurants within each neighborhood of Toronto. Afterwards, we will collect the rating and the number of tips for each restaurant through the Foursquare API.

Foursquare's API allows developers to search for venues a certain distance away from a specified latitude and longitude. We will collect our data based of looking for all venues .8km away from the latitude/longitude of each neighborhood center (which was collected prior to this analysis), and will filter for only venues whose Foursquare assigned category is Food or some sub-category of Food.

2.2 Data Processing

First, we will have to de-dup the data collected, to ensure venues that are close enough to multiple neighborhood centers are not double counted. Then, we will filter our all venues that Foursquare could not provide a rating for.

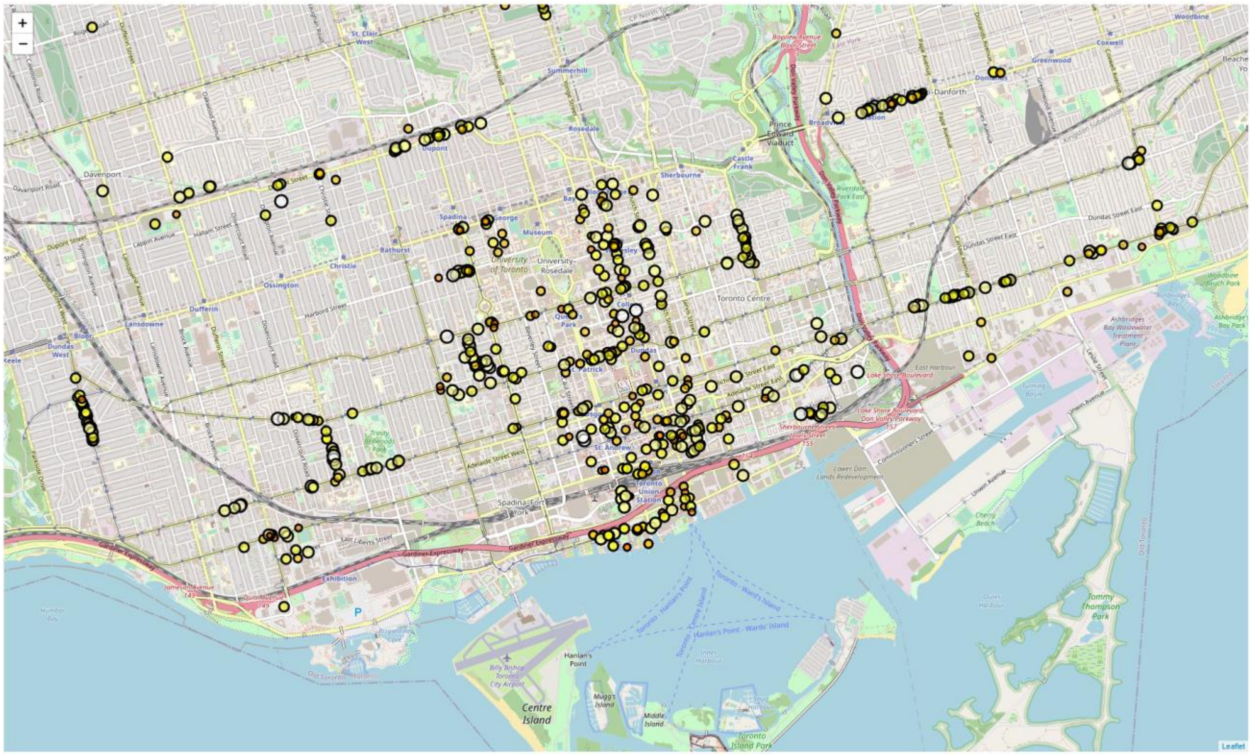
Rating data is often provided as an average of user ratings. To turn this value into a classification, we will convert each rating provided into an integer between 1-10 by rounding the number given.

Finally, we will use one-hot encoding to turn the neighborhood and category of each venue into a list of binary features usable to machine learning, and these encoded features will make up the feature set used by our model.

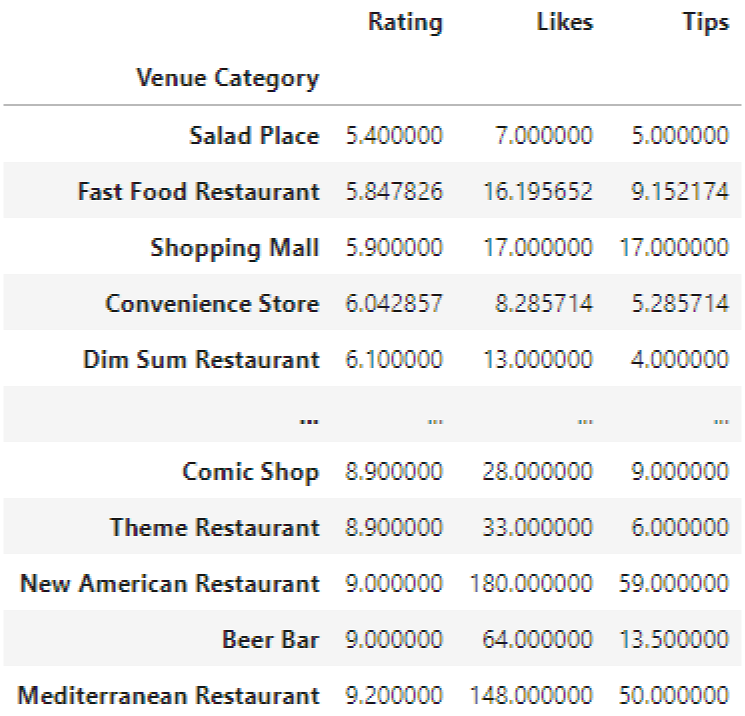
3. Methodology

3.1 Principal Analysis

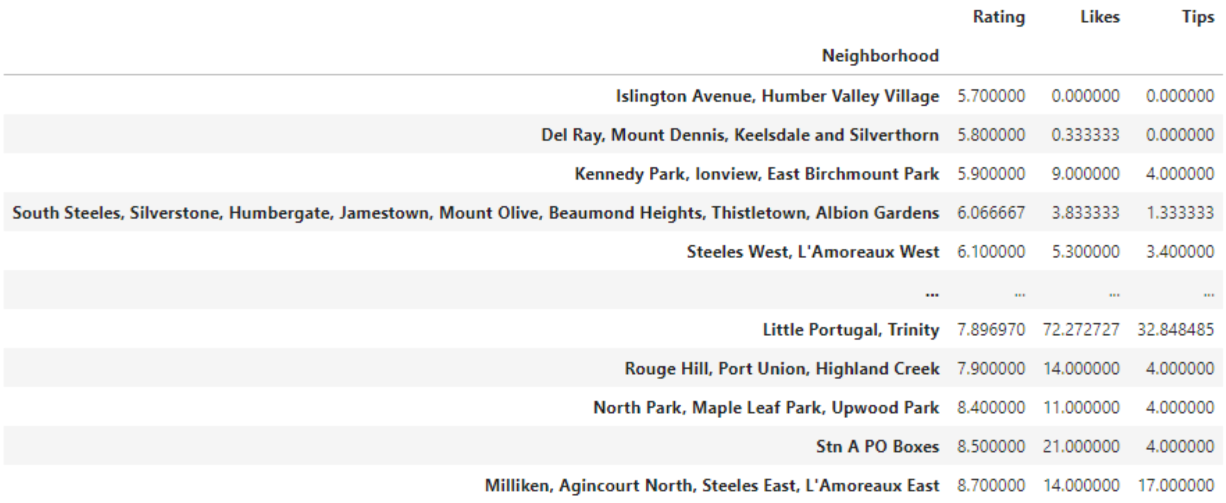
First, we mapped every restaurant with a rating. The larger and darker circles are the higher rated restaurants. This initial map doesn't show much of a relation, but does show that most of the restaurants with ratings are focused on the Toronto downtown.



Next, we examined the average ratings by neighborhood and category:



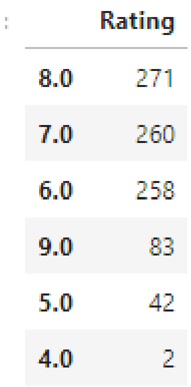
	Rating	Likes	Tips
Venue Category			
Salad Place	5.400000	7.000000	5.000000
Fast Food Restaurant	5.847826	16.195652	9.152174
Shopping Mall	5.900000	17.000000	17.000000
Convenience Store	6.042857	8.285714	5.285714
Dim Sum Restaurant	6.100000	13.000000	4.000000
...	...	...	...
Comic Shop	8.900000	28.000000	9.000000
Theme Restaurant	8.900000	33.000000	6.000000
New American Restaurant	9.000000	180.000000	59.000000
Beer Bar	9.000000	64.000000	13.500000
Mediterranean Restaurant	9.200000	148.000000	50.000000



	Rating	Likes	Tips
Neighborhood			
Islington Avenue, Humber Valley Village	5.700000	0.000000	0.000000
Del Ray, Mount Dennis, Keelsdale and Silverthorn	5.800000	0.333333	0.000000
Kennedy Park, Ionview, East Birchmount Park	5.900000	9.000000	4.000000
South Steeles, Silverstone, Humbergate, Jamestown, Mount Olive, Beaumond Heights, Thistletown, Albion Gardens	6.066667	3.833333	1.333333
Steeles West, L'Amoreaux West	6.100000	5.300000	3.400000
...	...	...	...
Little Portugal, Trinity	7.896970	72.272727	32.848485
Rouge Hill, Port Union, Highland Creek	7.900000	14.000000	4.000000
North Park, Maple Leaf Park, Upwood Park	8.400000	11.000000	4.000000
Stn A PO Boxes	8.500000	21.000000	4.000000
Milliken, Agincourt North, Steeles East, L'Amoreaux East	8.700000	14.000000	17.000000

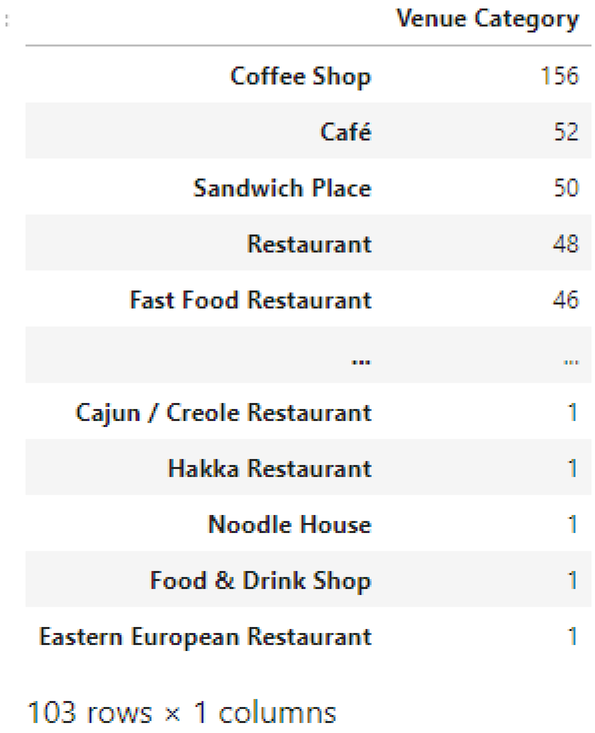
There does appear to be variation in rating for Venues and Neighborhoods. Mediterranean restauarants are highly rated, and Agincourt North, Steeles East, Milliken, and L'Amoreaux East appear to be good neighborhoods to be located.

After converting the ratings into classes, we examined the general distribution of the classes. Restaurant ratings appear to be focused around the 6/7 out of 10 mark.



	Rating
8.0	271
7.0	260
6.0	258
9.0	83
5.0	42
4.0	2

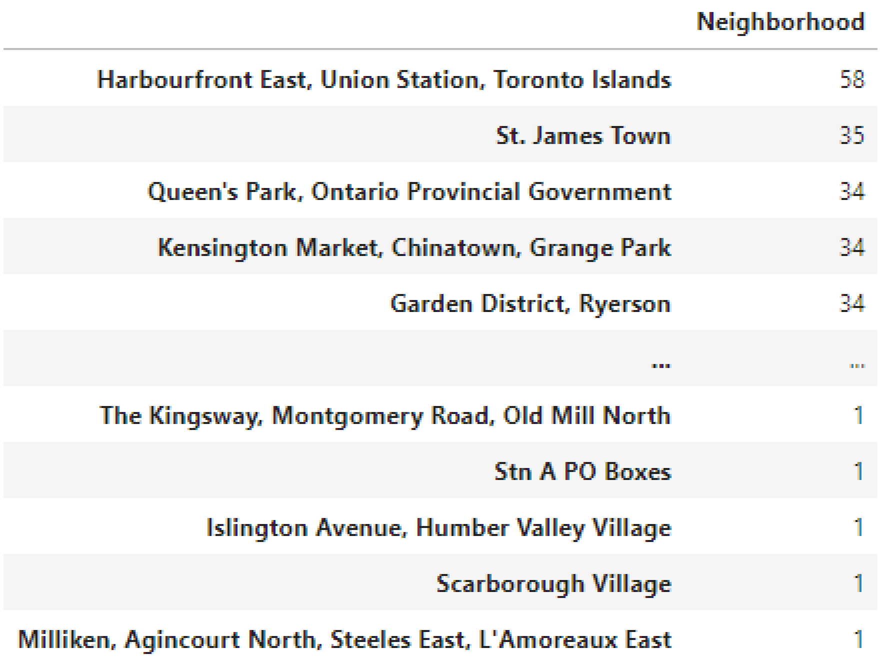
The distribution of venue categories is also heavily skewed, with coffee shops being overly represented by a wide margin, and other cuisines (like Cajun) only having a single representative.



	Venue Category
Coffee Shop	156
Café	52
Sandwich Place	50
Restaurant	48
Fast Food Restaurant	46
...	...
Cajun / Creole Restaurant	1
Hakka Restaurant	1
Noodle House	1
Food & Drink Shop	1
Eastern European Restaurant	1

103 rows × 1 columns

Distribution of neighborhoods in venues is equally skewed. As it turns out, the highly performing neighborhood is actually just a single highly rated restaurant.



	Neighborhood
Harbourfront East, Union Station, Toronto Islands	58
St. James Town	35
Queen's Park, Ontario Provincial Government	34
Kensington Market, Chinatown, Grange Park	34
Garden District, Ryerson	34
...	...
The Kingsway, Montgomery Road, Old Mill North	1
Stn A PO Boxes	1
Islington Avenue, Humber Valley Village	1
Scarborough Village	1
Milliken, Agincourt North, Steeles East, L'Amoreaux East	1

3.2 Predictive Models

We will examine two modelling types, decision trees and k-nearest neighbors. But first, we considered a few naïve approaches. One is to just randomly guess the rating classification between 1-10. We also guessed the average class of the restaurants, which is a class 7, and the most common class, which is 8.

For the decision trees, we examined the effects of gini vs entropy information gain calculations and max tree depth on accuracy. For kNN, we used up to 20 neighbors, and looked at the effects of weighting contribution to the classification uniformly or based on distance. We used 5-fold cross validation to test each parameter set, and chose the one with the best average over the 5 folds.

4. Results

We found a slight improvement of the two predictive models over the naïve approaches:

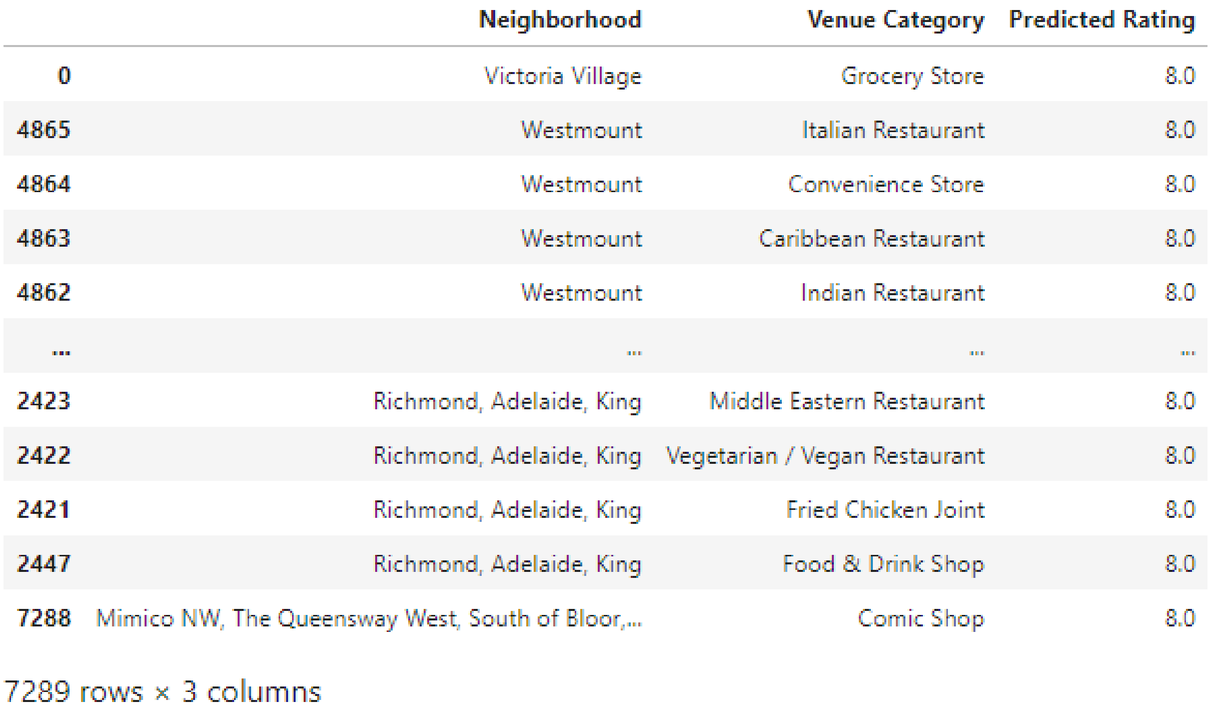
Approach	Accuracy	Optimal Parameters
Naïve Random (guess 1-10)	.10	N/A
Naïve Average (guess 7)	.283	N/A
Naïve Average (guess 8)	.296	N/A
Decision Tree	.363	Gini with max depth 2
k-Nearest Neighbors	.305	Uniform weight with 19 neighbors

5. Discussion

5.1 Predictions

Most Venue Category/Neighborhood pairings did not have a venue in the sample data. We can try to use the created models to predict what would be classifications of these pairings.

Both models, however, predicted 8’s exclusively for each category/neighborhood pairing.



	Neighborhood	Venue Category	Predicted Rating
0	Victoria Village	Grocery Store	8.0
4865	Westmount	Italian Restaurant	8.0
4864	Westmount	Convenience Store	8.0
4863	Westmount	Caribbean Restaurant	8.0
4862	Westmount	Indian Restaurant	8.0
...	...	...	...
2423	Richmond, Adelaide, King	Middle Eastern Restaurant	8.0
2422	Richmond, Adelaide, King	Vegetarian / Vegan Restaurant	8.0
2421	Richmond, Adelaide, King	Fried Chicken Joint	8.0
2447	Richmond, Adelaide, King	Food & Drink Shop	8.0
7288	Mimico NW, The Queensway West, South of Bloor,...	Comic Shop	8.0

7289 rows × 3 columns

5.2 Limitations/Concerns

These issues likely arise from the signification limitations of the data. First, of the around 2000 restaurants we found near enough to the neighborhoods, only about 950 had ratings. Preliminary data analysis showed that these ratings we also heavily concentrated around the ratings of 6,7,8, with no venues with a rating below 4. Without exemplars below 4, it will likely be impossible our models to predict a poor performing restaurant. Also, it is likely the models themselves are just guessing 8 based off how common the rating is.

In addition, most venue/neighborhood pairings do not have exemplars in the training dataset, so our predictive models were built off a set that did not reflect many of the possible scenarios. Without this information, it is hard for the models to see the specific impact and type of venue or the specific neighborhood may have.

6. Conclusion

The results of the accuracy test suggest that it is possible to build a model that can determine the rating of a restaurant based of where it is located and what type of cuisine it serves. The decision tree model was slightly more accurate that just using a statistically inferred model. However, as discussed in 5.2 the data is very sparse, and the accuracy gain is limited, so we cannot be sure that a predictive model is possible.

Before presenting these models as useful in a business sense, or definitively concluding that no relationship exists, we must first expand the data analysis. Two paths to expand should be taken.

First, we need more rating data. While Foursquare contains rating data, it is not a site dedicating to rating businesses. We should look to integrate data from sources like Yelp to expand the number of venues with ratings.

Next, we should attempt to collect more examples of each venue category. This is more difficult as our current investigation is limited to Toronto venues. However, we could attempt to expand beyond a single city. Instead of building a model based off the neighborhood itself, we could create a profile of each neighborhood based off certain kinds of data (age distribution, per-capita income, etc.). This profile could be created for other cities, and thus a more expansive training set with better coverage can be created.

In short, we are a long way from creating predictive models of restaurant ratings usable in a real-world scenario.