BTMA 559.03/797 FINAL PROJECT REPORT
Professor Hooman Hidaji

Michael Chui
Gabriel Guevara
Andrew Ly

Problem Formulation

Customers are increasingly looking to online review to help guide their meal decisions. A restaurant's low online rating can negatively impact customer preferences (Campbell, 2015) and financial results (Ding et al., 2017). The objective of this project is to assess if combining undergraduate level R skills and free online rating can be a useful tool for small chains to better a) differentiate its branch performances, b) provide operational insights, or c) assess likelihood of their store location's survival. We applied association rules, classification tree, multivariate regression, and logistics regression on free Yelp online data from Kaggle.

Data

We used a subset of Yelp's businesses, user reviews, and attributes data originally for Yelp's Dataset CHallenge made available on Kaggle for education purposes. The datasets contain over 150,000 business within 11 metropolitans areas in Canada and United States. Our original scope focused on textual analysis of user reviews which contained over 5.2 million user reviews. We ultimately limited our sample to summarized reviews embedded within the business information due to computational limitation.

Our main datasets are "yelp_business.csv" (business details) and "yelp_attributes.csv" (business attributes) from https://www.kaggle.com/yelp-dataset/yelp-dataset/data. The business details dataset contains approximately 170,000 unique stores along with name of the company, average ratings, number of ratings, the restaurant geographical details and a unique business identification (business id). The yelp_attributes.csv dataset provides categorical details of the store's (based on business id) physical and service offerings, ranging from a happy hour, to wheelchair accessible, to catering options, to children friendliness. We limited the attributes to factors conceived as "convenience" (Kincaid et. al., 2010) like accessibility, parking options, and those we have personally considered as important like offering Happy hours and allow dogs.

Exploratory Data Analysis

The combined dataset is filtered for a) food and restaurant, b) brands with more than 5 locations, and c) located in Canada (i.e. Quebec and Ontario). We first removed all business details outside of Quebec and Ontario. To simplify the computational process, we removed redundant and non-essential fields in the business detail: for example, city was used as the lowest level denominator. We further used the ifelse() function and str_detect(), to narrow down to restaurant and food businesses. Lastly, we utilize the count() and subset() function to limit brands with >=5 occurrences before combining the two datasets via business id. A few noteworthy issue is concentration of data in Toronto, accounting for 38% of the sample (n=1,483), and most rating landed around 3-stars.

Table 1. Dataset and variables (N= 3,865)

Name	**Variable type**	**Name**	**Variable type**
business_id	factor	name	Factor, food brands
city	factor	state	factor
stars	Numerical, average rating	is_open	binary/integer, 0 for closed, 1 for open
review_count	Integer, number of	categories	factor, generic

	online users		description
BusinessParking_ street	factor, TRUE for street parking	BusinessParing_ validated	factor, TRUE if parking is validated
Wheelchair Accessible	factor, TRUE if accessible	HappyHour	factor, TRUE if Happy hour offered
DogsAllowed	factor, TRUE if dogs allowed		

Table 1. Cumulative total of sample (n=3,865) by brand sizes

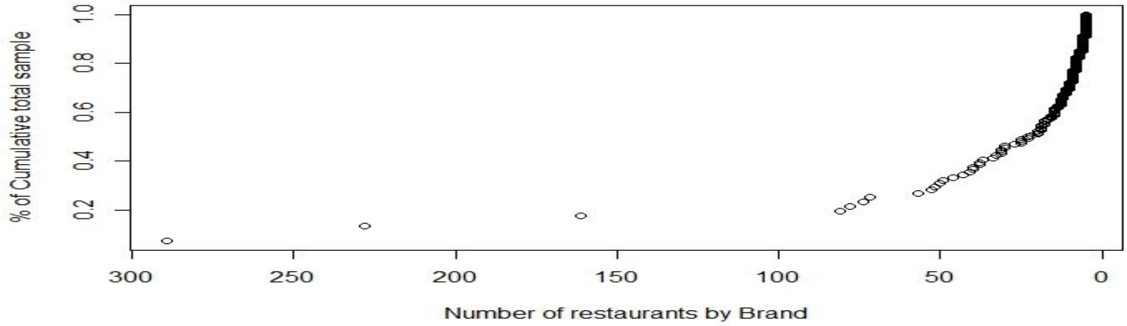
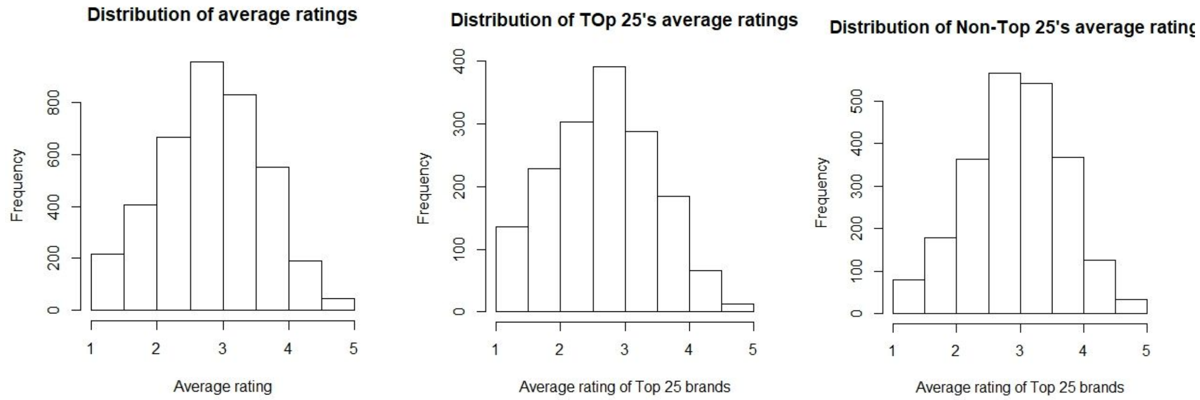


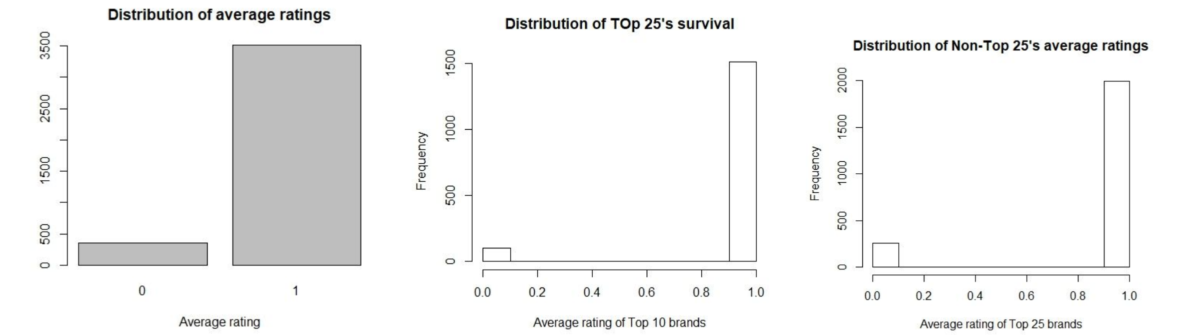
Table 2. Top 25 Brands with Store counts

Rank	Brand	n	Rank	Brand	n
1	"Starbucks" (=15.7% of N)	289	14	"Harvey's Restaurants"	49
2	"Tim Hortons" (=12.4% of N)	228	15	"Thai Express"	46
3	"McDonald's" (=8.8% of N)	161	16	"Domino's Pizza"	43
4	"Shoppers Drug Mart"	91	17	"Wild Wing"	42
5	"Second Cup"	81	18	"Boston Pizza"	40
6	"Pizza Pizza"	79	19	"Canadian Tire"	40
7	"Subway"	74	20	"Pizza Nova"	40
8	"Swiss Chalet Rotisserie & Grill"	72	21	"A&W"	38
9	"GoodLife Fitness"	67	22	"Sunset Grill"	38
10	"Popeyes Louisiana Kitchen"	57	23	"Metro"	37
11	"Tim Horton's"	53	24	"Aroma Espresso Bar"	34
12	"LCBO"	52	25	"Dollarama"	34
13	"Pizza Hut"	50	-	**Top 25 = 47.5% of N**	

Graph 1-3. Average ratings distributions



Graph 4-5. Average “Is still open”



Upon visual inspection, we identified skewness as Top 25 brands represented nearly half and top 3 represented over a quarter of the sample (N). Also, larger brands are more likely to remain open. As such, we need to caution our analysis is not skewed by large brands.

Average rating also confirms past studies on the upward bias on ratings due to generally above average rating by consumers (Kadry et al., 2011).
The average rating also presents an issue as the lack of variance information means we cannot validate if an average rating represent a true moderate review (i.e. numerous 3 stars)versus average of extreme reviews (i.e. average of 1- and 5-stars). However, Mudambi & Schuff (2011) suggests moderate review is more influential on consumer choices for experience goods - where consumers cannot ascertain quality without trying (like food) - means variance may not be significant to business.

Qualitative Analysis

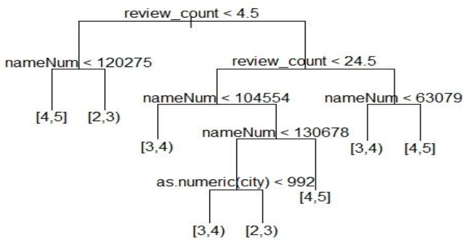
Based on literature review, we expects homogeneity among chains due to brand image and price expectation (De Langhe et. al, 2016) - this may help identify good vs. poor performers within brands. De Langhe et. al further concluded consumers form myopic perception based on average user ratings irrespective of number of reviews - business should take ratings seriously. A few other known issues beyond the scope of this project are malicious comments, controlling for survival bias, correcting or inferring missing data in attributes.

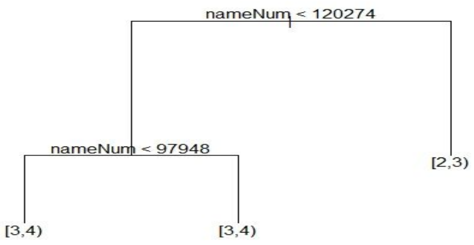
Main Analysis

Our main analysis focuses on results from our unsupervised learning, classification tree, multivariate regression, and logistics regression. Unsupervised learning using Association rules found no significance between any of the 13 variables after discretizing “stars”. We then applied the Decision tree to the 13 variables by converting categorical variables (i.e. city) as numerics. We also selected sub-sample of Top 10 and a specific brand as testing set to adjust for large brands.

For regression analysis, we wanted to test “stars = β0+β1*review_count+β2*name+β3*city” to predict performance difference and “𝐏(is_open=1|X) = review_countβ1+starsβ2” to assess likelihood of a store’s survival. We further separated the sample by open-only, top brands, and a random sample of a brand to test for potential bias.

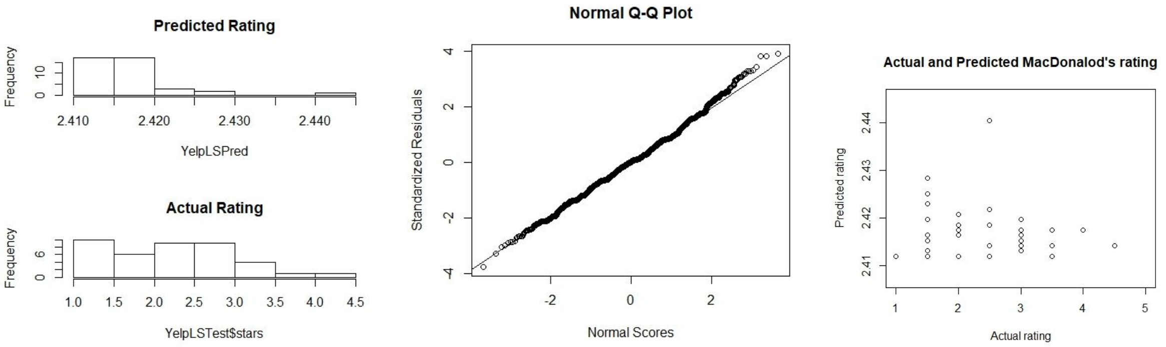
Results and Discussion





Unsupervised learning returned limited insights with 50% error rate for the tree: less than 5 reviews negatively affect ratings and brand effect is prominent.

Graph 4-6 Applying population model to predict McDonald’s rating



For regression analysis, brand and number of reviews for both average rating and survival were generally significant with R-square around 30%. The standardized residual were generally not normal at the tails. Predictive power for both were poor as the intercept (i.e. average rating) dominated the marginal improvement from an incremental review. Also, the explanatory power and significance did not improve when applying the population model to a specific brand. Most of the attributes were not insignificant due to large percentage of missing data.

Implications and future opportunities

Small chain owners needs to keep current on their restaurant attributes on rating sites to encourage patronage (Kincaid et. al., 2010) and allow more systematic analysis on attributes on customer satisfactions and store performances. Our analysis reaffirms restaurant owners’ needs to maintain and engage customers for regular feedback - good and bad. This may mean adding a mobile ordering platforms with robust feedback loops and analytic tools (Hirschberg et. al., 2016).

Future considerations should include a formal behavioral economic model to enhance the explanatory power of the dataset (Harper et. al., 2005) and utilize text analytics tools, like sentiment analysis, to reduce impact from averaging and upward bias of user ratings. More granular location model will also help assess survival rate due to neighbouring competition. In summary, the above recommendations likely is beyond scope of a small restaurant chains due to the computational requirements and advanced R or economics model skill needed.

Conclusion

In this study of online Yelp reviews, we do not believe an average small restaurant chain owners will be able to draw insights from Yelp dataset and R without help of more seasoned data analyst. Small restaurant chain owners can help by ensuring rating and attribute information are available and exploring adapting a mobile ordering platform.

References

Campbell, C. (2015). *Survey Says: 33% of Restaurant-Goers Never Dine Below 4-Stars*. Website, https://www.reviewtrackers.com/survey-says-33-restaurant-goers-dine-4-stars/, accessed 13 April 2018.

De Langhe, B., Fernbach, P., Lichtenstein, D. (2016). Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings. Journal of Consumer Research. 42.

Ding, D., Guan,C., Fang, Z.,Lee, P. (2017). Does Online Rating Affect Companies' Financial Performance? Evidence from Hotel in Singapore. *Journal of Accounting and Finance. 17(9), 60-75.*

Harper, F., Li, X., Chen, Y., Konstan, J. (2005). An Economic Model of User Rating in an Online Recommender System. *Proceeding*
UM'05 Proceedings of the 10th international conference on User Modeling. 307-316

Hirschberg, C., Rajko, A., Schumacher, T., Wrulich, M. (2016). The changing market for food delivery. Website. https://www.mckinsey.com/industries/high-tech/our-insights/the-changing-market-for-food-delivery, access 13 April 2018.

Kadry, B., Chu, L., Kadry, B., Gammas, D., Macario, A. (2011). Analysis of 4999 Online Physician Ratings Indicates That Most Patients Give Physicians a Favorable Rating. Journal of Medical Internet Research, 13(4).

Kincaid, C., Baloglu, S., Mao, Z., Bussers, J. (2010) What really brings them back? The impact of tangible quality on affect and intention for casual dining restaurant patrons. *International Journal of Contemporary Hospitality Management.* 22(2), 209-220.

Mudambi, S., Schuff, D. (2010). What Makes A Helpful Online Review? A Study Of Customer Reviews on Amazon.com. MIS QUarterly, 34(1).