Summary Report

Team 6 - Cohort A
Team Members: Xinman Liu, Missy Putur, Jiao Sun, Adil Wahab, Rui Xu
GitHub Repository: https://github.com/Rui210/BA888.git

Problem Statement

Finding restaurants through directory service and review forum apps has become a crucial part of business discovery. It has changed the way restaurants of different tiers attain brand loyalty and market themselves. For some restaurants, the proliferation of review forums in the food industry has even played a role in reorienting their business goals. The sharing economy has led to a massive rise in posting reviews of products and services with such ease that it has become a norm to rely on consumer reviews to help make purchasing decisions. The growing demand for consumer reviews triggered the integration of directory service apps and review forums. This integration enabled the introduction of apps that allow customers to seamlessly browse for services and reassure them of the quality of different services through proxies based on consumer reviews.

Yelp is one of the pioneers of the integration between directory service and review forum. The San Francisco-based company has played an influential role in the restaurant industry. Yelp continues to be a leader in its sector with the app/web site's ability to easily look up restaurants based on simple yet powerful attributes/filters and of course, users having access to publicly publish their experience and opinions about the businesses. Multiple studies have been conducted to understand the impact of the star rating of a restaurant on its sales. Furthermore, experts want to study the influence of the sentiment of reviews and the changes in star ratings on the chances of a restaurant being booked during peak hours. Leaders in the restaurant industry are digesting the importance of consumer reviews and the impact of directory service companies. Yelp's existence for over a decade has resulted in lakes of historical data. Given the effect of consumer reviews and ratings, and the use of the right tools can help uncover facets of the industry that experts have neglected in the past.

Yelp's terabytes of historical data and the new wave of data-driven decision-making through business intelligence tools and machine learning methods can revolutionize the way restaurants establish business goals and operate on a daily basis. The incorporation of the appropriate data-driven methods and the right combination of attributes can aid restaurants in understanding their impact on customers and what makes customers tick. This project looks to unpack the power of big data analytics in the restaurant industry by digging into the restaurant attributes and reviews. The foundation of this study is based on supervised and unsupervised machine learning techniques to answer the following questions. How can a restaurant leverage information on Yelp to increase their chances of attaining their business goals and preventing permanent closure based on the most significant attributes? What combination of attributes do customers value based on price level and cuisines that make them choose one restaurant over the other?

Dataset Overview

We downloaded our data from Kaggle (https://www.kaggle.com/yelp-dataset/yelp-dataset). The data is divided into four datasets; *business, check-in, reviews,* and *users*. All of them belong to the Yelp public database. We used Yelp's data from 2004 to 2018.

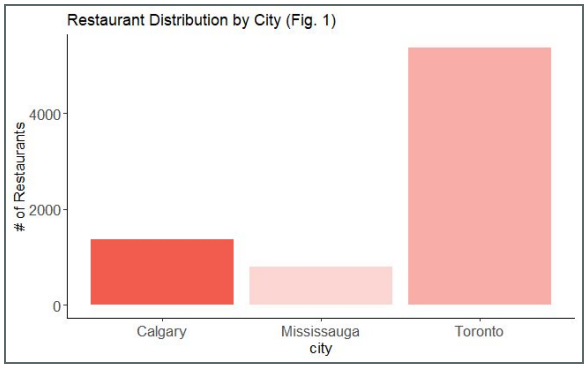
The business dataset contains all catering and other services on Yelp, with the description of *business type*, *characteristics (i.e. ambiance, attire, parking options)*, *location*, *review counts*, *star rating,* and *open hours* as variables. In this project, we are only interested in exploring *review*, *star rating,* and interactions with restaurants. To narrow down our dataset, at first, we manually reviewed the top 1000 phrases that businesses were tagged with to identify phrases that would be a good fit to be identified as a restaurant. We filtered the complete list to only include businesses that were tagged with the phrases "food" and/or "restaurant". Using these phrases we were able to identify 75,000 restaurants. Second, the dataset was filtered to only include major Canadian cities (Toronto, Mississauga, and Calgary) to balance between having a robust dataset that minimizes long compute time while preserving data quality. Montréal was excluded as a majority of the reviews are in French. Lastly, to ensure rich, non-sparse data for the remainder of restaurants in the dataset, we further reduced the restaurant list to include only the restaurants with the top 50% of reviews. With these criteria, only restaurants with 14 or more reviews were included in the study.

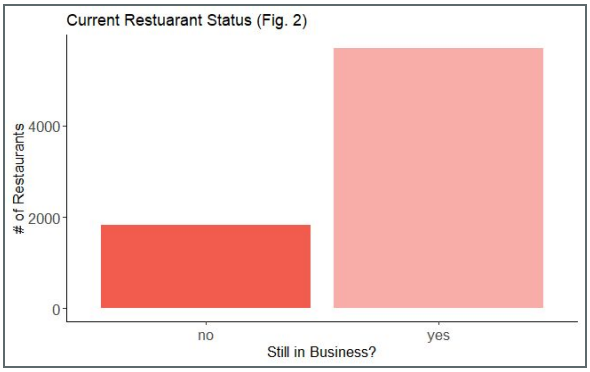
The *review* dataset and *user* dataset contains the information of all users. The review dataset is review-based, includes the review text of each business with *user ids, usefulness count,* and *reactions (useful, cool, funny)*. While the *user* dataset is user-based, recording the basic account information in every user's portfolio, such as the followers, total review counts, average star rating, and total favorite received for one account. Since we focused on the interaction between users' reviews and business management, we combined the two datasets by user id, keeping the users' id, total review counts, and average star rating from user dataset and every piece of review, every star rating and business id from review dataset. After we got this new user-based table, we filter it by the business id in the restaurant dataset (as identified by the above processing steps) to ensure the observation values are consistent in the two datasets. After this reduction, the dataset contained over 480,000 reviews.

The check-in dataset includes the exact timestamp a user checked into a business. It was structured so that each business was represented as a row with a comma-separated list of check-ins by date. To make the data easier to work with we rearranged it so that each check-in was represented as a row rather than a list. The output contains two variables: business id and check-in time of every Yelp check-in. For the supervised machine learning models, the check-ins were aggregated based on the average number of check-ins per month for each restaurant then it was grouped by seasons based on the total number of average check-ins for the months of the respective seasons.

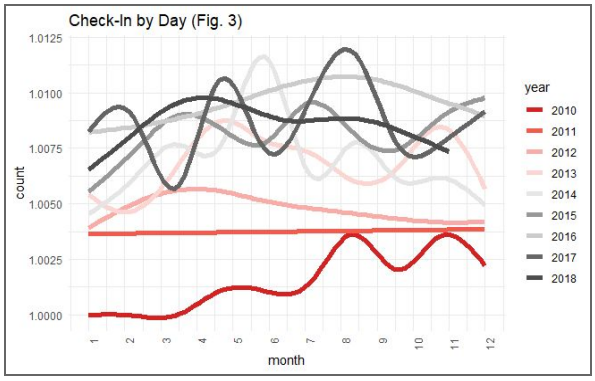
Exploratory Data Analysis

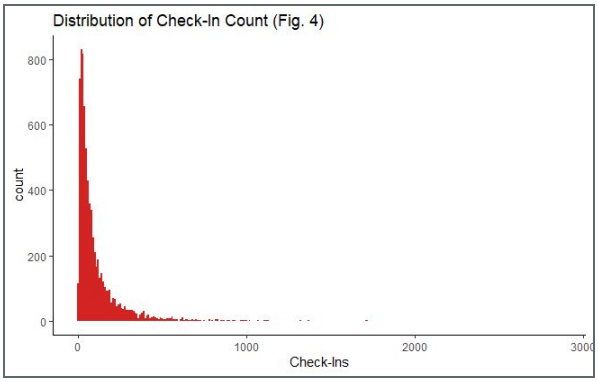
The dataset contained 7,530 restaurants from three Canadian cities. Toronto had the majority of restaurants 5,372 (70%), followed by Calgary with 1,365 (18%), and Mississauga with 794 (11%) (Fig. 1). The dataset contained both open and permanently closed restaurants; 5,699 (76%) restaurants were open and 1,831 (24%) restaurants had gone out of business (Fig. 2).



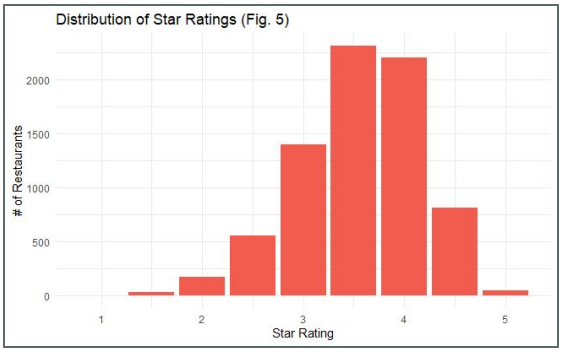


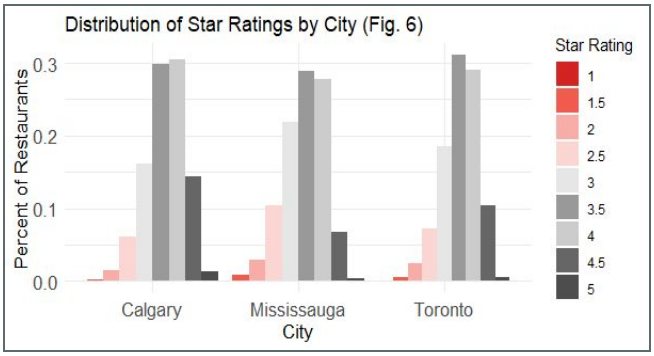
Check-In dates ranged from 2010 to 2018. The average restaurant had 111 check-ins with a median of 57 and a maximum of 2,872 check-ins (Fig. 4). Check-ins steadily increased year over year in the data until about 2015-2017 when they reached a peak (Fig. 3). The first year to see a decrease in check-ins from the previous year was 2018. The check-in data does exhibit some seasonality with visits appearing to typically be higher in the summer months than winter months.





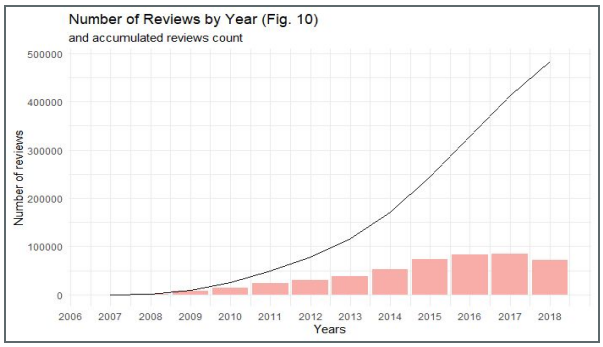
The distribution of star ratings for restaurants was skewed to the left (Fig. 5). The majority of restaurants had a four-star rating with the second most common rating being 3.5 stars. The distribution of star ratings between the three cities was fairly similar, however, Calgary had a slightly higher percentage of 4+ star restaurants than Toronto and Mississauga (Fig. 6).





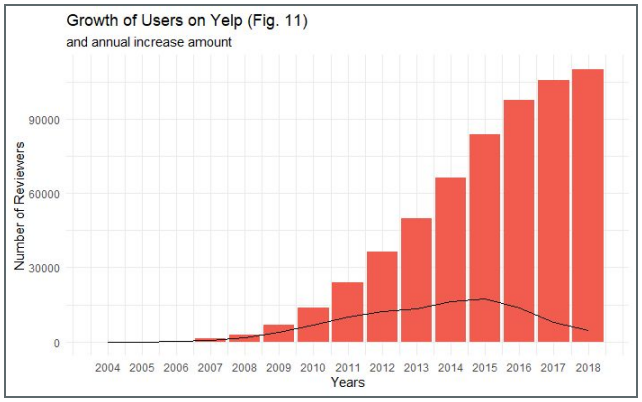
The dataset contained over 480k reviews. The mean number of reviews per restaurant was 64. The distribution of the review length is right-skewed and more than 50% of all the reviews were between 25 to 120 words. The average review length was around 130 words while the median was around 100 words. The most common bigrams found in reviews included “come back”, “ice cream” and “go back” (Fig. 8).

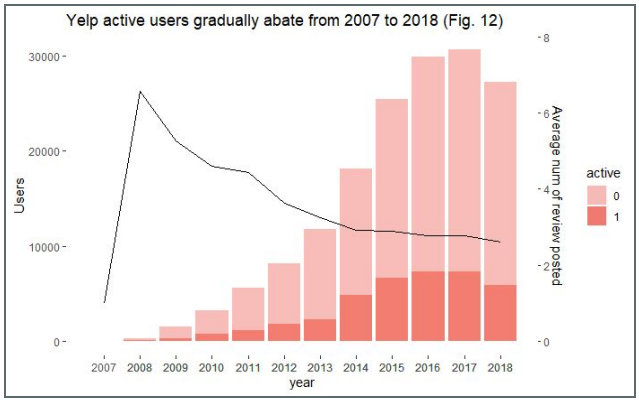
Reviewers submitting 1 and 2-star reviews wrote longer reviews on average than reviewers who gave 3, 4, and 5-star reviews (Fig. 9). The number of reviews increased year over year to a peak in 2017 of just under 100,000 reviews (Fig. 10).



By 2018, there were over 900,000 users who had written at least one review, or checked-in once at one of the restaurants in the dataset (Fig. 11). The number of new users increased year over year until 2015 when annual new users peaked (Fig. 11).

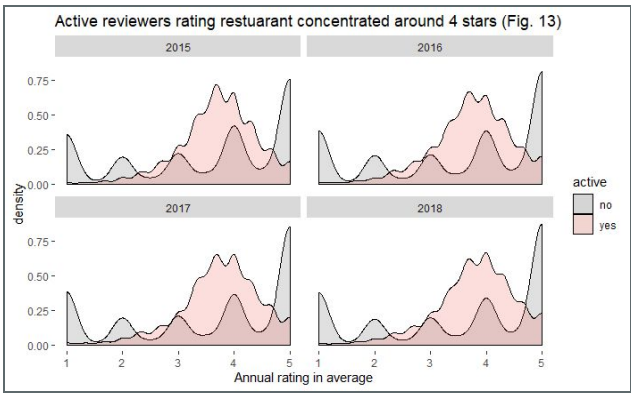
For the purposes of this project, an active user is defined as someone who posts more than the average reviewer in a year. The number of users who are considered active was increasing year over year until 2018 (Fig. 12). The percent of users who were considered active in 2017 was 24% compared to 22% in 2018. Although the total number of users grew year over year, the number of reviews written on average steadily decreased from a peak of around 7 reviews per user in 2007 to just over 2 reviews in 2018 (Fig. 12).

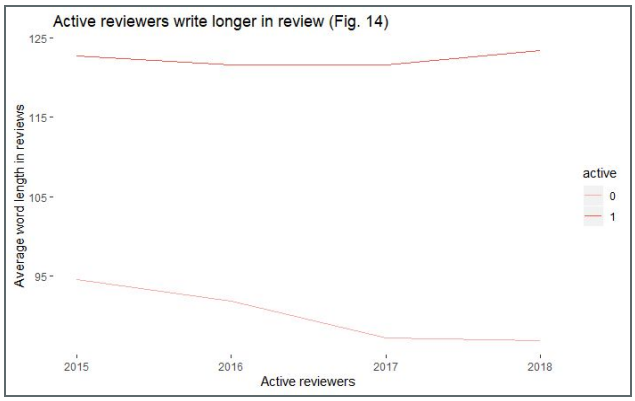




Not only did active users write more reviews on average, but the length of their reviews were also longer. The average active reviewer wrote reviews containing just under 125 words compared to the average inactive user whose reviews typically contained under 95 words (Fig. 14).

Active and inactive users gave different star ratings on average. Active users rate in a more normal way with a peak at about four stars, while inactive users are more extreme in their rating. Inactive users give 1-star and 5-star reviews more often than active users (Fig. 13).





Methods

Our project has two main objectives:

A. Descriptive and Diagnostic Analytics

Our first goal is to study the most essential aspects of a restaurant based on price level to understand what customers value most when rating a restaurant (e.g. Do customers at low-price restaurants value the same "things" as mid-price and high-price restaurants).

To better understand which aspects of the restaurant and dining experience customers value most we will use topic modeling to see if themes differ depending on if the restaurant is categorized as inexpensive (labeled as "$") or expensive (labeled as "$$$$"). An initial hypothesis is that customers in $ and $$ restaurants might care more about the speed of service and getting their money's worth while patrons of $$$ and $$$$ might value attributes like ambiance and customer service more.

B. Predictive Analytics

The next topic we are interested in exploring is whether or not we can accurately predict that a restaurant is out of business. Through this portion of the project, the goal is to help restaurants by creating a way to identify if some sort of changes in their metrics (star ratings, ambiance, amenities, check-ins) indicate that there is a downturn in customer satisfaction with their restaurant.

To achieve this goal we will use various supervised machine learning methods to predict and identify what attributes of a restaurant and check-in months have the strongest influence on the outcome variable of whether a restaurant is open or permanently closed.

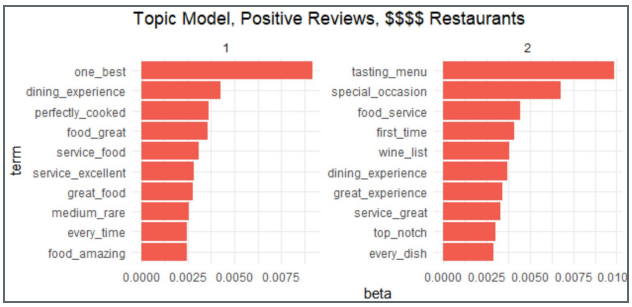
Our final goal is to create a recommendation system that predicts users' preferences based on other reviewers that are similar to them in the way they rate restaurants they visit.

Results

1. **Topic Modeling**

To explore if customers at inexpensive restaurants value different aspects of their dining experience than customers at expensive restaurants we performed topic modeling. We attempted grouping

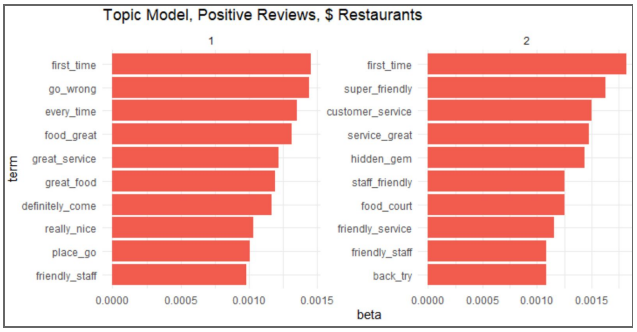
restaurants and reviews in several different ways including comparing 1-star and 2-star reviews to 4-star and 5-star reviews and comparing “$” and “$$” restaurants to “$$$” and “$$$$” restaurants. After exploring several subsets we found we were able to find the most unique topics between groups when we limited the dataset to only include “$” (inexpensive) and “$$$$”(expensive) restaurants and 1-star and 5-star reviews written about those restaurants.

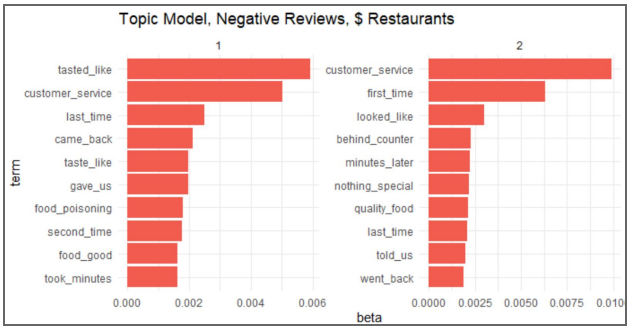


$$$$ Restaurants Reviews written by customers that had a positive experience at expensive restaurants typically fall under two topics. The first topic is around “food preparation”. Customers that write about this topic commonly use phrases like “perfectly cooked”, “great food” and even use specific phrases about how the food was prepared like “medium rare”. The second topic focuses on “special occasions”. These customers use phrases like “tasting menu”, “special occasion”, and “first time” to describe their experience. These customers do not appear to regularly dine at expensive restaurants but are looking to be wowed with a special experience when they do.

When exploring 1 and 2-star reviews for expensive restaurants, unique topics were difficult to find. Top phrases in negative reviews included neutral phrases like “customer service”, “tasting menu”, and “first time”, however within the top fifteen phrases things like “minutes later”, “quality food”, and “food poisoning”.

$ Restaurants We found two core topics for positive reviews written about inexpensive restaurants. While “first time” is again a popular phrase, more patrons for these restaurants indicated that they regularly visit the establishment. The first topic can be labeled as “Consistent Food”. These customers use phrases like “can’t go wrong”, “every time” and “great food”. They are regulars at these restaurants and enjoy that they consistently have a good meal, no matter what they choose on the menu. The second topic can be labeled as “Friendly Service”. While these customers do enjoy and mention the food, they use phrases like “super friendly”, “staff friendly” and “friendly service” when giving five-star reviews indicating that a welcoming and accommodating staff sets the place apart.





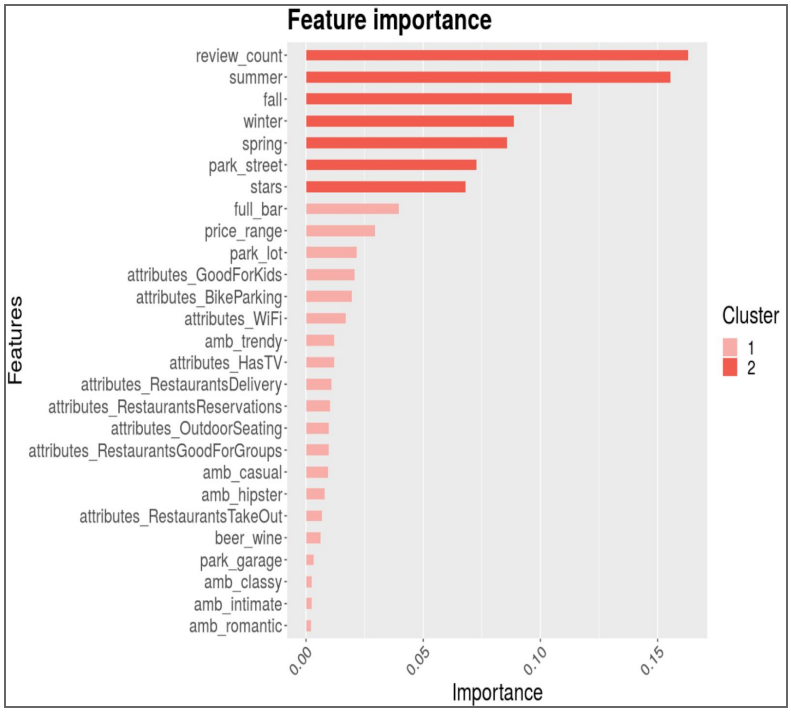
For negative reviews, customers discussed two topics, “Food” and “Service”. Comments around food that frequently came up when discussing a negative experience with food were comments regarding bad taste as well as the fact that the food may have given them food poisoning. Customers who discussed service mentioned things like “minutes later” and “behind the counter”. These customers

were dissatisfied with the length of time it took their food to be ready. Also, a problem that is more unique to inexpensive vs. expensive restaurants is that customers are able to watch employees behind the counter interact and/or prepare their food. Customers with a negative experience often discussed behavior they saw behind the counter that detracted from their experience. Finally, these customers more frequently used phrases like “came back” and “last time” indicating that it was not their first time dining at the restaurant.

2. Supervised ML Models

Dissecting the attributes that play the most influential role in predicting whether a restaurant is in business or not through predictive modeling is a multi-faceted study. Utilizing supervised machine learning techniques of distinctive features enables us to uncover not only the features of a restaurant holding the most weight in predicting permanent closure but also the direction of the impact. Certain models are well-designed for determining the variation in levels of importance for prediction, whereas other models attempt to balance between variable importance and settling on the exact weights (positive and negative). *Random Forest* and *Extreme Gradient Boosting (XGBoost) models* were deployed for the purpose of variable importance. *Logistic regression models (Simple, Lasso, Ridge)*, and *Stepwise regression models* examined the direction of the relationship between attributes and chances of staying open. Studying all the seven models revealed that **a mix of three models (XGBoost, Lasso, and Forward Stepwise)** accomplished the goal of identifying critical attributes and how changes to those attributes may correlate with the chances of closure.

2a. XGboost



For XGBoost, we constantly adjusted the number of classes and rounds to balance the variables’ importance and weights. When we predicted whether restaurants opened or not based on the features and seasons, we got 27 variables that influenced the results. According to the importance of variables, it can be roughly divided into two clusters. For the first cluster, summer was the peak season, and spring was the slack season at most restaurants. Meanwhile, review count and star rating were the most

important features. Because stars and reviews were all the feedback that customers gave to the restaurants, they can reflect the service quality of a restaurant, then the restaurant can make corresponding improvements based on the reviews to reduce the risk of closing. What's more, people relied on star ratings to select restaurants. For the second cluster, whether the restaurant had a bar was very important, because alcohol sales have higher gross margins than food, you can improve your bottom line with a full bar, and alcohol sales account for around 30 percent of the revenue in most restaurants. For people, they would also choose restaurants according to the price range. Furthermore, was there a parking lot near the restaurant that was still important, because people did not want to spend lots of time finding parking lots, so the park lot feature and park street feature were showing high correlation on the chance of restaurants staying in business. Finally, the predicted accuracy rate of XGBoost is 77.2%.

2b. Lasso Regression

A Lasso regression model's unique ability to perform both regularization and variable selection to enhance the prediction accuracy and interpretability seamlessly aligns with the goal of the study. The model's prediction accuracy is around 77%. There are fifteen variables that correlate with the chances of whether a restaurant will stay open or not, but only a few variables appear to hold the most weight. Customers having access to a parking lot associated with the restaurant has the highest positive correlation on the chances of a restaurant staying in business. However, a restaurant that only has street parking appears to have the strongest negative correlation with staying in business. Note that this study looks at some of the most populated cities in Canada: Toronto-Mississauga, and Calgary. A customer's inability or hassle to find parking seem to be detrimental to restaurant sales and possibly reputation. Customers may not want to deal with the headache of metered-parking street parking while trying to enjoy a meal. They may also not want to wait for a spot to open up close to the restaurant or park far from the restaurant requiring them to walk more than they may be willing to.

Variable	Coefficient
park_lot	0.257
attributes_GoodForKids	0.1689
stars	0.1494
attributes_RestaurantsTakeOut	0.0653
attributes_WiFi	0.0175
review_count	0.0086
summer	0.0029
attributes_OutdoorSeating	-0.0081
amb_classy	-0.0167
attributes_RestaurantsReservations	-0.0376
amb_trendy	-0.0402
attributes_RestaurantsAttire	-0.091
attributes_BikeParking	-0.1829
full_bar	-0.3119
park_street	-0.4779

A higher star rating and being accommodating towards kids have a relatively similar strong positive correlation with a restaurant staying in business. The importance of star rating is undeniable as more and more customers start to rely on ratings from directory service and review forum apps to decide on a restaurant. This may prove to restaurant managers and PR teams the value of investing in strategies to improve their presence on apps like Yelp. Restaurants being accommodating towards kids may be an indicator that many parents who have kids may be limited or more inclined to dine at restaurants where they do not have to worry about where to leave the kid/s if they want to go out to eat.

2c. Forward Stepwise Regression

As expected, many of the variables describing the characteristics of the restaurants were included in the forward regression model. The final forward stepwise model contained 22 variables, but only a few variables are significant. The model's prediction accuracy is around 73.35%. From the results on the table to the right, it is clear that the

Variable	Coefficient
park_lot	0.0521
stars	0.0455
attributes_GoodForKids	0.0379
park_valet	0.0333
amb_touristy	0.0282
attributes_RestaurantsDelivery	0.0278
attributes_WiFi	0.0179
attributes_HasTV	0.0171
attributes_RestaurantsGoodForGroups	0.0139
amb_upscale	0.0062
review_count	0.0019
fall	0.001

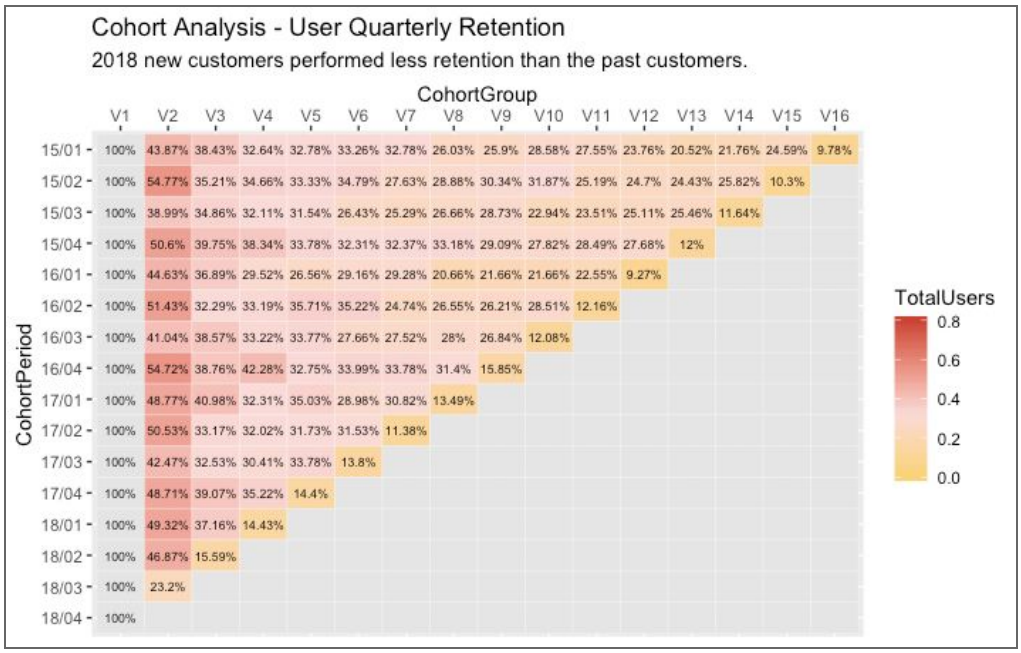
presence of a parking lot, star ratings, kid-friendly setting, and valet availability have higher positive correlations to whether the restaurant is open or not. While just street parking, a full bar, and a romantic ambiance have strong negative correlations to whether a restaurant stayed in business. I would say that customers pay more attention to parking in restaurants. Because we are studying four metropolitan areas, where it was difficult to find parking spaces. Therefore, customers can find parking spaces seems to be a favorable trend for restaurant sales. This is why restaurants with parking lots have a positive correlation with the probability of restaurants continuing to operate, while restaurants with only street parking lots seem to have the strongest negative correlation with continuing to operate. Also, star rating also plays a vital role in the restaurant business. Yelp has now become a reference app for many people to find restaurants. When restaurants have a high star rating, it will be regarded as a good choice by people by default. Therefore, an endless stream of people will go to restaurants to taste. Besides, customers usually choose a child-friendly restaurant, so they don't need to pay more attention to looking after their children, and they can enjoy a meal well.

Variable cont.	Coefficient
winter	-0.0011
price_range	-0.0098
attributes_OutdoorSeating	-0.0175
park_validated	-0.0482
beer_wine	-0.0512
attributes_BikeParking	-0.0555
amb_trendy	-0.0576
amb_romantic	-0.0612
full_bar	-0.087
park_street	-0.105

3. Cohort Analysis

Customer reviews are an important reference for restaurants to design menus, develop dishes, and improve services. Cohort Analysis on Yelp users helps us to inspect whether Yelp is a vibrant local community with active reviewers contributing to high-quality content. It determined whether our review research could significantly represent the opinion held by the local customer population, which is the true reference for local business development.

3a. User retention

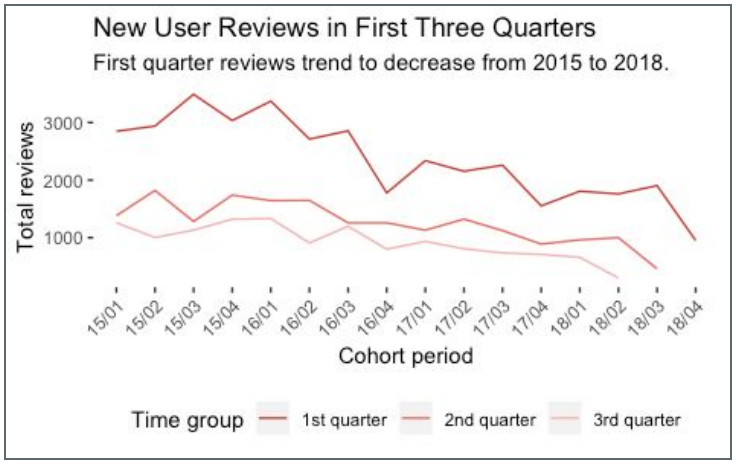


Dividing years 2015 to 2018 into 16 quarters, we calculated quarterly retention rates for users coming in each period. In the local area, the heat map shows over 40% of new users would return to review

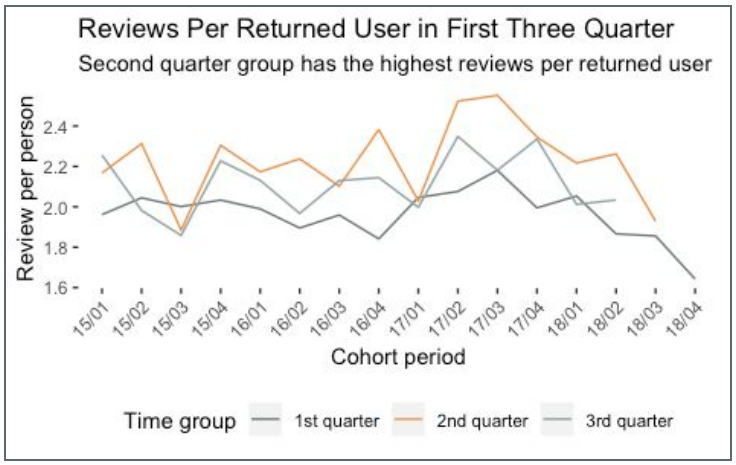
restaurants on Yelp in the second quarter, 30% ~ 40% of new users kept actively reviewing within 4 to 8 quarters, while the review group would decrease to 20% or less after 8 quarters. Converting our findings to the general time unit, it reported that yelp has around 40% user retention in the first year, around 30% user retention in the second year, and less than 20% user retention after the second year. As the increase in the number of reviews posted annually started to slow down, users who joined within two years can be Yelp's target group for review engagement.

3b. Review Posting

We researched reviews posted by returning users on a quarterly basis. The first three cohort groups are emphasized since these groups have relatively large user retention [30% ~ 40% new users]. The large population makes the observed trends representative and reliable.



The line plot at the top right shows the total review amount in the first three cohort groups by 16 cohort periods. Reviews in all three groups tend to decrease period by period, this trend is the most obvious for the first quarter group, which means the posted contents by new users in the first quarter decreased from 2015 to 2018. It is a warning of recent content contribution in the Yelp community, users still kept a retention rate as same as previous while posting fewer reviews totally.



The line plot at the right bottom shows the average number of reviews written by individual contributors in the first three cohort groups by 16 cohort periods. The line that presents the second-quarter group trends above other group lines among all time periods. It indicated that Yelp users tended to post more frequently in their second quarter after joining than they did in their first quarter or at any later time. Combined with the user retention result, the second quarter in the Yelp user lifetime circle is a core period to engage users to contribute to reviews. It is a time that has both a high retention rate and most reviews posted individually.

4. Recommendation System

Providing recommendations for users is an engagement tactic which not only personalizes users' experiences on the site but also encourages customers to rate more restaurants to receive new and more accurate recommendations. A User-Based Collaborative Filtering model was implemented as our recommender algorithm to predict both new restaurants' star ratings and a recommended restaurant list for sample individuals. It works by clustering users based on similar rating profiles. If two users have the same star ratings for two or more restaurants, the algorithm will recommend new restaurants that are liked by one user to the other.

We took 13,807 reviewers' rating profile to 680 restaurants in Mississauga as sample data. To reduce noise, we only included Yelpers who had rated at least 10 restaurants. We used 75% of observation for the training group and 25% for the test group. The final sample data was a 490 * 670 matrix.

Recommendation results give every observation a 10-restaurant list in which the algorithm detected as the items the user would like. We estimated the prediction accuracy for 117 users in the test group. 96.70% of prediction is corrected, while only 7.9% of prediction results are relevant to the actual user rating profile. Having a few true positive predictions in our data caused the low relevant prediction.

Discussion

While topic modeling is a good way to get an initial idea of the aspects of a restaurant reviewers are most frequently commenting on, it is hard to use topic modeling to make concrete claims. For example, we stated that Yelpers reviewing $-restaurants commonly discussed the topics of friendly staff and having a consistently good meal. We think because these comments are frequently mentioned that they are very important to customers and how they rate restaurants. However, to prove these characteristics make a significant impact on how a customer applies a star rating to a restaurant, these ideas would need to be tested through experiments.

For the supervised machine learning algorithms, we were mainly working with categorical variables. Those variables were made up of 1's and 0's to indicate whether or not a restaurant had a certain feature or the variables contained several levels of categories like noise which could be rated as "none', "quiet", "average", "loud", and "very loud". Additionally, there were many numerical features we would have liked to include in the models that we did not have access to like the number of items on the menu, the average price of menu items, the number of seats available in the restaurant, etc. Our objective of implementing supervised machine learning models was to predict if a restaurant is in business or not based on the amalgamation of restaurant attributes available, which served as the basis for studying the degrees of correlation among these attributes and a restaurant's success in staying in business. Using a dataset that contained an equal amount of closed and open restaurants with greater niche attributes would make a more robust dataset for achieving a highly accurate model.

Creating an accurate recommendation system was very challenging with the current subset of data we were using. At first, we were pleased because the accuracy score of the recommender system stated that it had accurately predicted whether a customer liked or disliked a restaurant 96% of the time. However, after examining the precision and recall scores and finding that they were both <8% we realized the accuracy score was misleading and did not paint the full picture. For the customer population we were examining, only 3.6% of the more than 13,000 Yelpers had rated at least 10 or more of the 600+ restaurants. Because of this, if the recommender system was to guess a rating of 0 for every restaurant, it would automatically have an extremely high accuracy rate. To create a more useful recommender system Yelp would need to either only recommend for customers with extremely high rating frequency, which doesn't currently exist, or only include restaurants that are already extremely frequented which is not as useful for helping Yelpers discover new places.

Conclusion

User retention on Yelp tends to drop from 40% to 20% in the first two-year lifetime. The second quarter is the core time period to encourage users to post reviews. New user activity went down in 2018.

Yelper's do not visit $$$$-restaurants regularly, and when they do so they are often visiting the first time, to celebrate a special occasion and to experience something unique like a tasting menu. High-end restaurants should ensure they have a signature item that customers can order to help them feel celebrated. Customers that visit $ value restaurants where they can consistently have a good meal and are greeted by friendly staff. Inexpensive restaurants should focus on training and hiring employees that are friendly as their interactions with customers are shorter and they have less time to make a good impression; they should also be aware that how employees are working behind the counter, when not interacting with customers, can have an impact on the customer's perception of service.

High review count and star ratings have the strongest significance in predicting that a restaurant will stay in business. The presence of a parking lot has a high positive correlation with staying in business, while just having street parking is highly negatively correlated. Kid-friendly restaurants and a high volume of check-ins during the summer also appear to be positively correlated with staying in business.

Restaurant industries all around the world have been severely rocked by the current COVID-19 pandemic. Many restaurants have begun to or will likely need to restructure their business models for once the situation improves and regular business hours resume. We hope this study can serve as a reference for restaurants that plan to reorient themselves in the market or new restaurants looking to enter the market.