Decoding Restaurant Survival: The Impact of Online Reviews and Pricing

Mustafa Elahi

October 5, 2024

1 Introduction

The proliferation of online review platforms–such as Yelp, Google Reviews, and TripAdvisor–has fundamentally altered the landscape of the dining industry, placing a significant emphasis on online consumer feedback and reviews. Studies have shown that a one-star increase in Yelp ratings can lead to a 5-9% increase in revenue for independent restaurants, underscoring the impact of consumer ratings on business success (Luca 2016). However, the literature has yet to fully explore how the interplay between customer ratings, the volume of reviews, and other factors such as price level ($, $$, $$$, $$$$), contributes to a restaurant's longevity in the market. This gap highlights a critical area of inquiry, given the high failure rates within the hospitality industry and the increasing competition fueled by the digital economy.

This research project seeks to address this gap by examining how customer ratings and the volume of reviews, alongside price level, influence a restaurant's ability to remain operational in the competitive dining market. This study specifically focuses on star ratings, review count, and price level as primary variables due to their direct visibility to consumers and reported influence in existing literature. By employing logistic regression, Lasso regression, and Ridge regression, this study aims to dissect the relationship between these factors and restaurant survival. The choice of these methods is strategic; logistic regression will allow for modeling the probability of business survival as a binary outcome, while Lasso and Ridge regressions will allow for the identification of the most predictive factors by penalizing the less significant ones. This approach is expected to provide a nuanced understanding of the determinants of restaurant success.

This study delves into the complex impact of digital feedback on restaurant survival, subtly suggesting that the intuitive link between high ratings, extensive reviews, and business longevity may not always align with reality. Such findings invite a reevaluation of conventional beliefs, unveiling the intricate influences at play in determining a restaurant's success in the digital age.

2 Data Sources and Descriptions

This study uses the Yelp Open Dataset, a publicly accessible dataset designed for academic research. It includes millions of reviews, business profiles, and user interactions from various global cities, providing detailed insights into consumer behavior and business performance, especially in the restaurant industry. Our research focuses on the $'yelp_{a}cademic_{d}ataset_{b}usiness.json'$ file, which offers comprehensive information on businesses such as ratings, review counts, price levels, and operational status. During preprocessing, we filtered the dataset to include only restaurants with complete data.

To analyze the impact of consumer reviews and ratings on restaurant longevity, the following variables were extracted and utilized:

Stars: The average rating a restaurant receives on Yelp, ranging from 1.0 to 5.0.

Review Count: The total number of reviews a restaurant has received.

Price Level: The average cost of dining at the restaurant, denoted on a scale of 1 to 4, with 1 being low cost and 4 being high cost.

Restaurant Status: Whether the restaurant is open (1) or not(0).

The summary statistics of these variables are presented in Table 1, offering a snapshot of the dataset's characteristics.

Access to this dataset was obtained directly through the Yelp Dataset website (https://www.yelp.com/dataset), where Yelp periodically releases updated versions of the dataset for research purposes.

3 Method

This research employs a combination of logistic regression, Lasso regression, and Ridge regression models to analyze the impact of online consumer feedback on restaurant longevity. The primary challenge in this prediction problem lies in the multifaceted nature of the data and the inherent complexity of the restaurant industry's dynamics.

Logistic regression (1) is used to model the odds of a restaurant's operational status, effectively connecting consumer feedback (stars, review count) and pricing strategies to the pivotal question of whether a restaurant stays in business.

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \tag{1}$$

In this regression model, $y_i \in \{0,1\}$ represents the operational status of the restaurant (1 for open, 0 for closed), X_{i1} is the average star rating, X_{i2} is the review count, X_{i3} is the price level, and ε_i is the error term. The coefficients β_1, β_2, and β_3 measure the impact of each independent variable on the likelihood of a restaurant being operational.

The assumption of conditional independence, expressed as $E[\varepsilon_i|X_i] = 0$, is critical for the validity of our analysis. This assumption implies that the error term is uncorrelated with the independent variables, ensuring that our estimates of β are unbiased and consistent.

To pinpoint the key determinants for restaurant longevity, we utilize Lasso (2) and Ridge (3) regression techniques. These models are particularly useful for penalizing the magnitude of coefficients, thus reducing the risk of overfitting and enhancing the model's predictive accuracy. The Lasso model is known for its ability to shrink some coefficients to zero, effectively performing variable selection. The Ridge model, on the other hand, shrinks coefficients towards zero but does not set them to zero, which is beneficial in cases where multicollinearity is present among the predictors.

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3})^2 + \lambda \sum_{j=1}^{3} |\beta_j| \right\} \tag{2}$$

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3})^2 + \lambda \sum_{j=1}^{3} \beta_j^2 \right\} \tag{3}$$

The main parameter of interest in our models is the set of coefficients β_1, β_2, and β_3, which correspond to the average star rating, review count, and price level, respectively. The estimated coefficients, denoted as $\hat{\beta}$, reveal the direction and magnitude of the relationship between each independent variable and the probability of a restaurant being operational. A positive $\hat{\beta}$ indicates that an increase in the independent variable is associated with a higher likelihood of the restaurant remaining open, while a negative $\hat{\beta}$ suggests the opposite.

We employed logistic regression, Lasso, and Ridge regression due to their relevance to our study's objectives. Logistic regression models the binary outcome of restaurant operation, making it suitable for our analysis. Lasso regression aids in variable selection, focusing on the most significant predictors for restaurant survival, while Ridge regression checks the stability of these variables against multicollinearity. This methodological combination helps answer our research question and enhances the accuracy and reliability of our findings.

4 Results

The results from running the logistic regression revealed that higher customer ratings (stars) were paradoxically associated with a slightly decreased likelihood of a restaurant remaining operational ($\beta_1 = -0.1324$). This suggests that while positive ratings are desirable, they may not be the sole determinant of a restaurant's success. In contrast, a higher volume of reviews positively correlated with the likelihood of a restaurant staying in business ($\beta_2 = 0.0049$), indicating that customer engagement through reviews might be more critical to restaurant survival than previously thought. Moreover, restaurants with higher price levels were found to have a lower probability of staying open ($\beta_3 = -0.4066$), suggesting that affordability might be a key factor in attracting and maintaining a stable customer base.

Lasso regression provided a more refined view by applying a penalty to the regression coefficients, effectively shrinking less significant predictors to zero. This method identified the volume of reviews as a particularly strong predictor of restaurant survival, reinforcing the notion that active customer engagement is vital. The negative impact of higher price ranges was also confirmed, albeit the effect was smaller than what logistic regression suggested.

One thing to note from the results of the Lasso regression is that none of the coefficients were shrunk to 0, hence all the variables provided must have been considered relevant by the Lasso model in predicting the operational status of restaurants.

In Figure 1, as a result of running a Ridge regression, we see that the mean squared error (MSE) increases with alpha for the larger alpha values. This indicates that there could be under-fitting as a result of over-penalization.

On the left, the Coefficients vs. Alpha graph gives us insight into the Ridge regression coefficient values for each of the independent variables. These coefficient values confirm the general relationship trends identified by the logistic and Lasso regressions; stars and price level are negative predictors whereas review count is a positive predictor. The stability of these predictors shows that they must have a robust effect on the outcome variable (operational status) that isn't influenced by multicollinearity.

5 Conclusion

Our research successfully addressed the initial question regarding the impact of online consumer ratings, the volume of reviews, and price levels on restaurant longevity. We discovered that, contrary to expectations, higher star-ratings do not guarantee business survival, but a higher volume of reviews positively impacts business longevity. Additionally, our analysis confirmed that affordability is crucial for attracting a stable customer base, as higher price levels negatively affect a restaurant's chance of staying open.

The strength of this study lies in its analytical approach, utilizing logistic regression, Lasso regression, and Ridge regression to dissect the multifaceted relationships between consumer feedback and restaurant success. This methodology allowed us to isolate and identify the most significant factors contributing to restaurant longevity.

Although we were able to answer the initial research question, this study is not without limitations. Complete reliance on the Yelp dataset introduces potential bias, as it may not fully represent the diversity of the global restaurant industry. Additionally, the study's scope was limited to factors readily available in the dataset, excluding potentially influential variables such as menu diversity, social media presence, and level of competition. The low predictive accuracy of our model suggests that a broader range of variables may need to be considered to fully address the research question.

Future research could build upon our findings by incorporating a broader set of variables and exploring the role of external economic and social factors. Expanding the dataset to include other review platforms and geographical locations would also enhance the generalizability and depth of the analysis.

In conclusion, while our study advances the understanding of the digital feedback's role in restaurant longevity, it is certainly not enough to make broad generalizations about what really matters when it comes to restaurant longevity. We hope that this research opens the door for further exploration into the complex web of factors that define success in the hospitality industry.

References

Luca, Michael. 2016. ”Reviews, Reputation, and Revenue: The Case of Yelp.com.” Harvard Business School Working Paper, No. 12-016.

6 Tables and Graphs

6.1 Summary Statistics

See Table 1 for the summary statistics of the dataset used in this study.

Table 1: Summary Statistics

Variable	Count	Mean	Std Dev	Min	Median	Max
Stars	56430.0	3.498627	0.827543	1	3.5	5
Review Count	56430.0	85.652100	184.142038	5	33	7568
Price Level	56430.0	1.617969	0.589207	1	2	4
Restaurant Status (is open?)	56430.0	0.667641	0.471063	0	1	1

Note: This table summarizes the key statistics for Ratings, Review Counts, Price Level, and the dependent variable, Restaurant Status, of restaurants in the Yelp Open Dataset. Restaurant Status indicates whether the restaurant is open (1) or not (0).

6.2 Logistic Regression Results

Refer to Table 2 for the results of the logistic regression analysis.

Table 2: Logistic Regression Results

Variable	**Coef.**	**Std.Err.**	**z**	**P> $\|z\|$**	**[0.025**	**0.975]**
Intercept	1.4880	0.045	32.780	0.000	1.399	1.577
Stars	-0.1324	0.011	-11.747	0.000	-0.154	-0.110
Review Count	0.0049	0.000	39.048	0.000	0.005	0.005
RestaurantsPriceRange2	-0.4066	0.016	-25.720	0.000	-0.438	-0.376

Note: Each of the coefficients for the independent variables are statistically significant.

6.3 Lasso Regression Results

Refer to Table 3 for the results from the Lasso regression analysis.

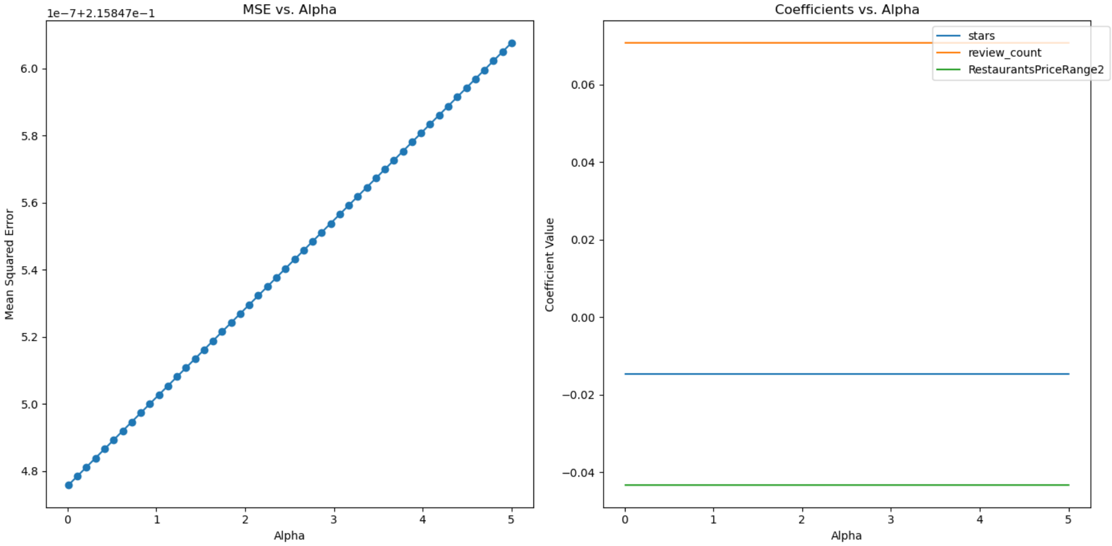
Table 3: Lasso Regression Results

Variable	**Coeff LASSO**
Stars	-0.015435
Review Count	0.072171
Price Level (RestaurantsPriceRange2)	-0.043818

6.4 Ridge Regression Results

For insights into the Ridge regression analysis, see Figure 1.

Figure 1: Ridge Regression Results


MSE vs. Alpha
1e−7+2.15847e−1
Mean Squared Error
Alpha
Coefficients vs. Alpha
stars
review_count
RestaurantsPriceRange2
Coefficient Value
Alpha