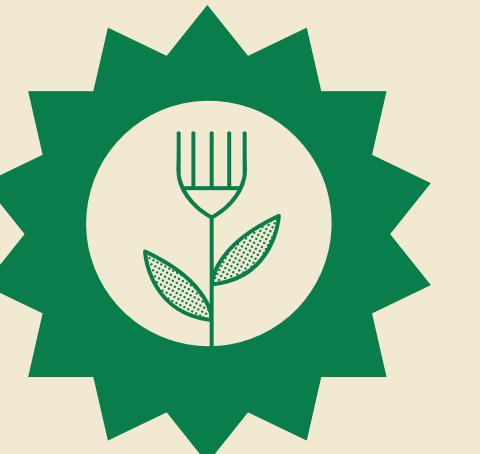


Data Analytics Project by Yennie Lee

Data-Driven Insights into Restaurant Success





CONTENT

Overview

3

About the dataset

5

Preprocessing

7

Define custom performance metrics

9

Statistical Testing

13

Machine Learning

16

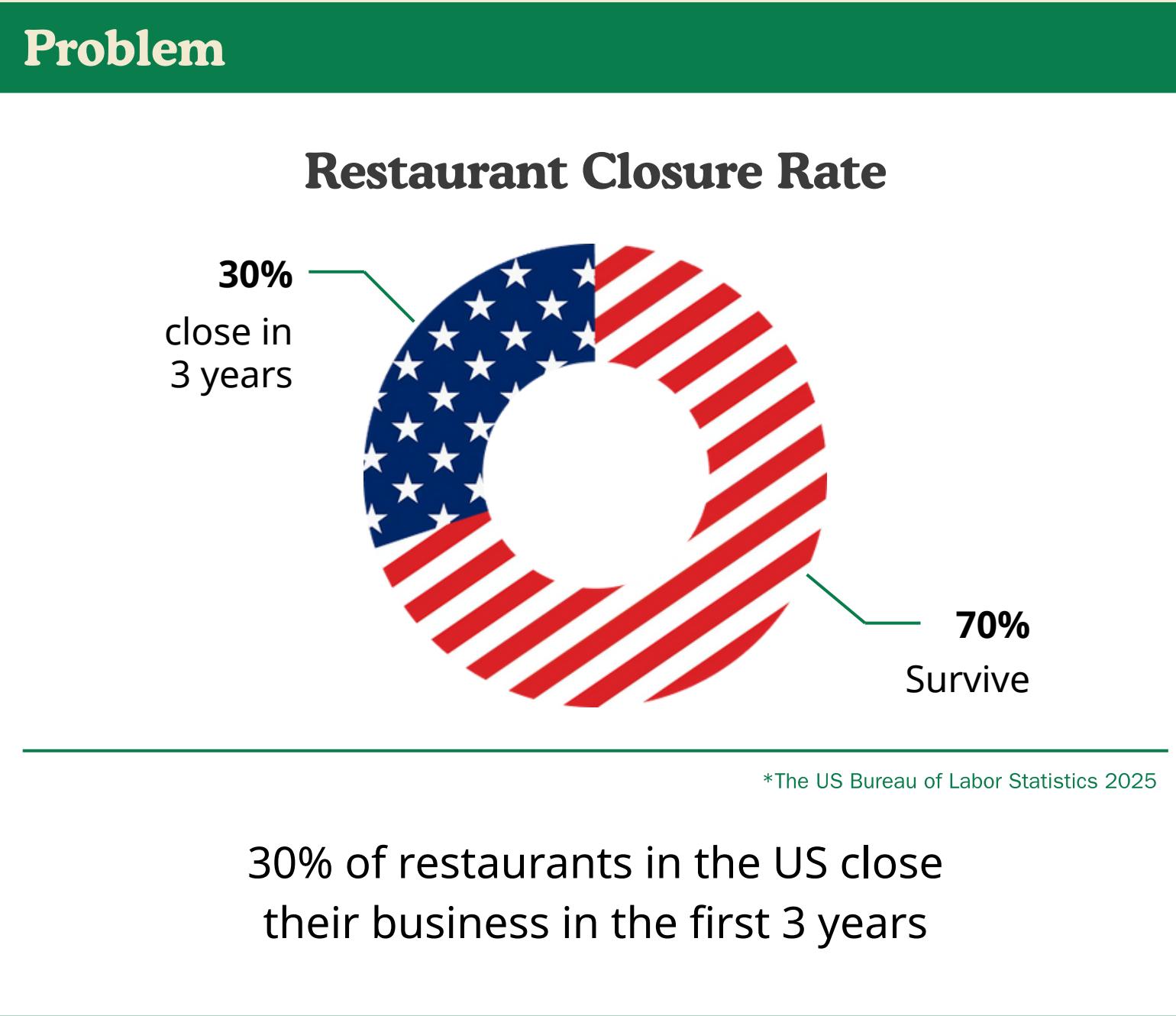
Insights Deep Dive

18

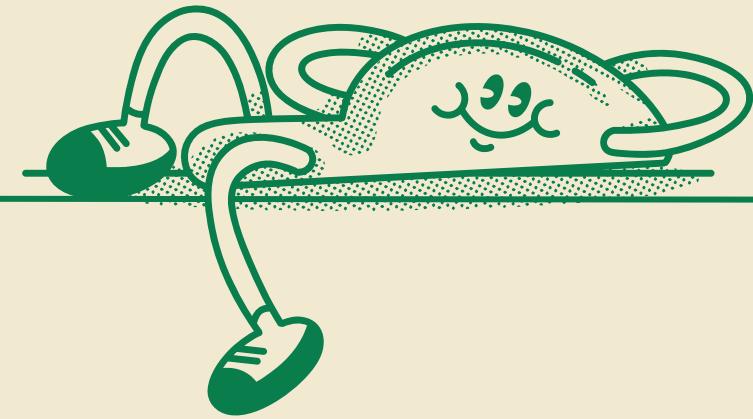
Recommendations for business

24

Project Background



Major Steps



- 1 Define custom performance metrics
- 2 Validate custom metrics using statistical testing
- 3 Identify key features that positively or negatively impact each metric
- 4 Translate analytical insights into actionable recommendations



Target Audience

Small- to mid-sized independent restaurant owners who make decisions on menu design, operations, and positioning

yelp* open dataset



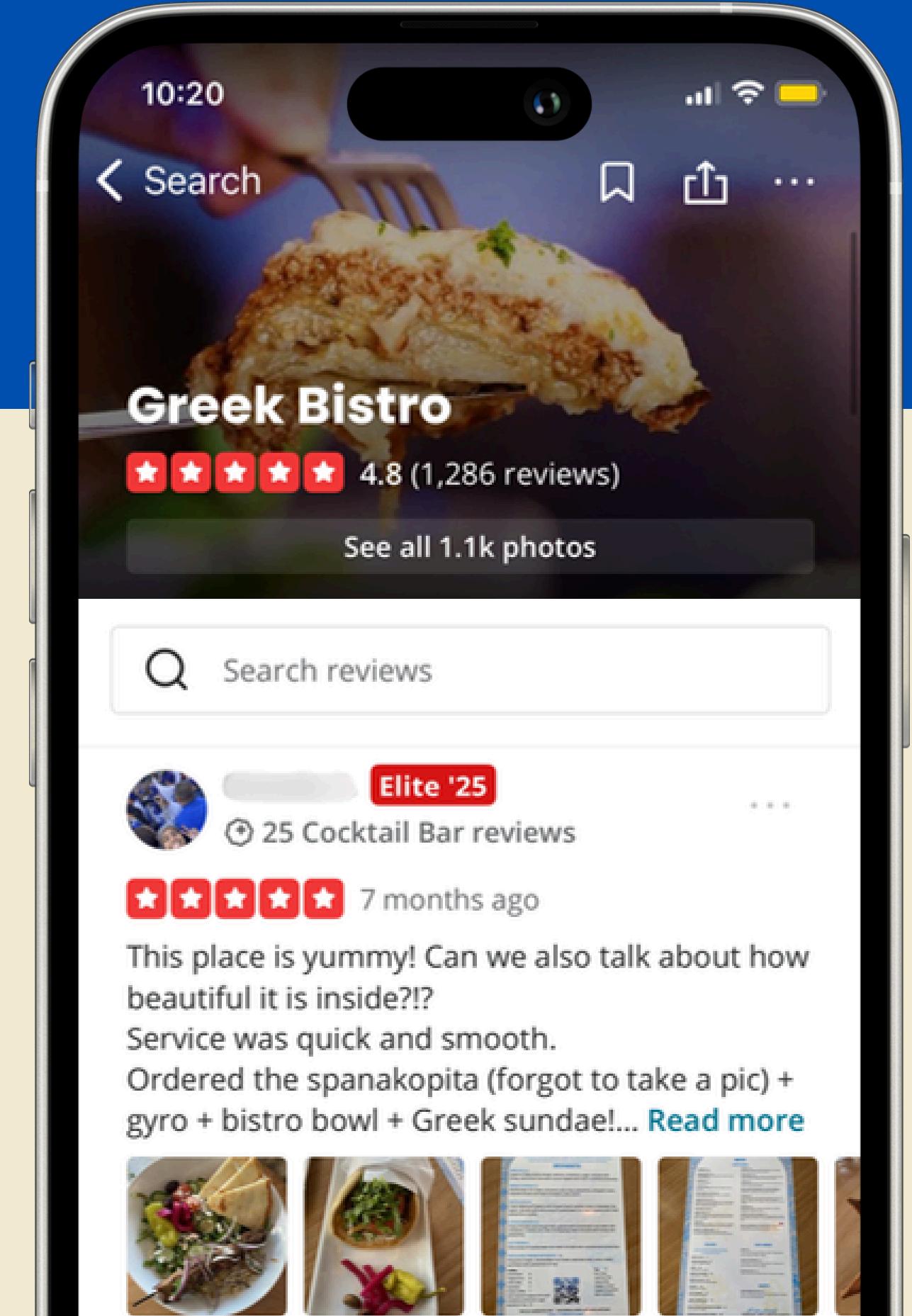
What is Yelp?

- Online review platform in the US
- Offers user review data for businesses



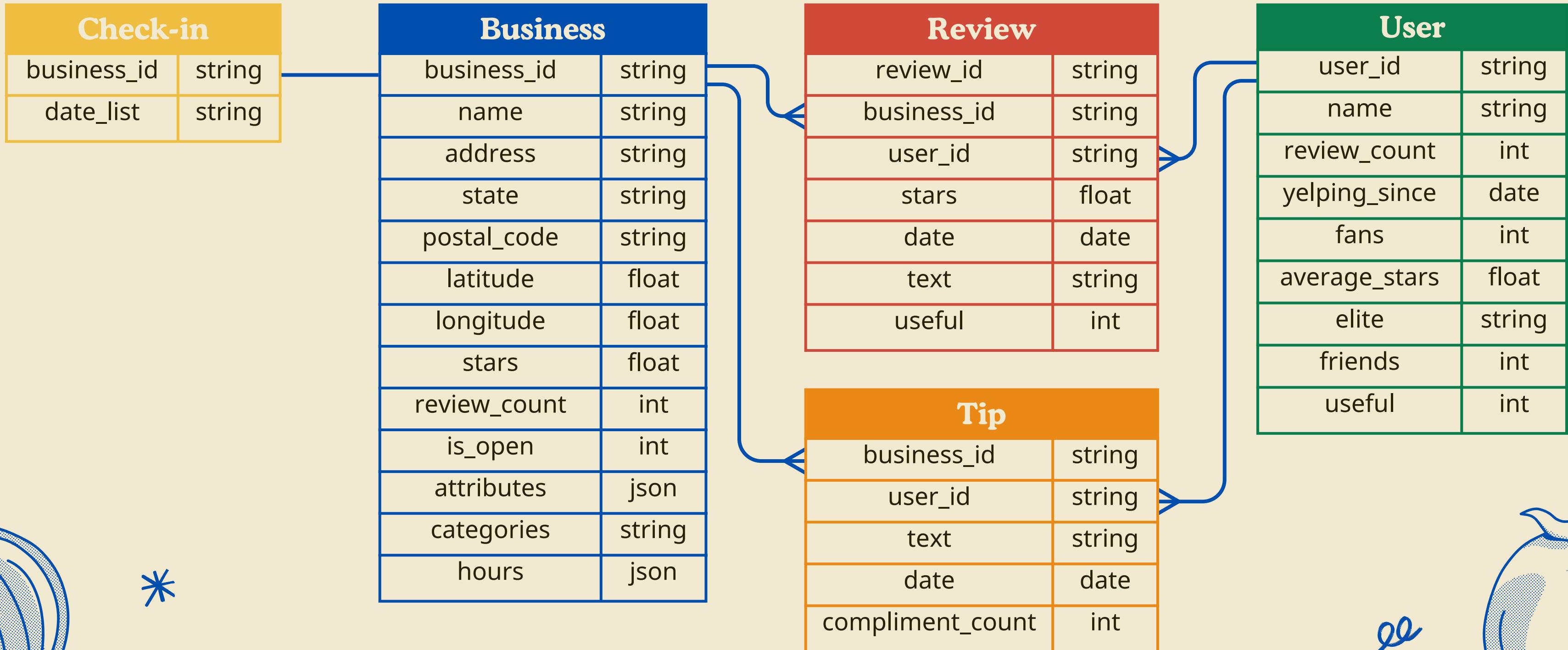
About the dataset

- 150K + business information
- 6.9M + review records





Dataset Structure



Data Preprocessing



- ✓ Filtered restaurant businesses only
- ✓ Selected restaurants with 25+ reviews to ensure data reliability
- ✓ Excluded businesses outside the US
- ✓ Performed feature engineering:
 - Created binary attributes
 - Generated location-based features

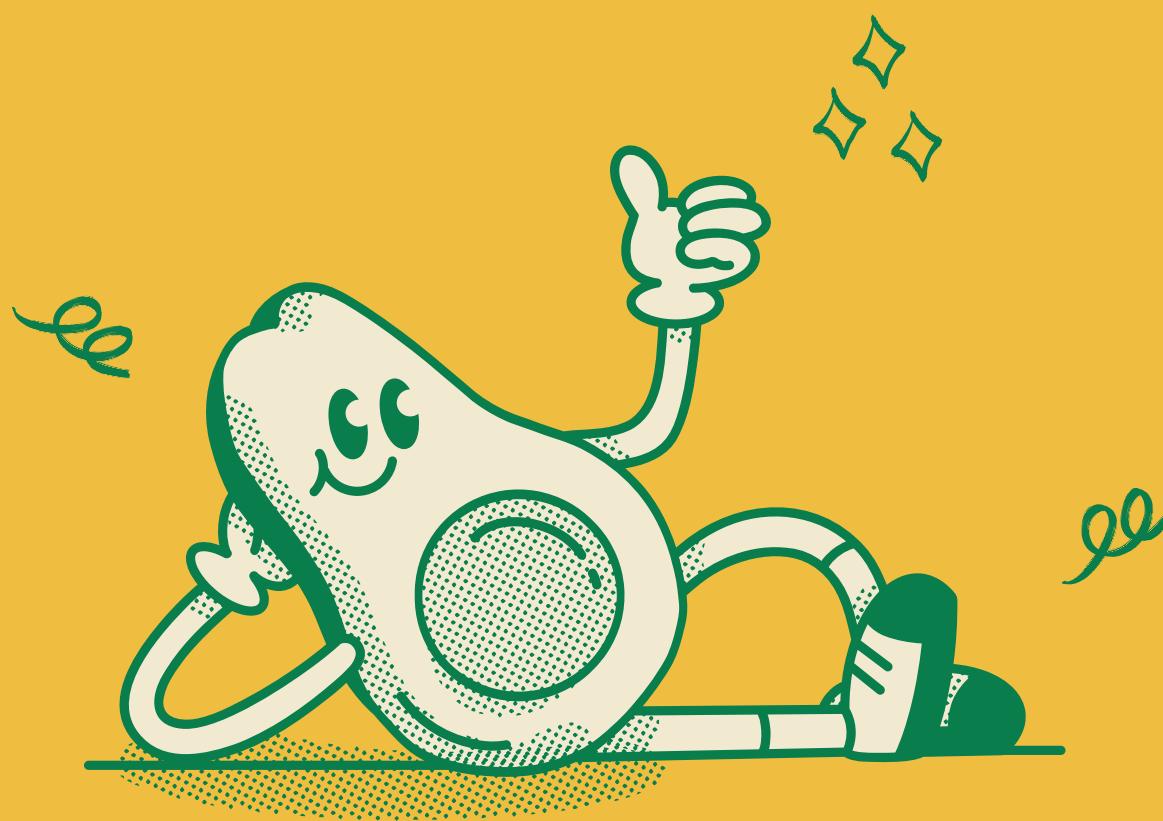


Final dataset size

19,500 restaurants × 92 features

Feature Engineering

: Examples of features generated through feature engineering



Attribute Derived

a_parking_space	If the restaurant offers a parking space
a_outdoor_seating	If the restaurant has outdoor seating
a_price_range	Average price range of the restaurant's menu
a_ambience	Overall ambience of the restaurant

Operating Hours Derived

open_weekend	If the restaurant operates on weekends
open_night	If the restaurant operates at night (after 12 a.m.)
total_weekly_hours	Total operating hours per week

Neighborhood information Derived

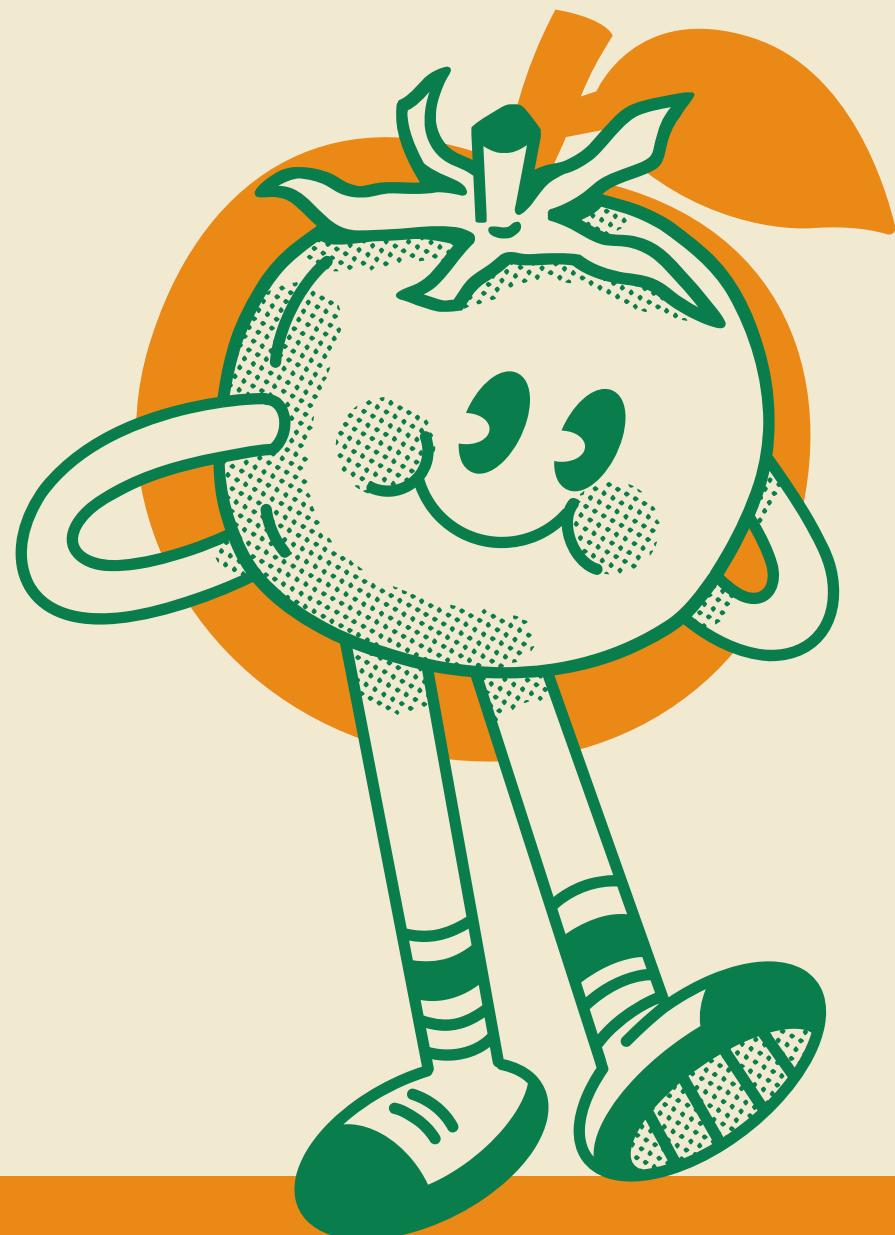
neighbor_density	Number of restaurants located near the restaurant
neighbor_similarity	How similar the restaurant is to neighboring stores
neighbor_avg_stars	Average ratings of neighboring restaurants

Review Derived

avg_sentiment	Average sentiment of reviews (positive vs. negative)
avg_review_length	Average length of reviews (word count)

Define performance metrics

Key Metrics



01

Stability Score

- Measures a restaurant's operational resilience
 - Its ability to withstand and recover from negative events such as rating drops or critical reviews

02

Loyalty Score

- Reflects how beloved a restaurant is
 - How many regular customers it has and how frequently they visit

03

Reliability Score

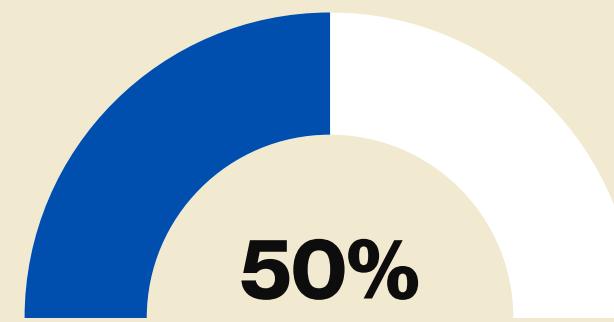
- Captures how credible a restaurant appears to potential customers
 - Measures the share of reviews written by influential reviewers.

Stability Score



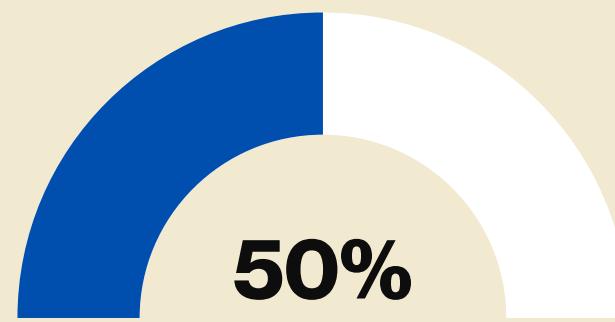
: measures a restaurant's operational resilience

✓ Used Features



1 Base Score

- Represents the fundamental stability of the business



2 Recovery Score

- Measures how quickly and strongly the restaurant's rating rebounds after a "low-rating shock"

⬇️ Low-rating shock

- It refers to a rapid decline in ratings of more than 0.8 points, caused by negative reviews

Formula :

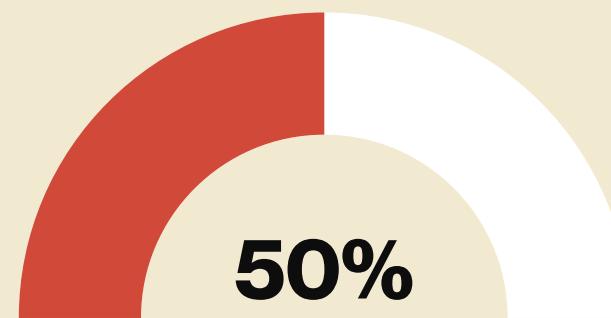
- $\text{base_score} = (\text{avg_stars} \times \log(\text{review_count})) \times (1 / (1 + \text{stars_std})) \times (1 - \text{low_star_ratio})$
- $\text{recovery_score} = 0.6 \times \text{recovery_speed} + 0.4 \times \text{recovery_breadth}$

Loyalty Score



: reflects how beloved a restaurant is

Used Features



50%

1 Regular Customer Score

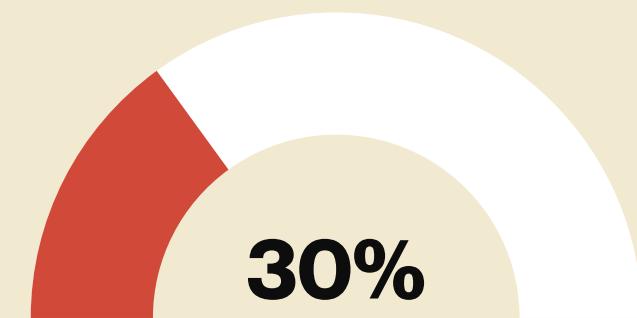
- Represents the ratio of repeat customers and their ratings



20%

2 Check-in Count

- Measures how many times the restaurant has been visited



30%

3 Check-in Interval

- Indicates visit frequency
- Complements check-in count, which can be biased by long operating periods

Formula :

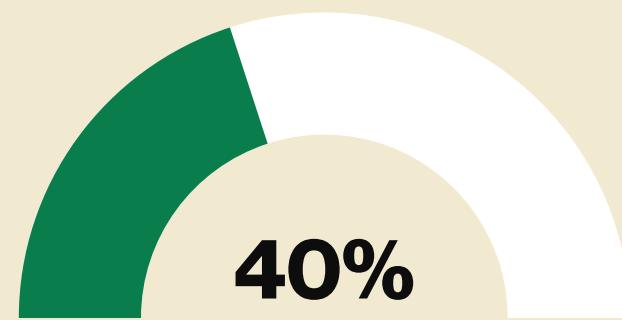
$$\text{regular_customer_score} = 0.5 + \text{regular_customer_ratio} \times 0.3 + (\text{regular_customer_rating} - 3) \times 0.1$$

Reliability Score



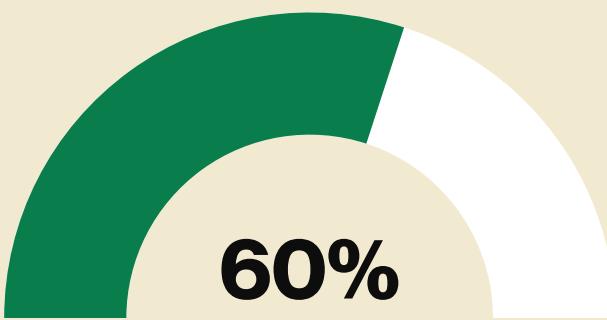
: captures how credible a restaurant appears

✓ Used Features



1 Expert Review Ratio

- Reflects the ratio of expert reviews to total reviews



2 Elite Review Ratio

- Yelp selects “elite” users each year based on engagement
- Measures the ratio of elite reviews to total reviews.

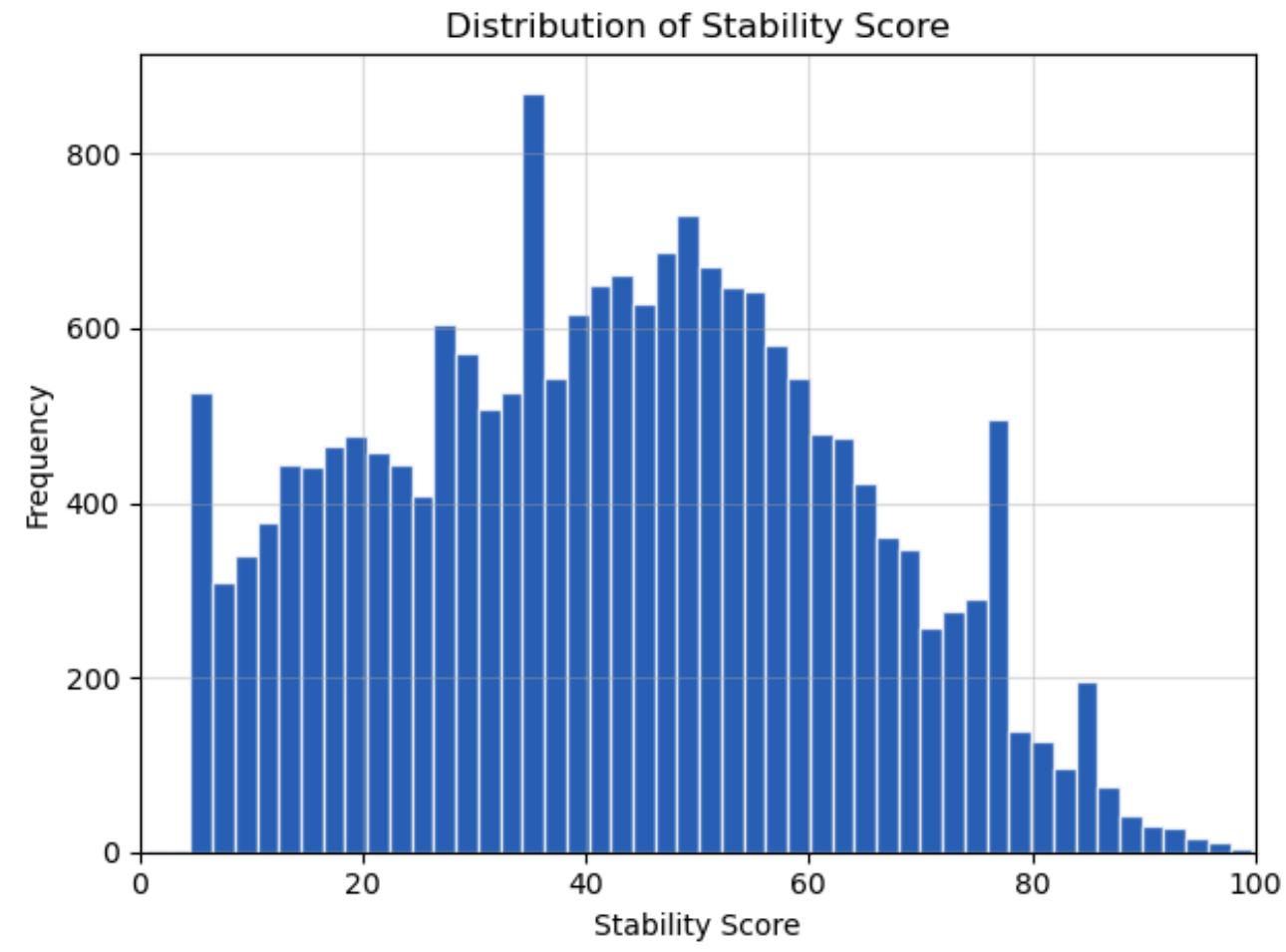


Expert User

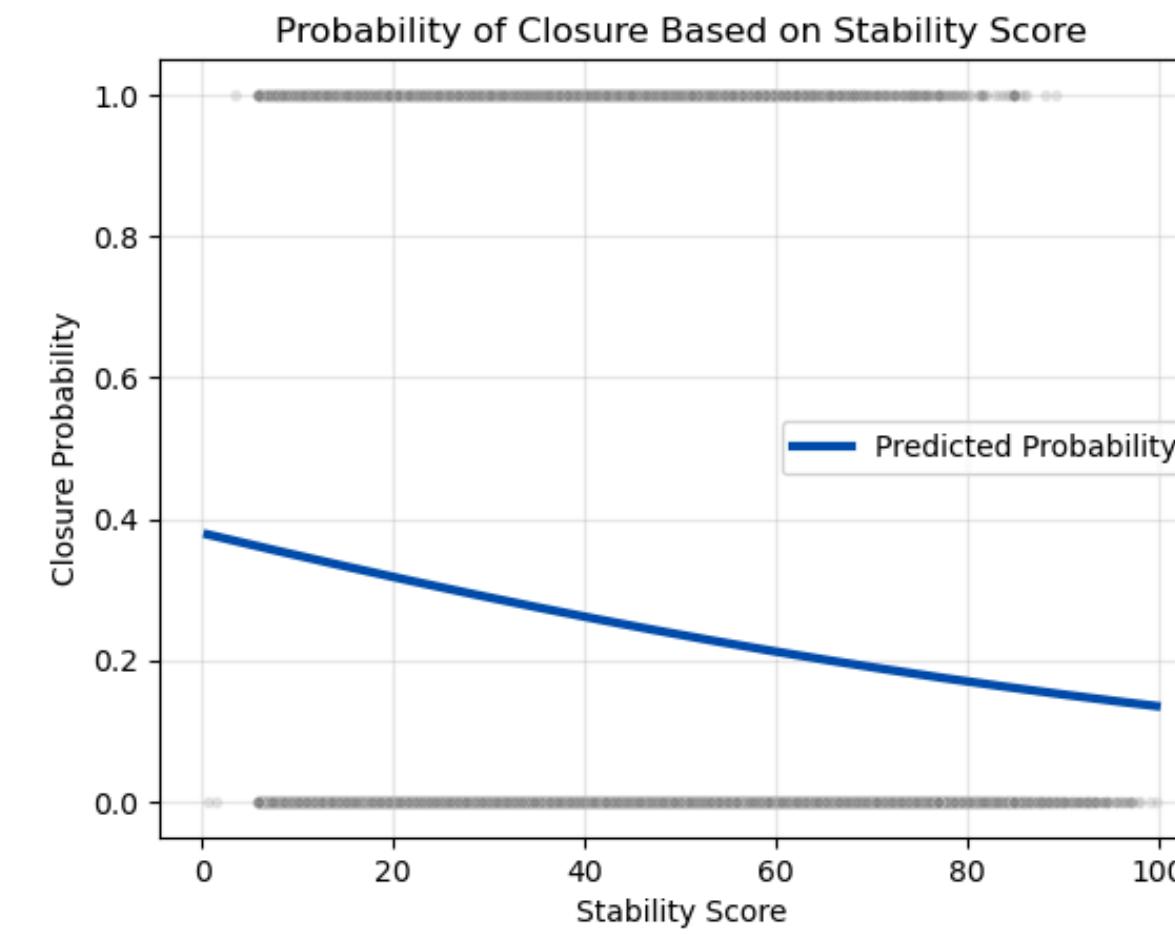
- “Experts” are users who focus on reviewing specific food categories
- If the restaurant’s category matches one of a user’s top 3 categories, that review is considered an expert review

Stability Score

Distribution



Statistical Testing



- Higher stability scores are associated with lower closure probabilities

X

**Stability
Score**

y

**Business
closure**

Method

**Logistic
Regression**

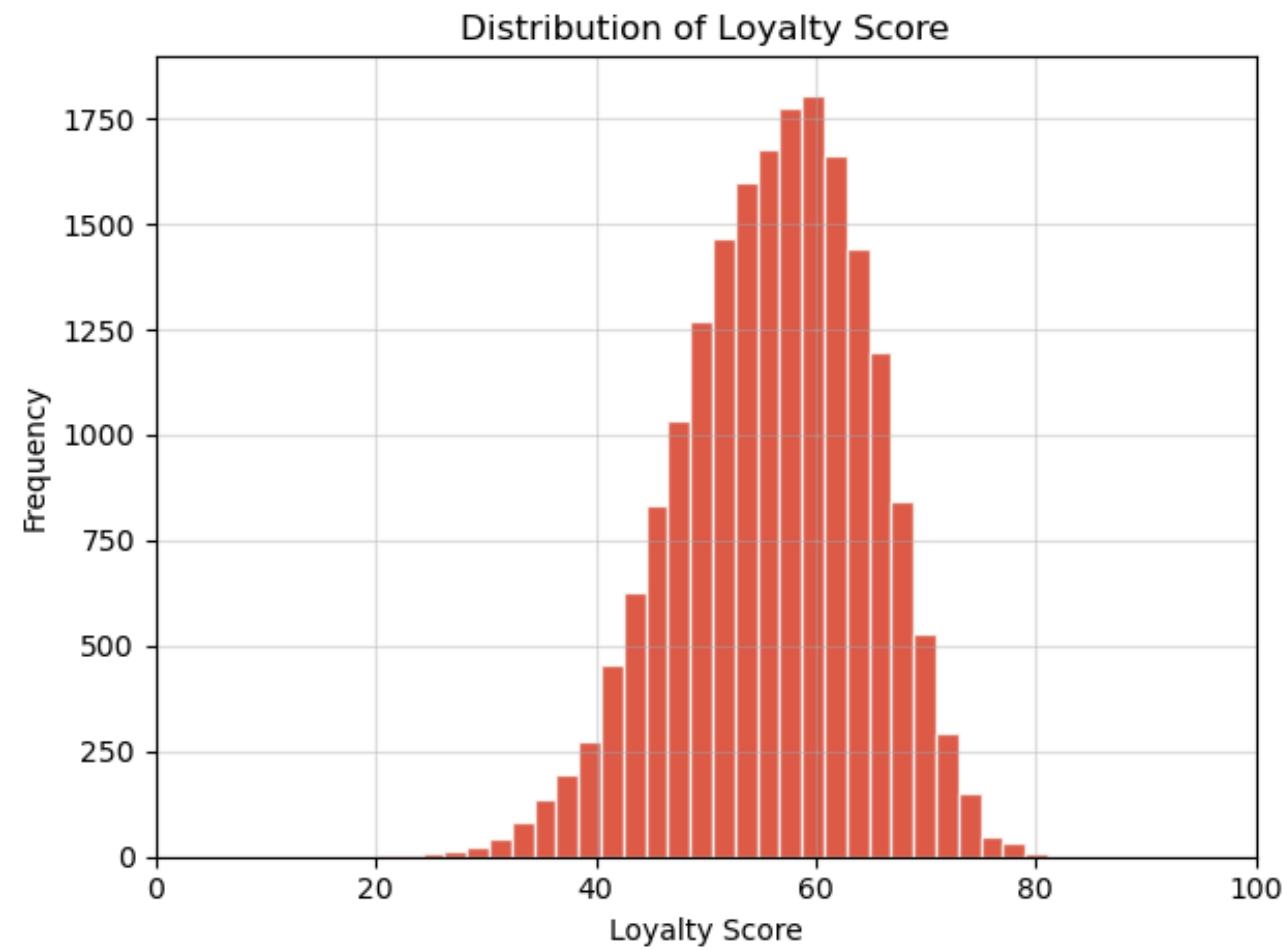
P-value

< 0.001

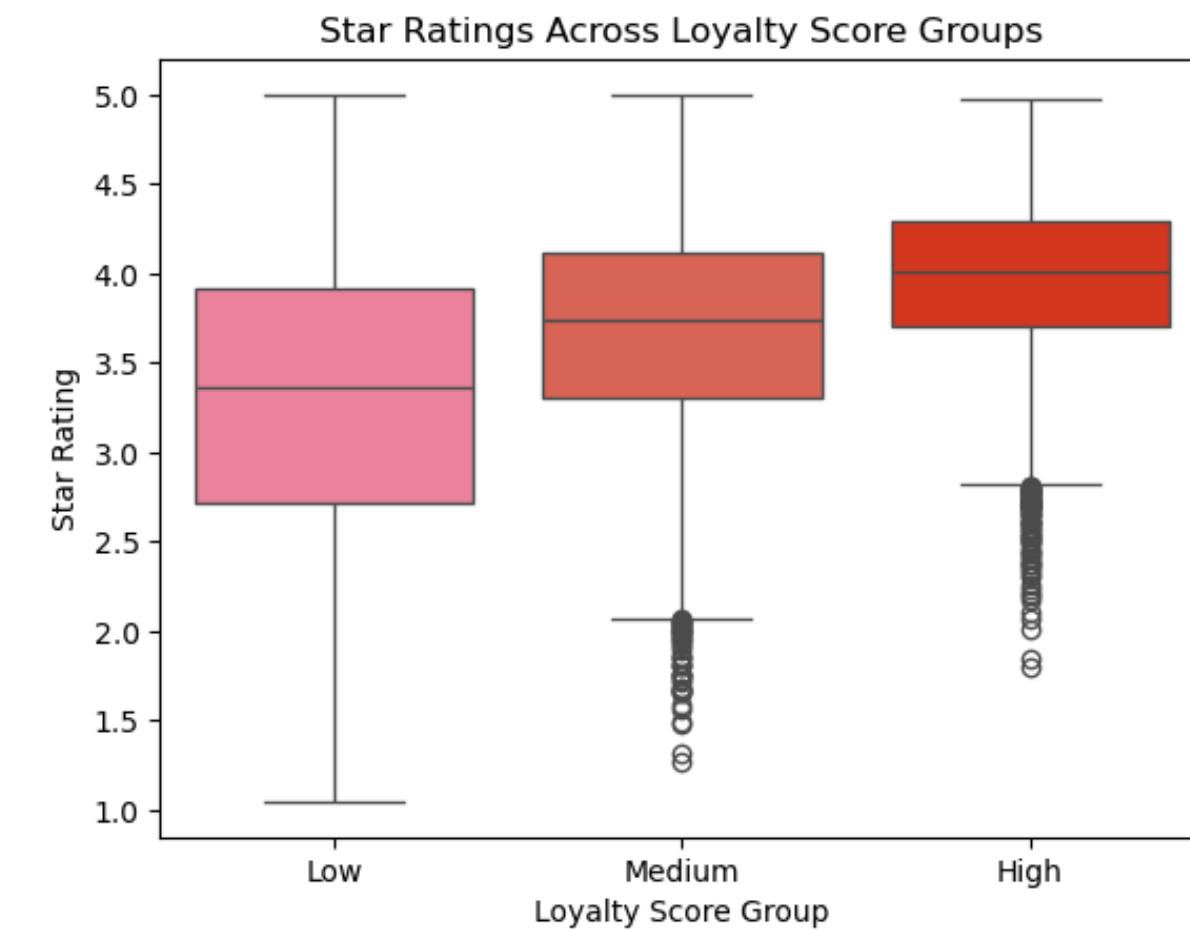
Loyalty Score



Distribution



Statistical Testing



- Restaurants with higher loyalty scores tend to maintain higher average ratings

X

Loyalty score group

y

Star rating

Method

Welch's ANOVA

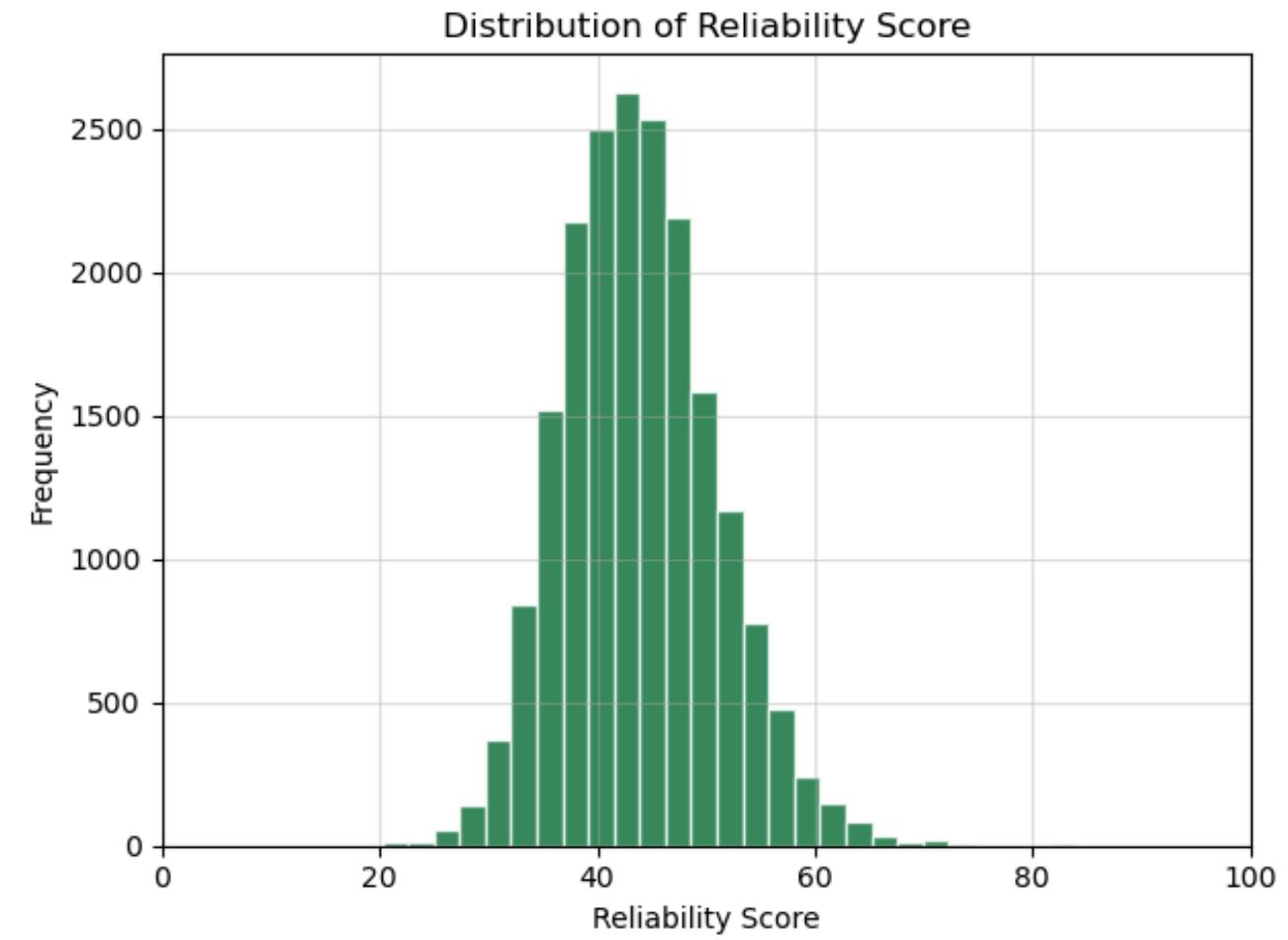
P-value

< 0.001

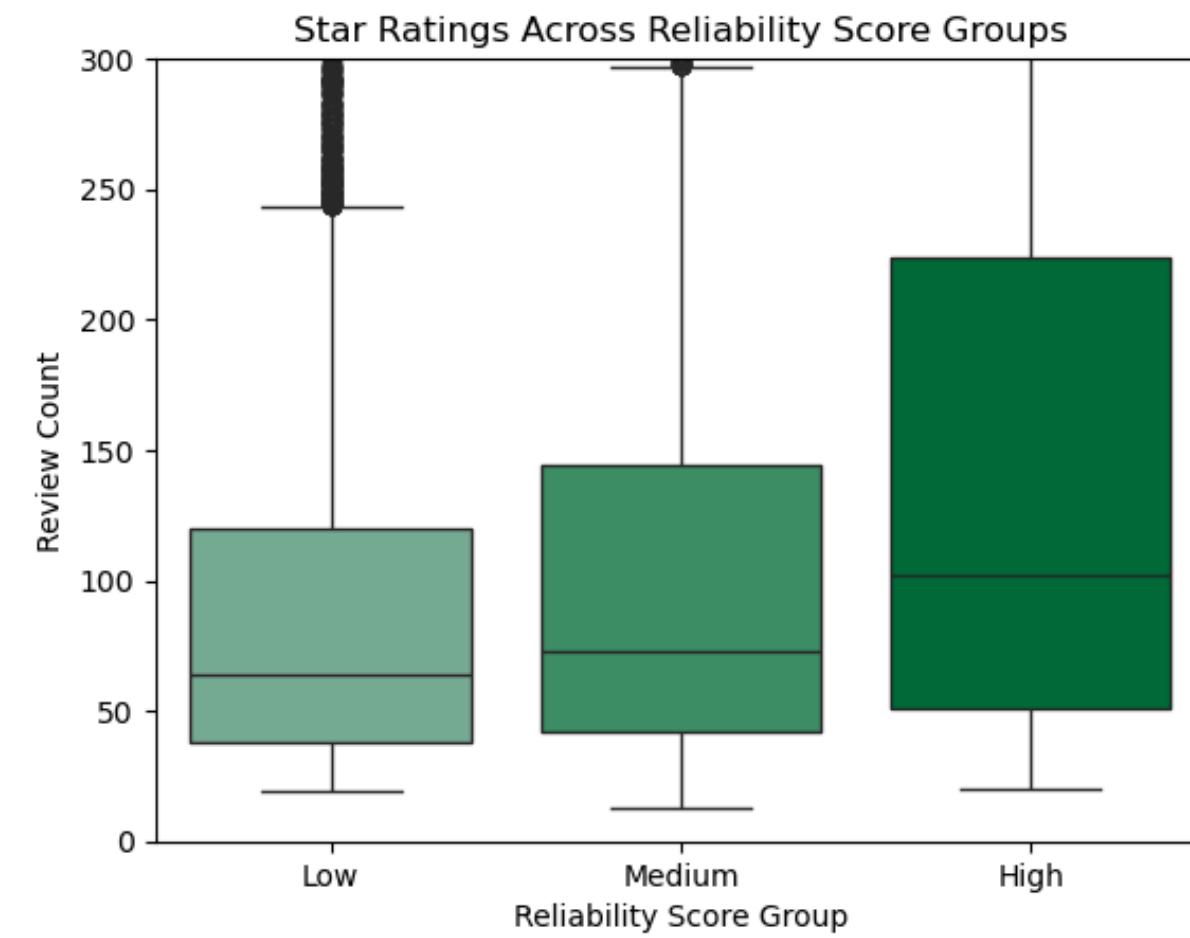
Reliability Score



Distribution



Statistical Testing



- Restaurants with higher reliability scores generate more first-time visitors and reviews

X

Reliability score group

y

Review count

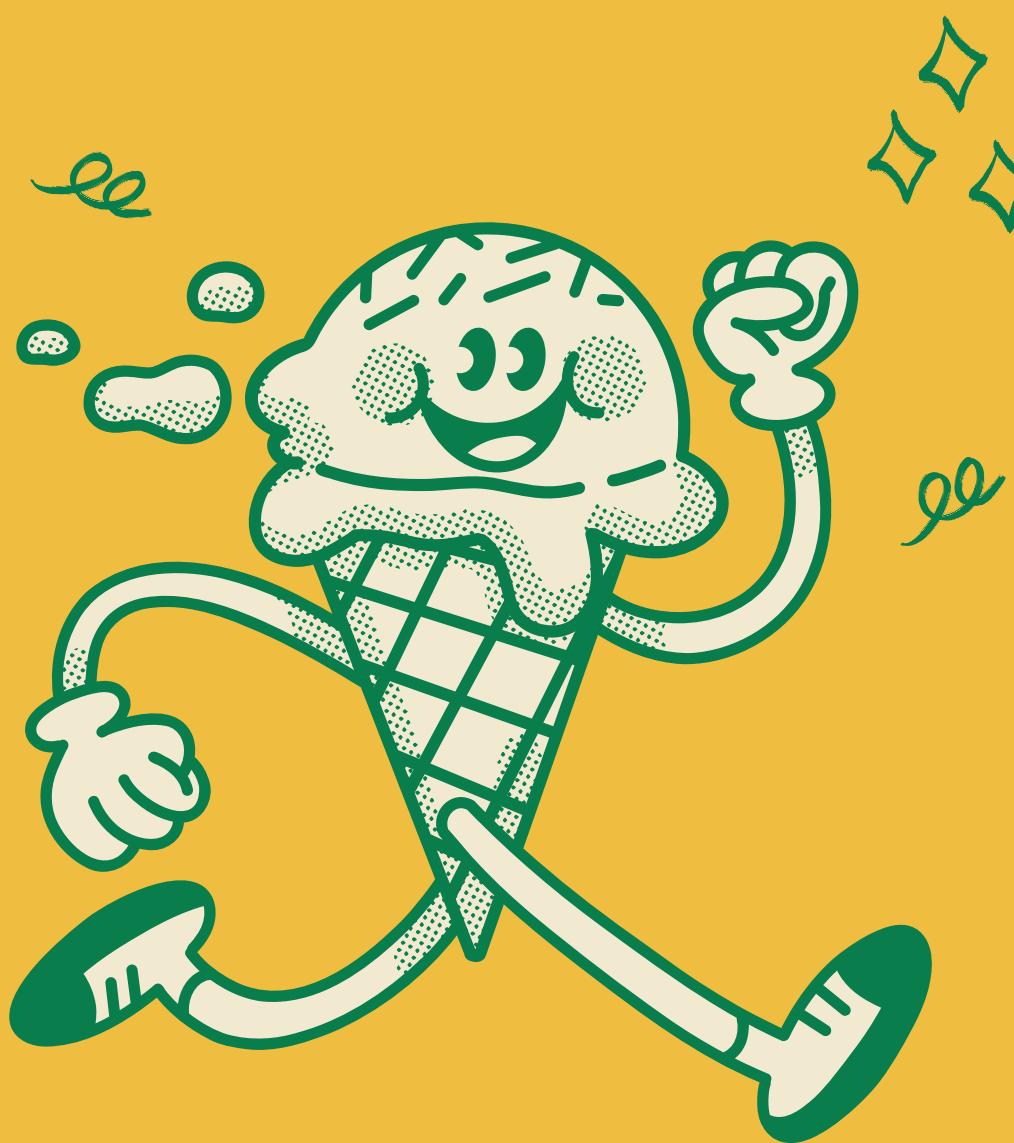
Method

Welch's ANOVA

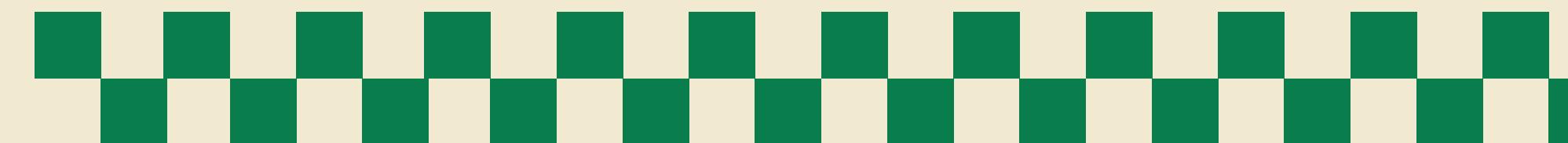
P-value

< 0.001

ML Overview



- ✓ Build regression models to predict the three custom metrics
- ✓ By identifying key features that increase or decrease each scores, **drive business insights** and **develop actionable recommendations**
- ✓ Selected CatBoost for its stable performance across all scores
- ✓ Used Optuna to optimize CatBoost hyper-parameters (minimizing MAE)



ML Performance

✓ Evaluate baseline model

Model	RMSE	MAE	R ²
CatBoost	4.66	3.56	0.695
LightGBM	4.65	3.57	0.696
Gradient Boost	4.66	3.58	0.695
Random Forest	4.72	3.63	0.687
SVM	4.85	3.73	0.669
XGBoost	4.87	3.74	0.667
AdaBoost	5.16	4.11	0.628
KNN	6.00	4.72	0.494
Decision Tree	6.94	5.31	0.323

✓ Final Performance (Optimized)

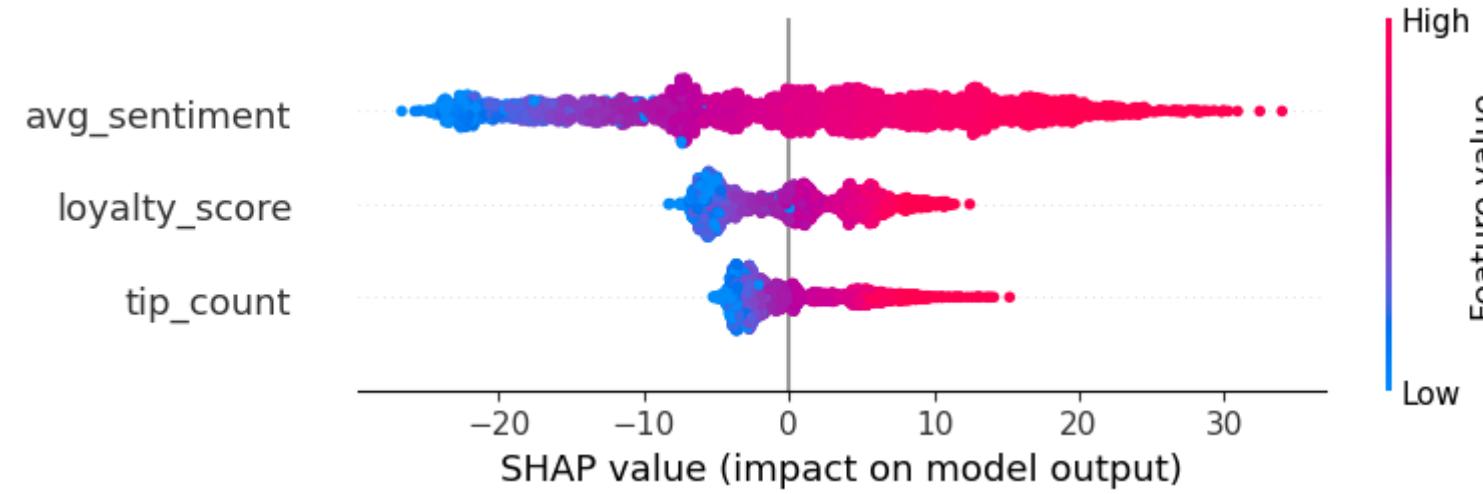
Stability Score	
RMSE	11.09
MAE	8.67
R ²	0.70

Loyalty Score	
RMSE	11.09
MAE	8.67
R ²	0.70

Reliability Score	
RMSE	11.09
MAE	8.67
R ²	0.70

Stability Score

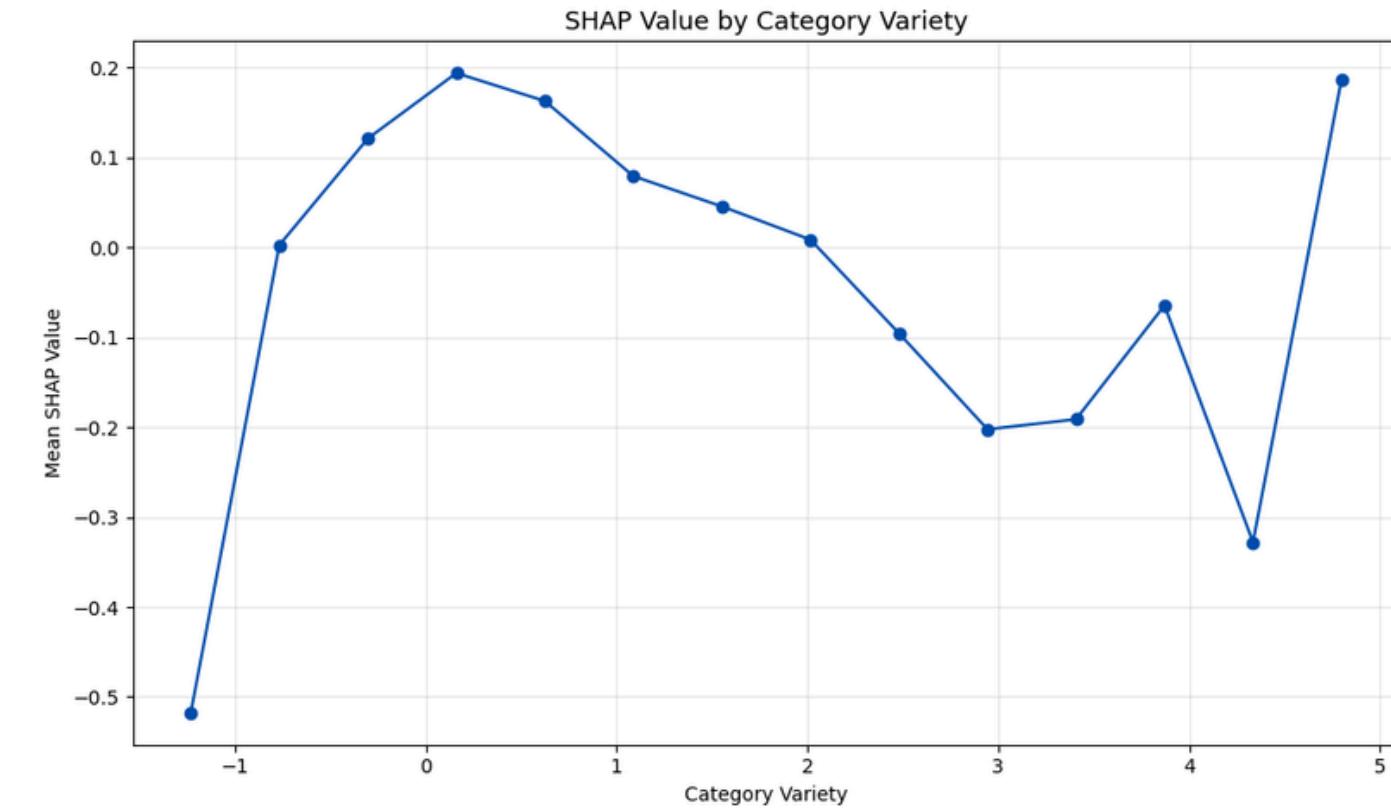
Insight ①



* sentiment : if the reviews are positive or negative

- ✓ Sentiment of reviews has the strongest impact

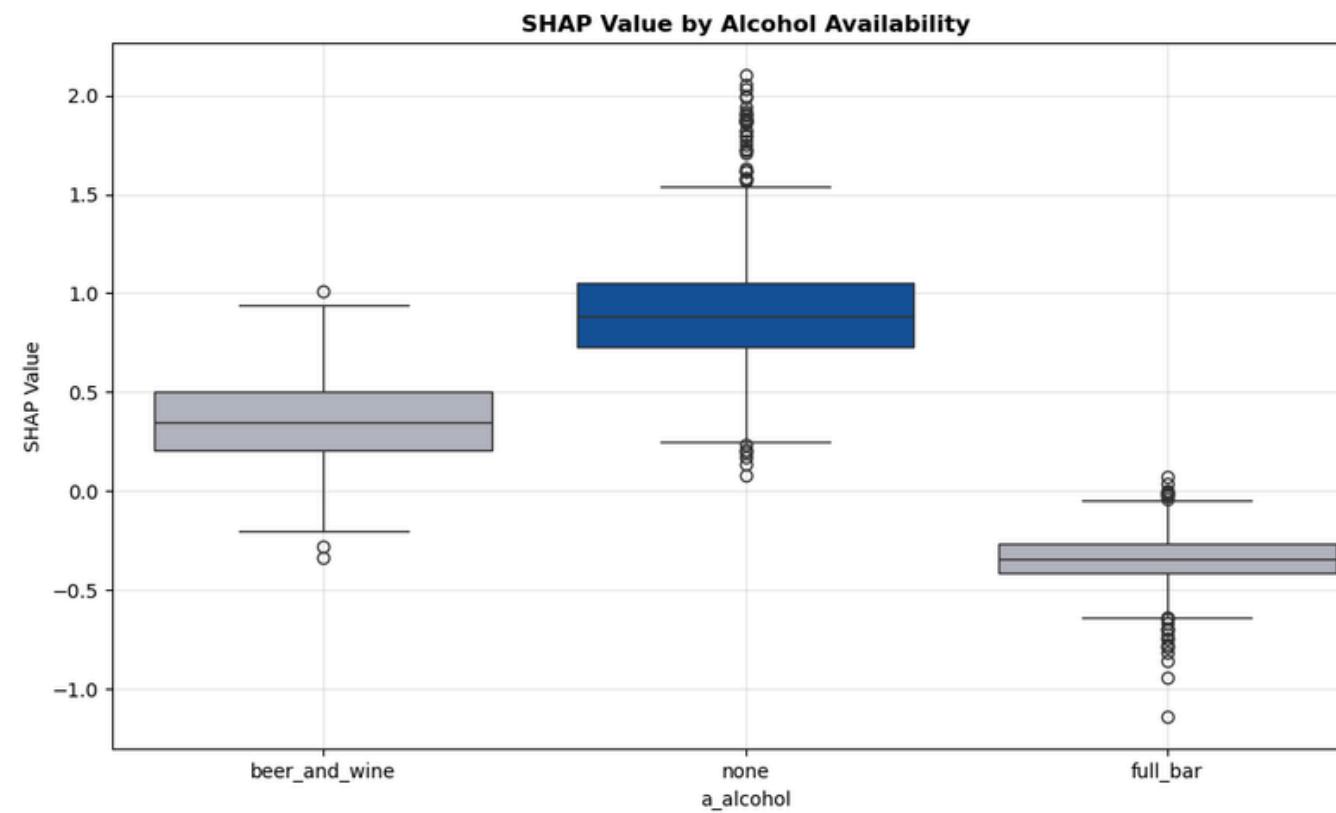
Insight ②



- ✓ Moderate category variety improves stability

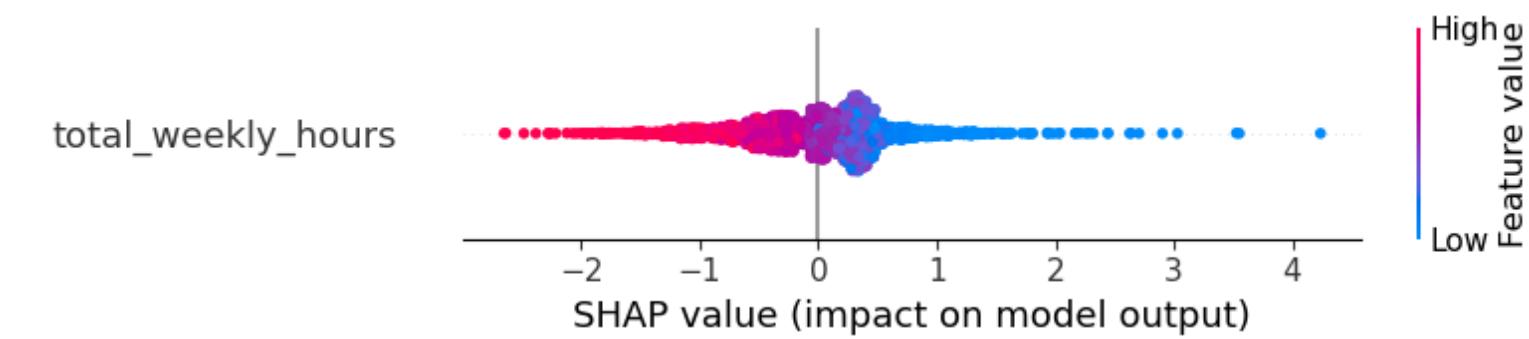
Stability Score

Insight ③



✓ Restaurants without alcohol service tend to be more operationally stable

Insight ④

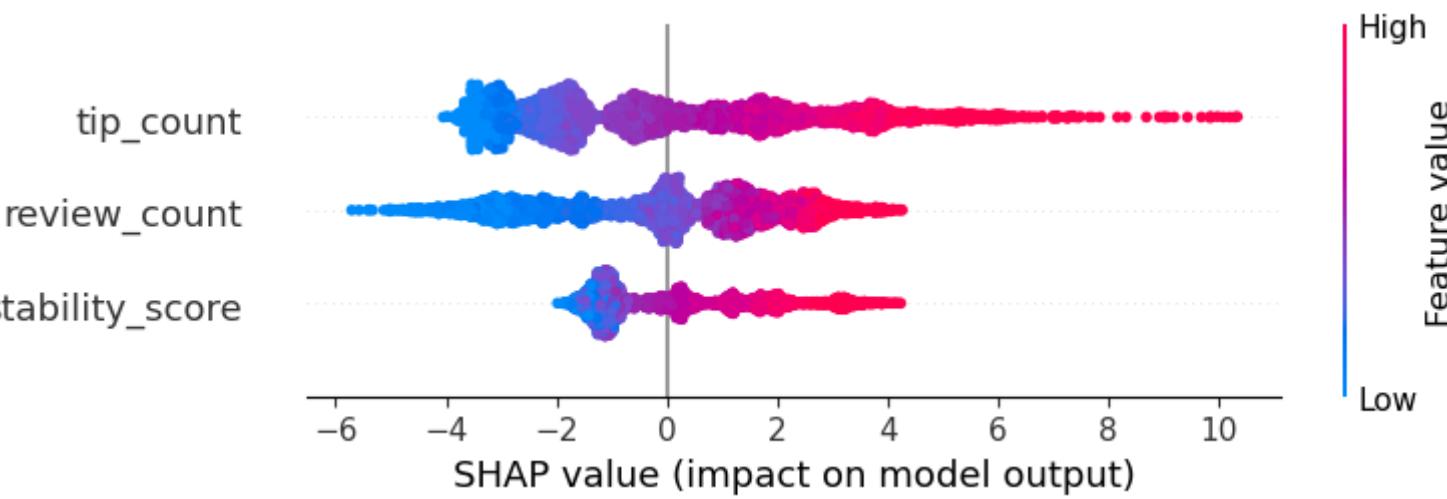


✓ Excessively long operating hours negatively affect stability

Loyalty Score



Insight ①

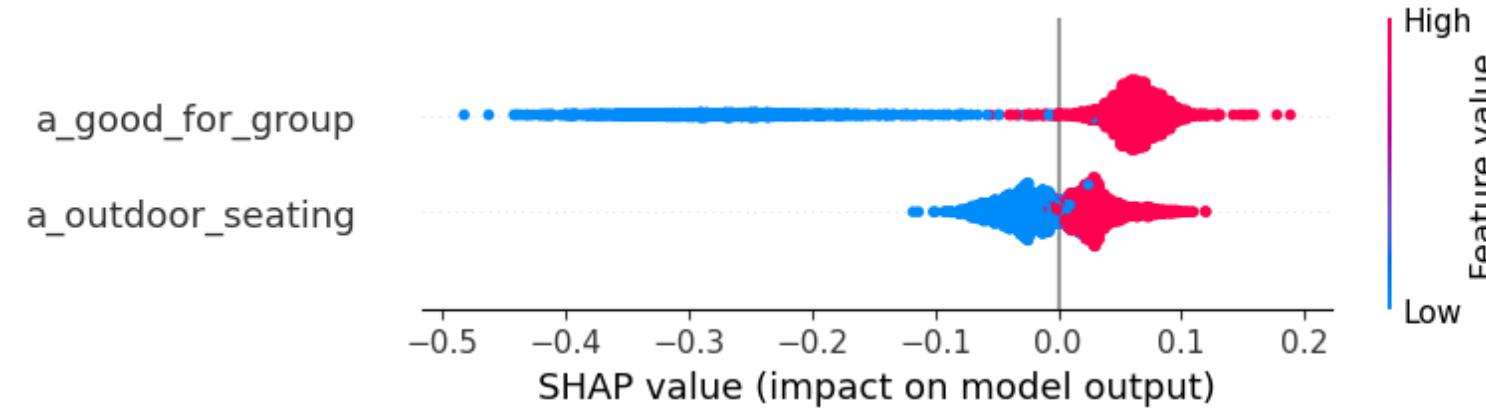


* tip : short review for a restaurant



Review volume is the strongest predictor of loyalty

Insight ②

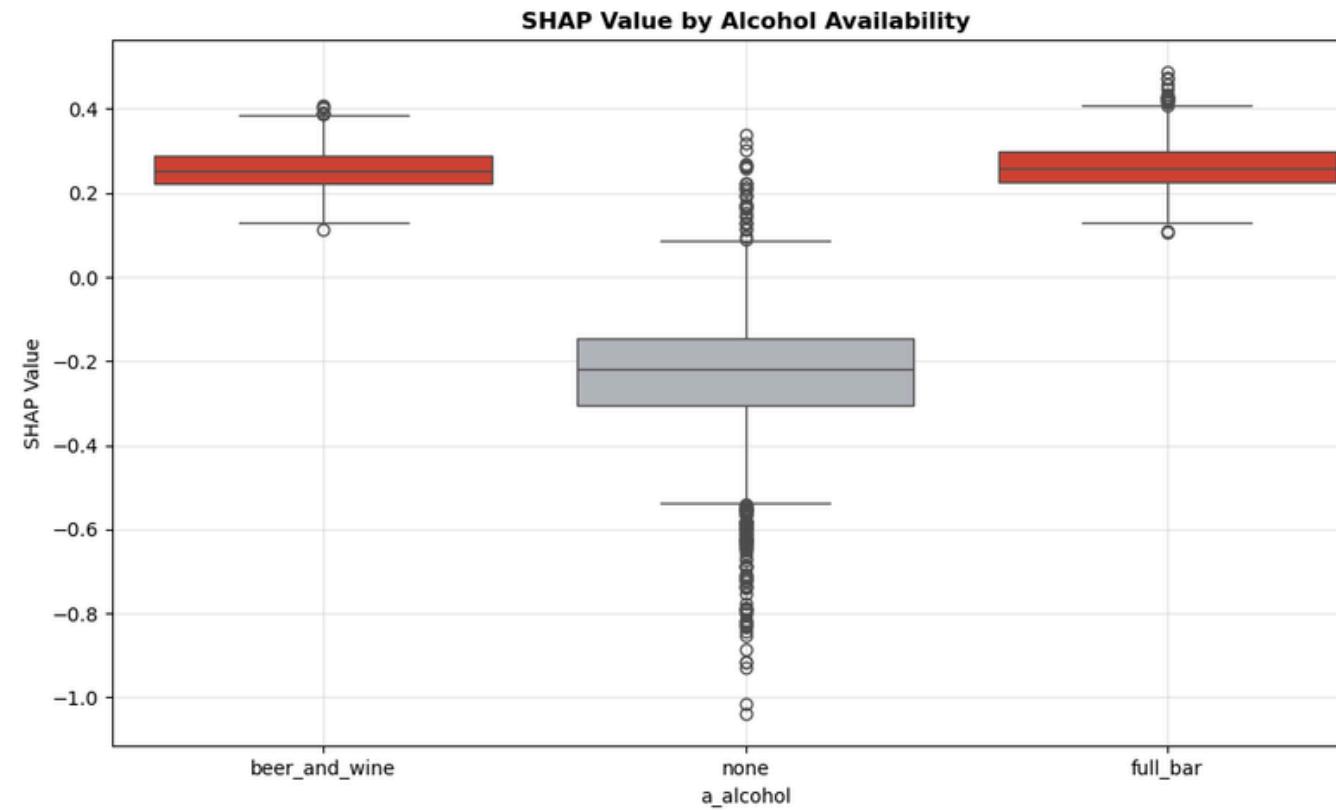


Outdoor seating and group-friendly environments increase repeat visits

Loyalty Score

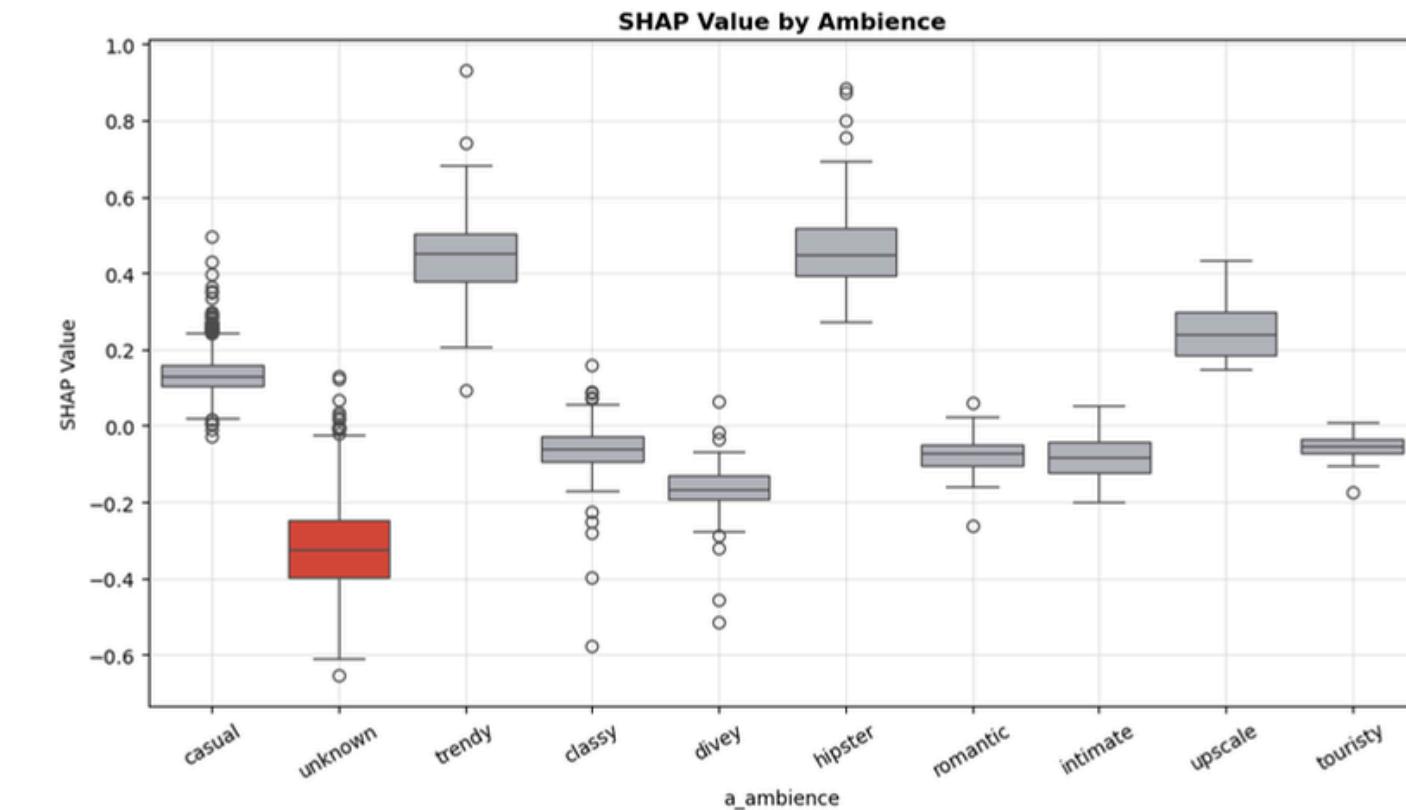


Insight ③



✓ Alcohol service increases loyalty, despite reducing stability

Insight ④

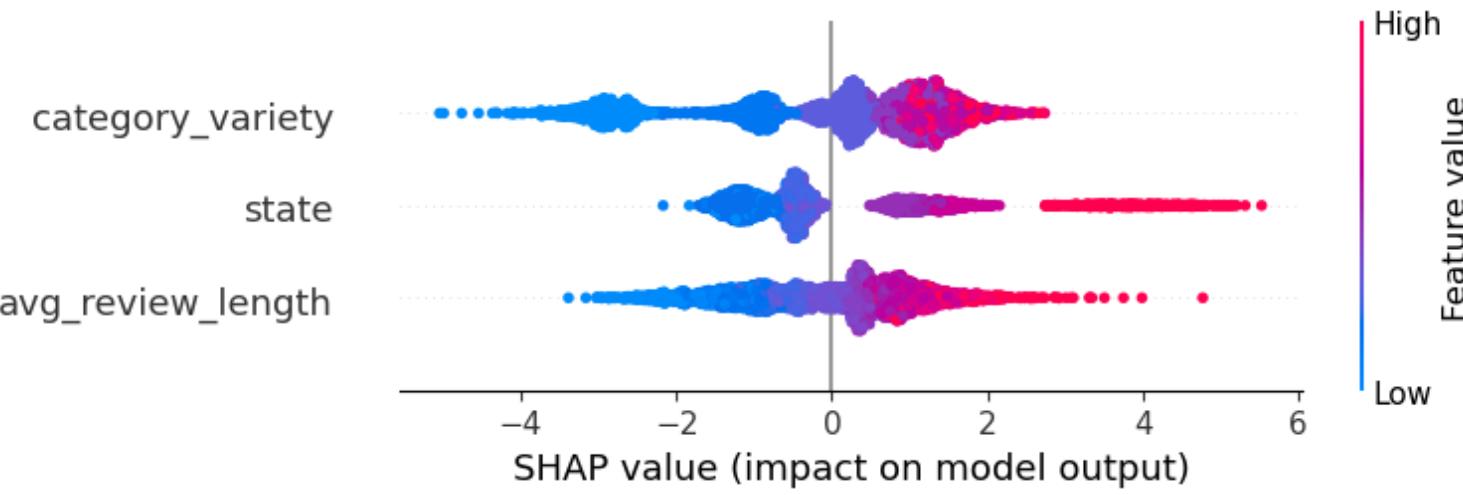


✓ Unclear ambience correlates with lower loyalty

Reliability Score

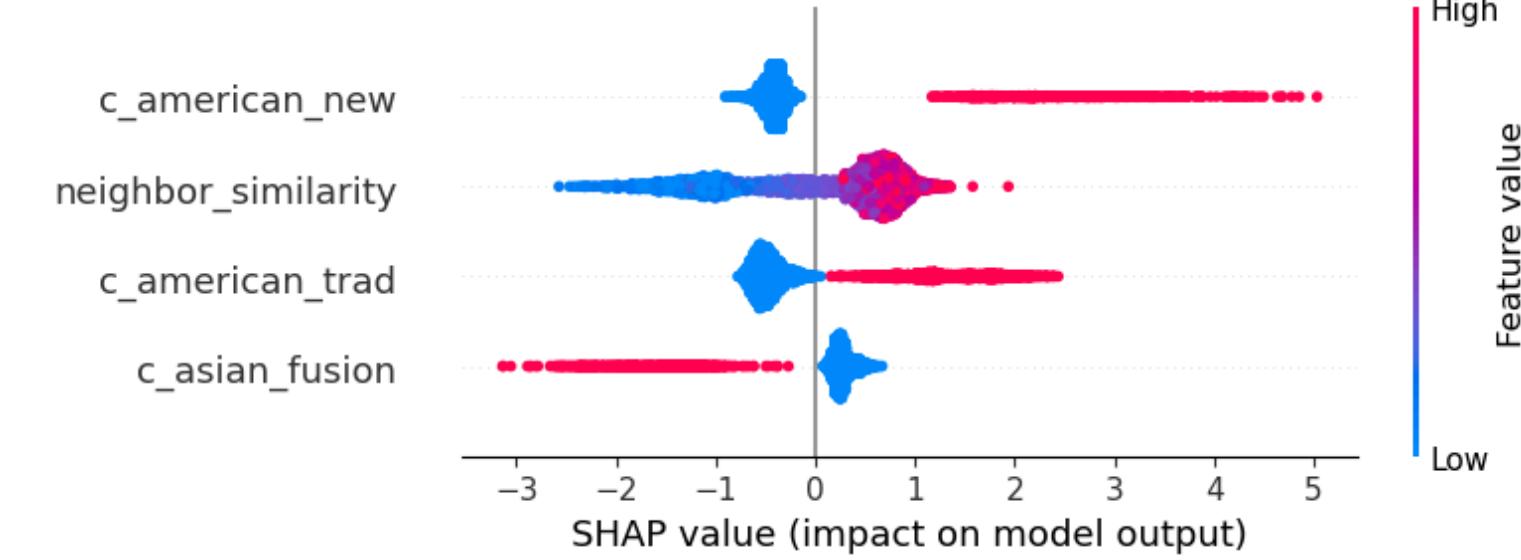


Insight ①



- ✓ Breadth of food categories is the most influential factor

Insight ②

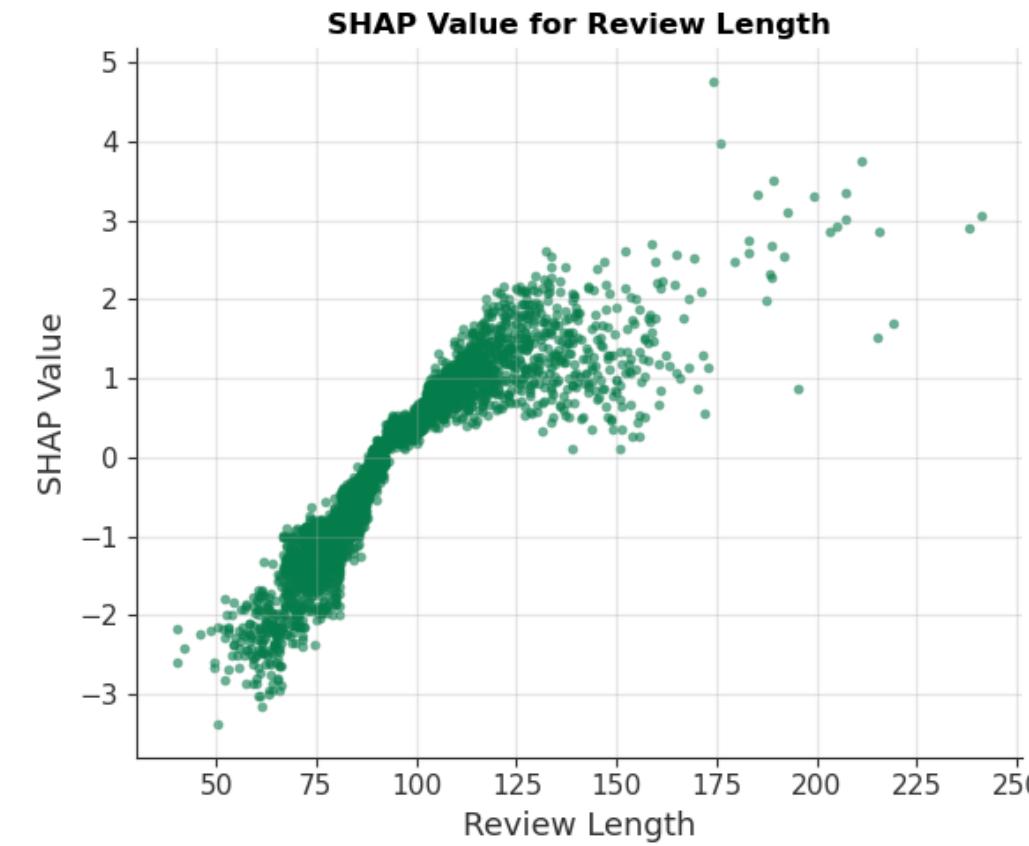


- ✓ Restaurants serving familiar, mainstream cuisines achieve higher reliability than niche or experimental concepts

Reliability Score

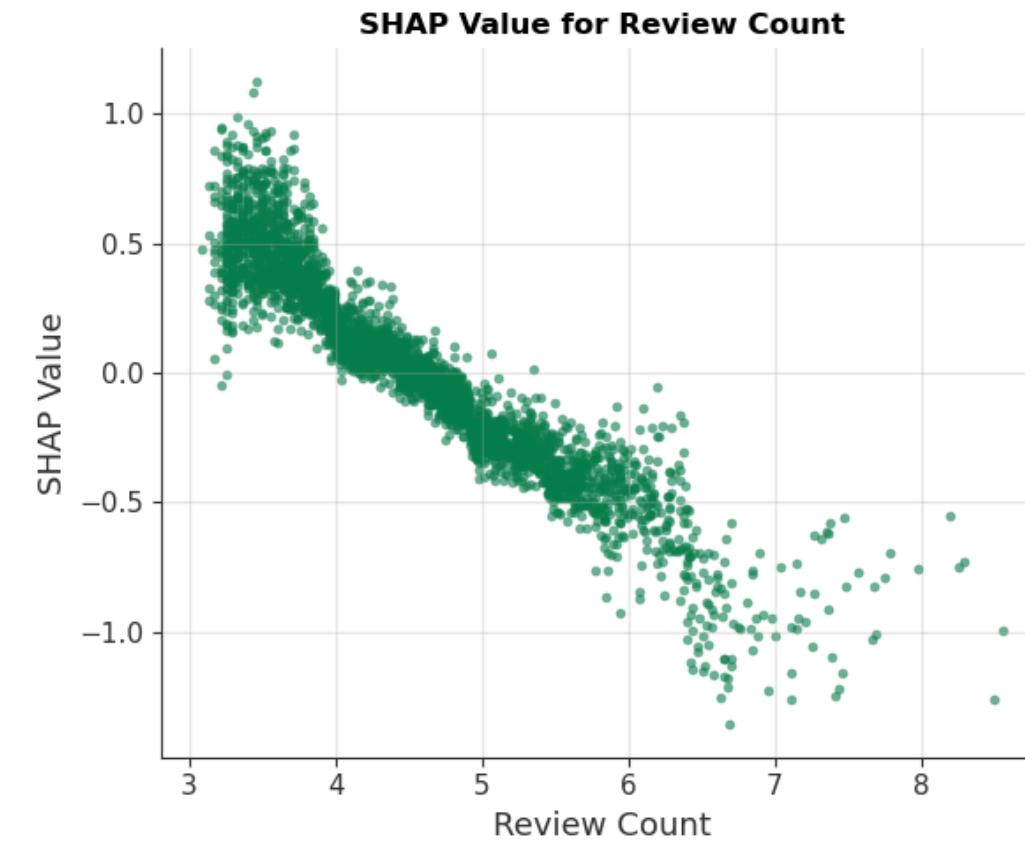


Insight ③



Quality of reviews matters more than sheer quantity

Insight ④



A high number of low-quality reviews negatively affects the reliability score

Improving Stability



- ✓ Focus on core menu offerings instead of excessive variety
- ✓ Monitor customer feedback closely and address recurring complaints
- ✓ Consider regular off-days for more stable operations
- ✓ Be cautious with alcohol service depending on target customer base

Building Loyalty



- ✓ Invest in outdoor seating and group-friendly layouts
- ✓ Develop a clear and consistent brand atmosphere or concept
- ✓ Offer alcohol where it aligns with the restaurant's concept and audience

Enhancing Reliability



- ✓ Encourage high-quality, thoughtful customer reviews
- ✓ Balance uniqueness with familiarity in menu design
- ✓ Prioritize accessible cuisine before expanding into niche categories
(e.g., fusion concepts)