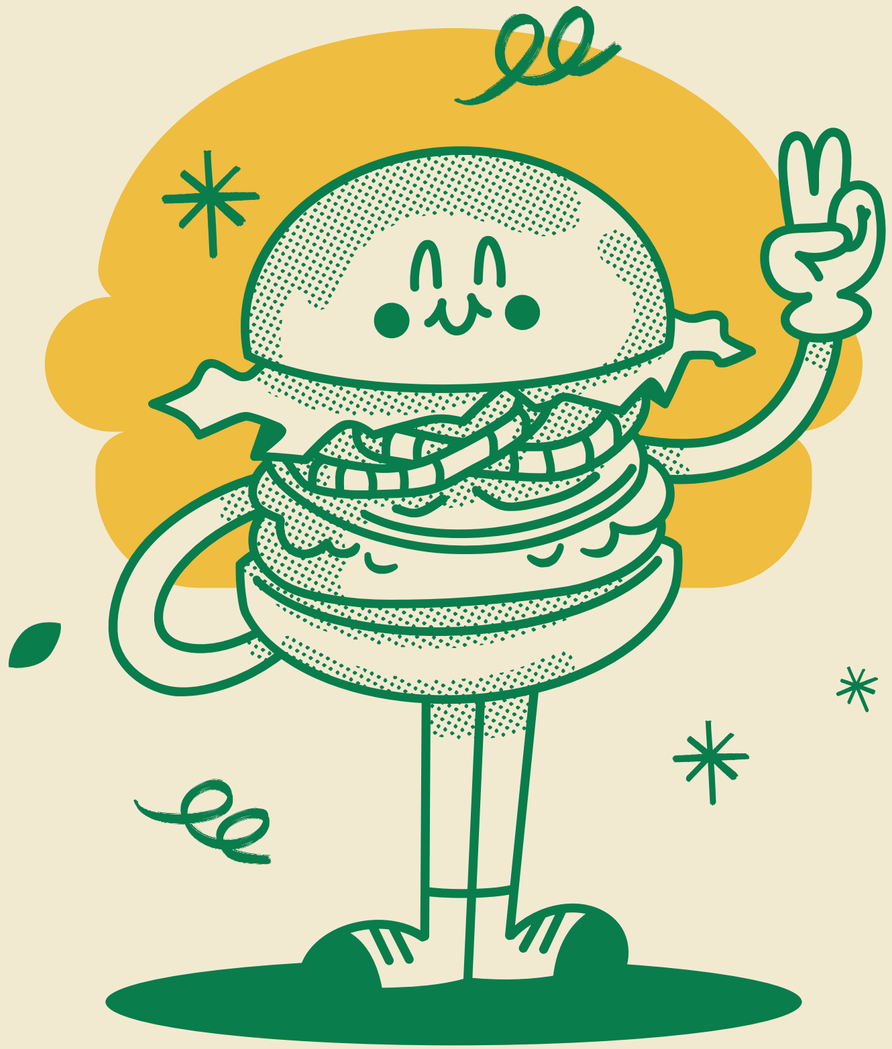
Data Analytics Project by Yennie Lee

Data-Driven Insights into Restaurant Success



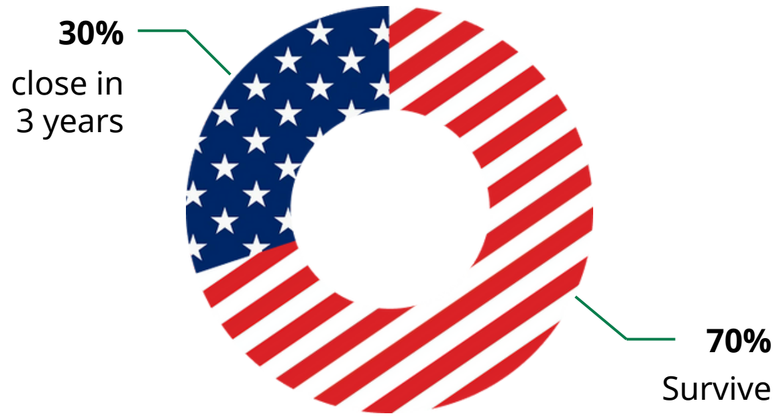
CONTENT

Overview

Project Background

Problem

Restaurant Closure Rate



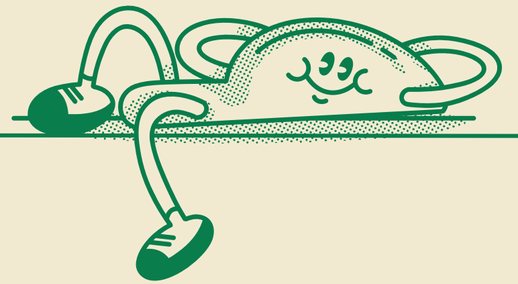
*The US Bureau of Labor Statistics 2025

30% of restaurants in the US close
their business in the first 3 years

Goal

By analyzing review data,
get **insights for restaurant owners**

Major Steps

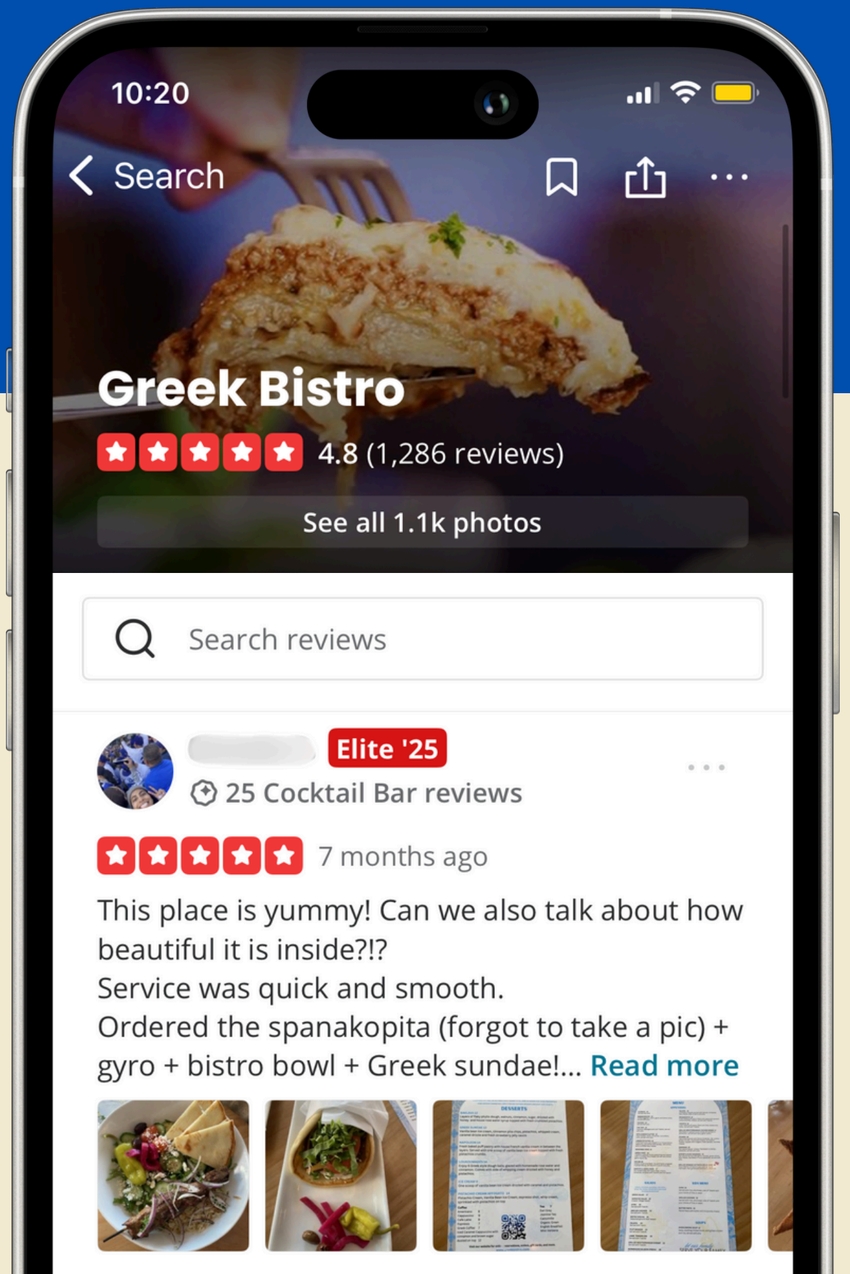


1. Define custom performance metrics
2. Validate custom metrics using statistical testing
3. Identify key features that positively or negatively impact each metric
4. Translate analytical insights into actionable recommendations

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

Check-in	
business_id	string
date_list	string

Business	
business_id	string
name	string
address	string
state	string
postal_code	string
latitude	float
longitude	float
stars	float
review_count	int
is_open	int
attributes	json
categories	string
hours	json

Review	
review_id	string
business_id	string
user_id	string
stars	float
date	date
text	string
useful	int

Tip	
business_id	string
user_id	string
text	string
date	date
compliment_count	int

User	
user_id	string
name	string
review_count	int
yelping_since	date
fans	int
average_stars	float
elite	string
friends	int
useful	int

Data Preprocessing

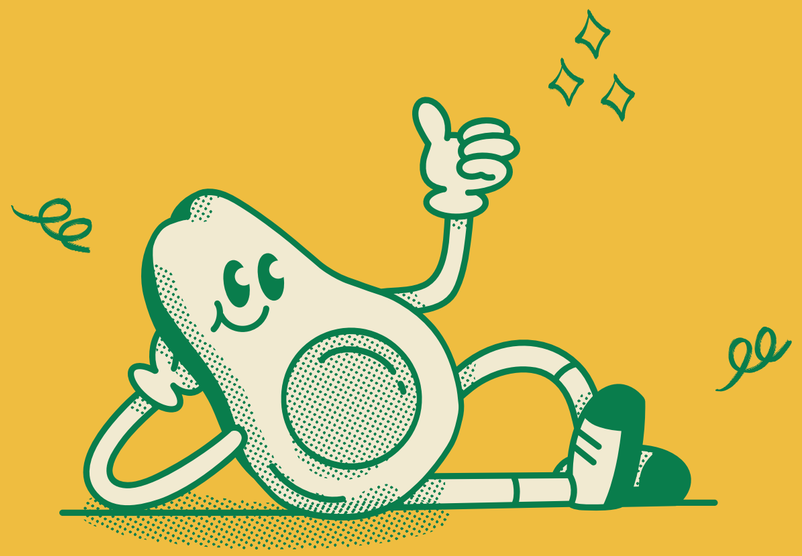
- Filtered restaurant businesses only
- Selected restaurants with 25+ reviews to ensure data reliability
- Excluded businesses outside the US
- Performed feature engineering:
  - Created binary attributes
  - Generated location-based features

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

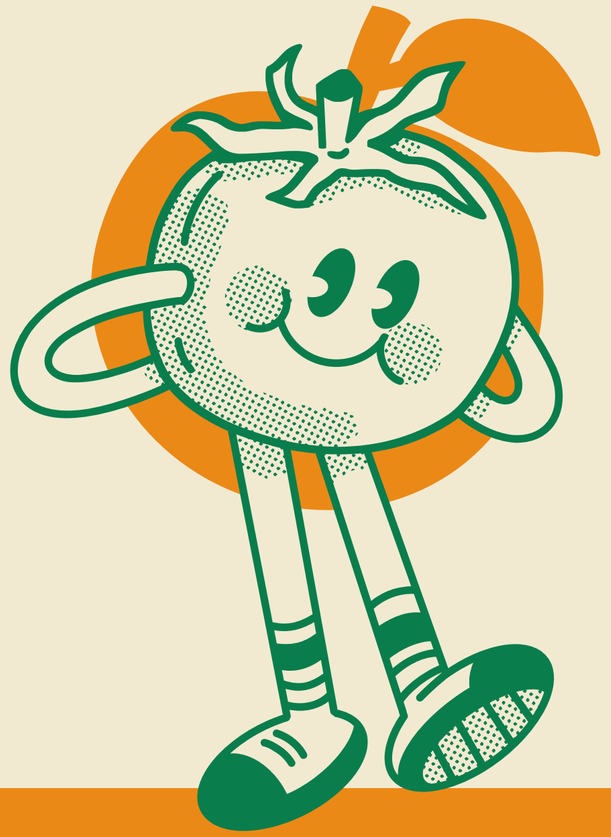
Operating Hours Derived	
open_weekend	If the restaurant operates on weekends
open_night	If the restaurant operates at night (after 12 a.m.)
total_weekly_hours	Total operating hours per week

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

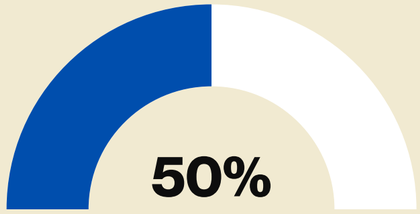
- Captures how credible a restaurant appears to potential customers
- Measures the share of reviews written by influential reviewers.

🍽 Define performance metrics

Stability Score

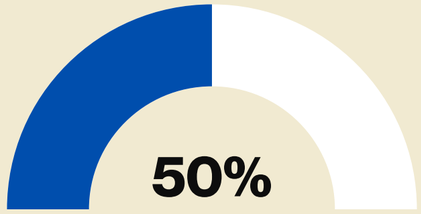
: measures a restaurant’s operational resilience

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 :

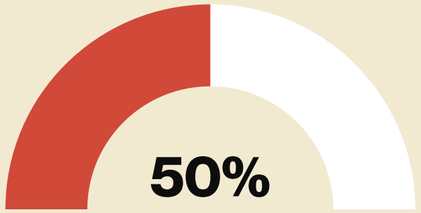
- base_score = (avg_stars × log(review_count)) × (1 / (1+stars_std)) × (1 - low_star_ratio)
- recovery_score = 0.6 × recovery_speed + 0.4 × recovery_breadth

🍽 Define performance metrics

Loyalty Score

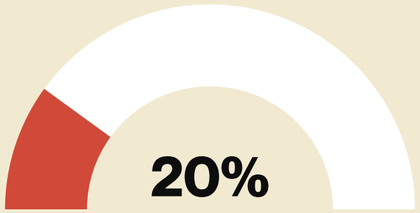
: reflects how beloved a restaurant is

Used Features



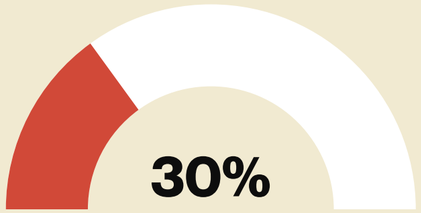
1 Regular Customer Score

- Represents the ratio of repeat customers and their ratings



2 Check-in Count

- Measures how many times the restaurant has been visited



3 Check-in Interval

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

Formula :

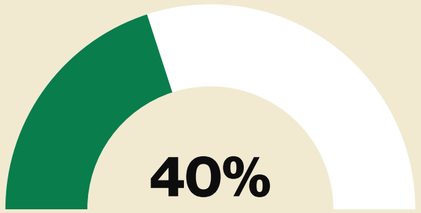
- regular_customer_score = 0.5 + regular_customer_ratio × 0.3 + (regular_customer_rating - 3) × 0.1

🍽 Define performance metrics

Reliability Score

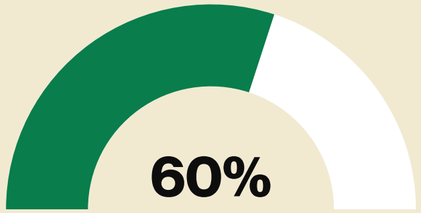
: captures how credible a restaurant appears

Used Features



1 Expert Review Ratio

- Reflects the ratio of expert reviews to total reviews



2 Elite Review Ratio

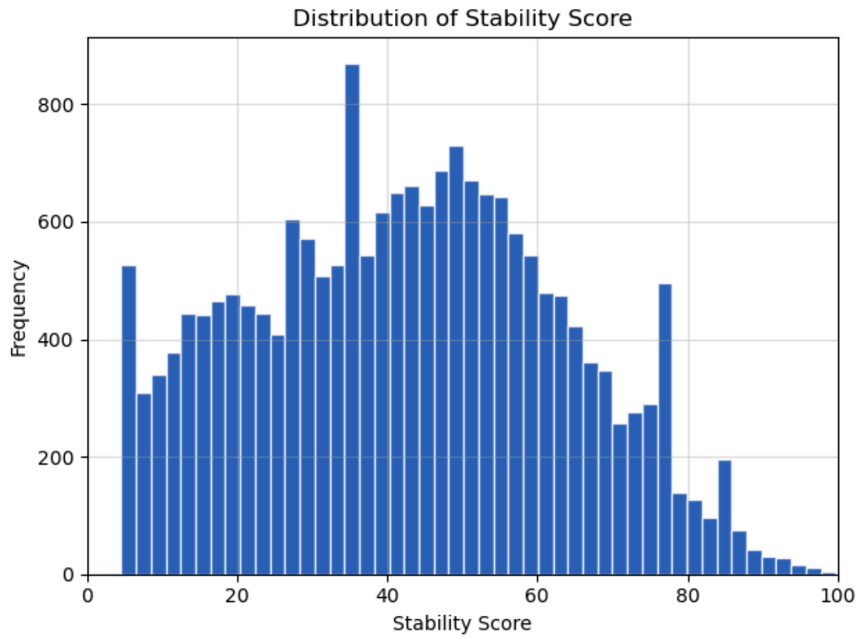
- Yelp selects “elite” users each year based on engagement
- Measures the ratio of elite reviews to total reviews.

Expert User

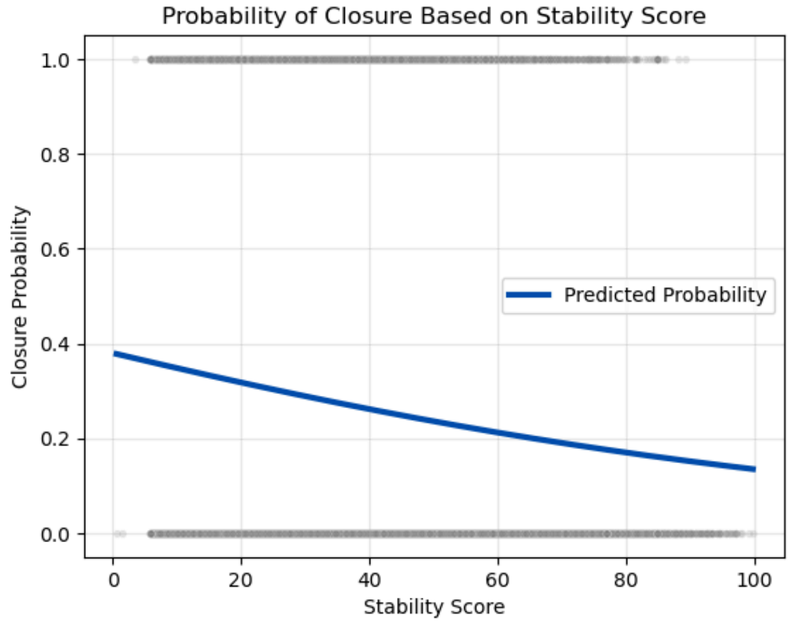
- “Experts” are users who focuses on reviewing specific food categories
- If the restaurant’s category matches one of a user’s top 3 categories, that review is considered an expert review

Stability Score

Distribution



Statistical Testing



- Higher stability scores are associated with lower closure probabilities

x

Stability Score

y

Business closure

Method

Logistic Regression

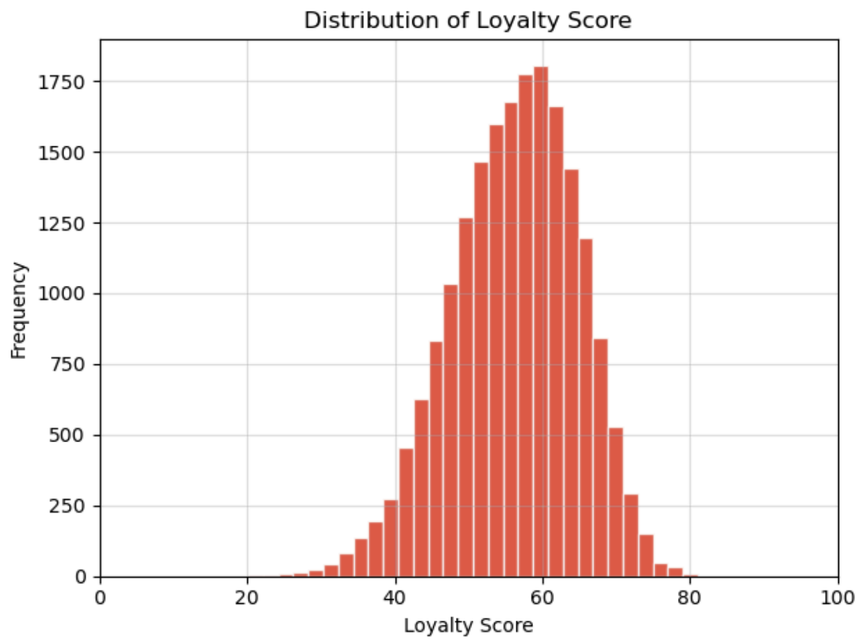
P-value

< 0.001

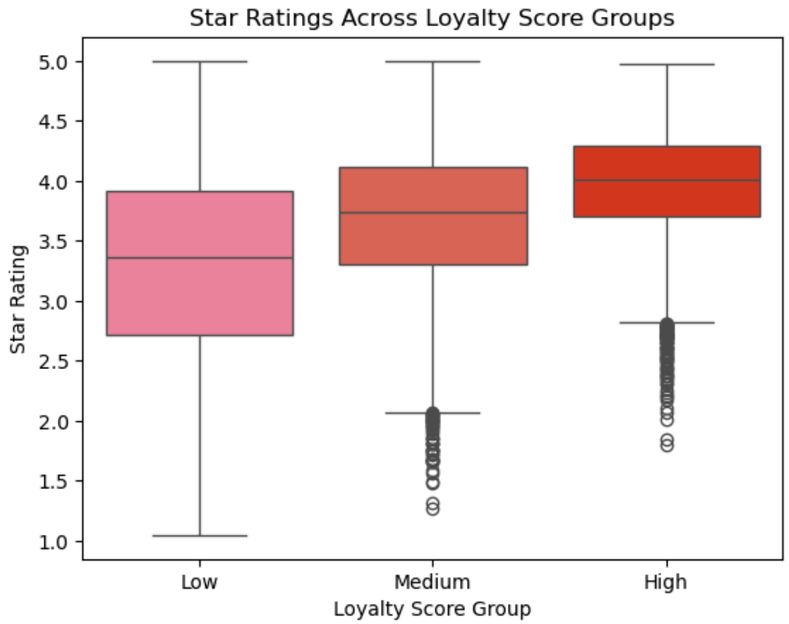
Statistical Testing

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

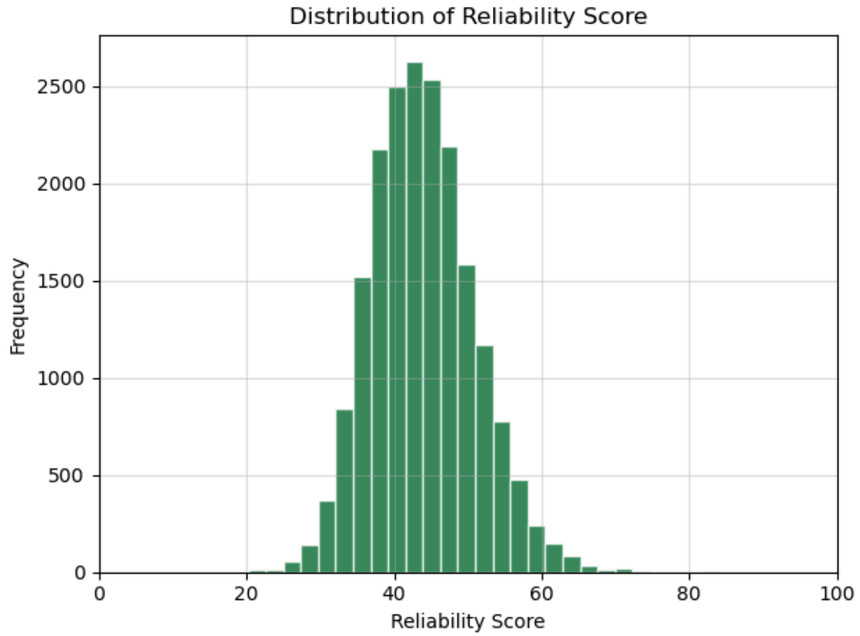
P-value

< 0.001

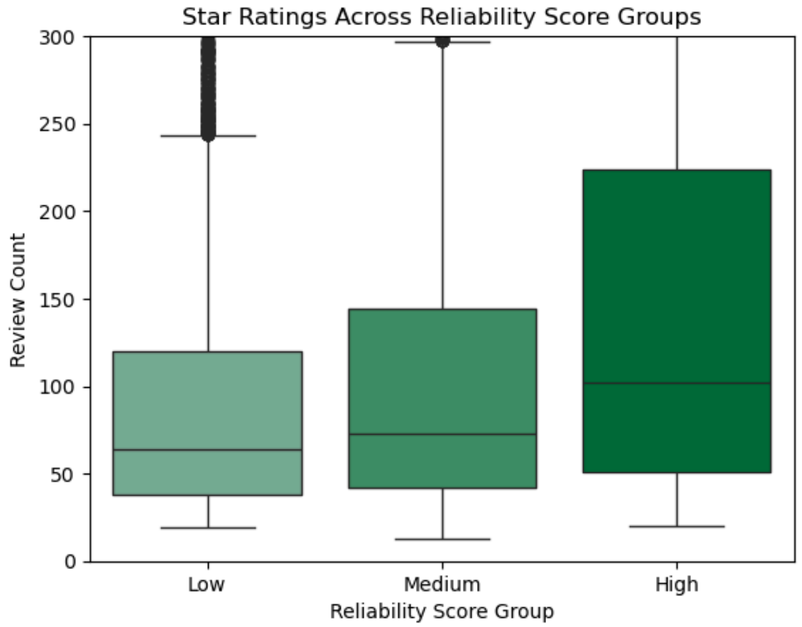
Statistical Testing

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

Machine Learning

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)

Machine Learning

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 (Oprimized)

Stability Score	
RMSE	11.09
MAE	8.67
R^2	0.70

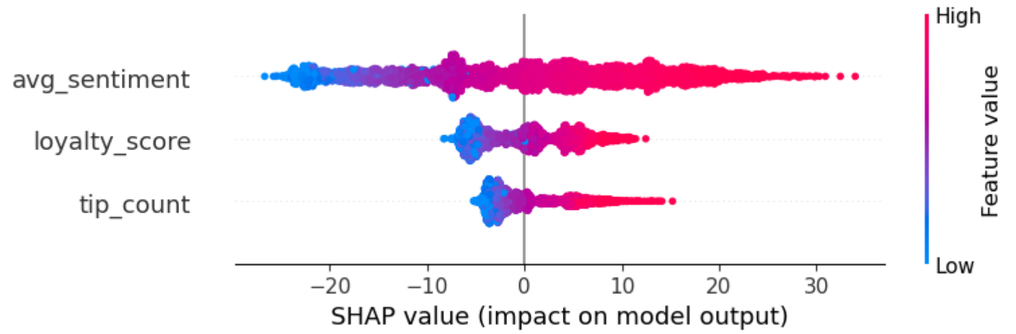
Loyalty Score	
RMSE	11.09
MAE	8.67
R^2	0.70

Reliability Score	
RMSE	11.09
MAE	8.67
R^2	0.70

Insights Deep Dive

Stability Score

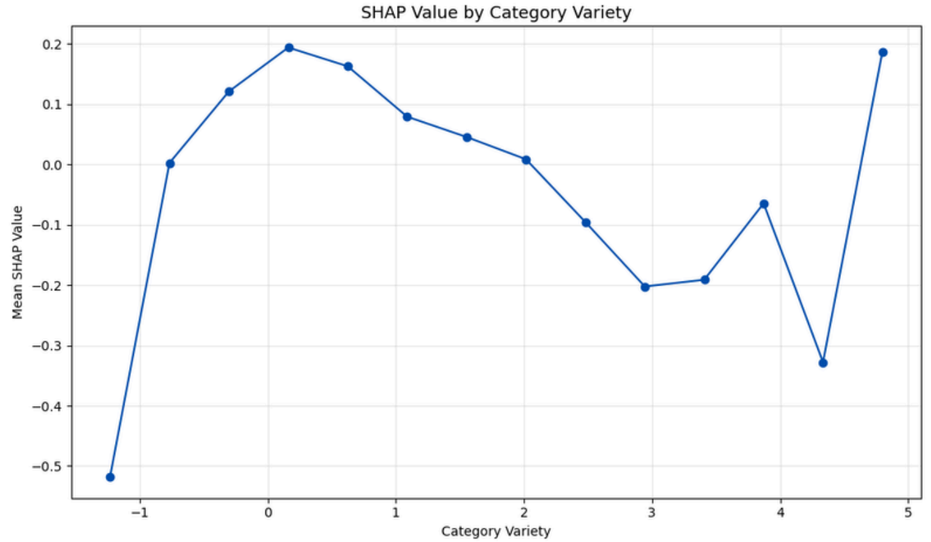
Insight ①



* *sentiment : if the reviews are positive or negative*

- Sentiment of reviews has the strongest impact

Insight ②

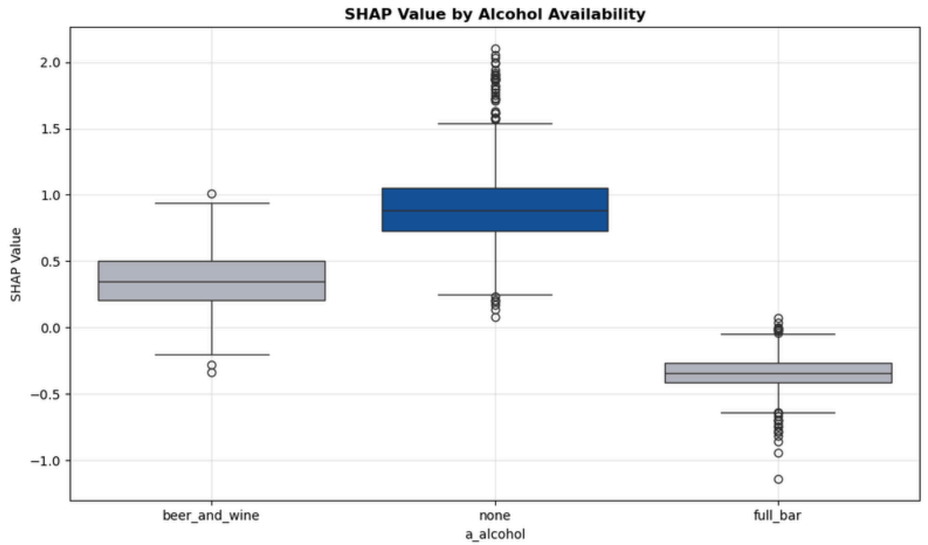


- Moderate category variety improves stability

Insights Deep Dive

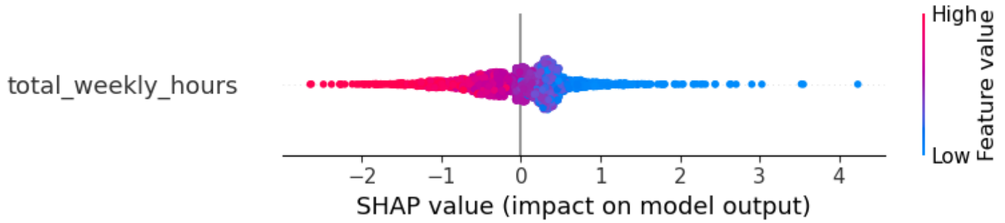
Stability Score

Insight ③



Restaurants without alcohol service tend to be more operationally stable

Insight ④

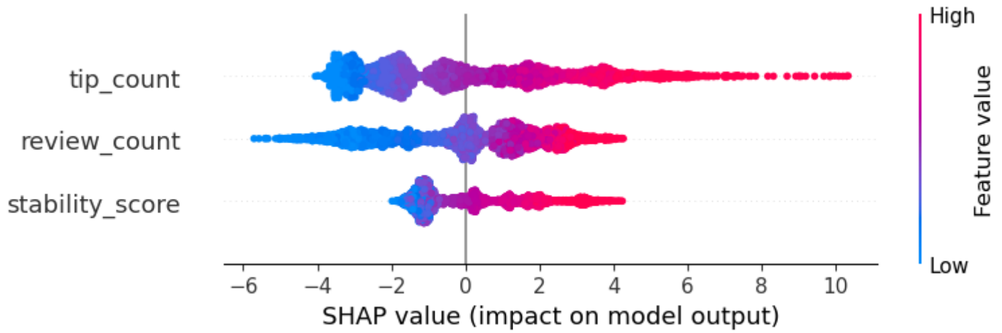


Excessively long operating hours negatively affect stability

Insights Deep Dive

Loyalty Score

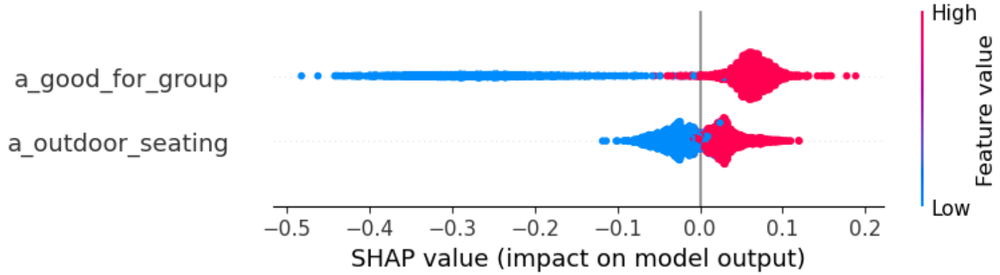
Insight ①



∗ *tip : short review for a restaurant*

Review volume is the strongest predictor of loyalty

Insight ②

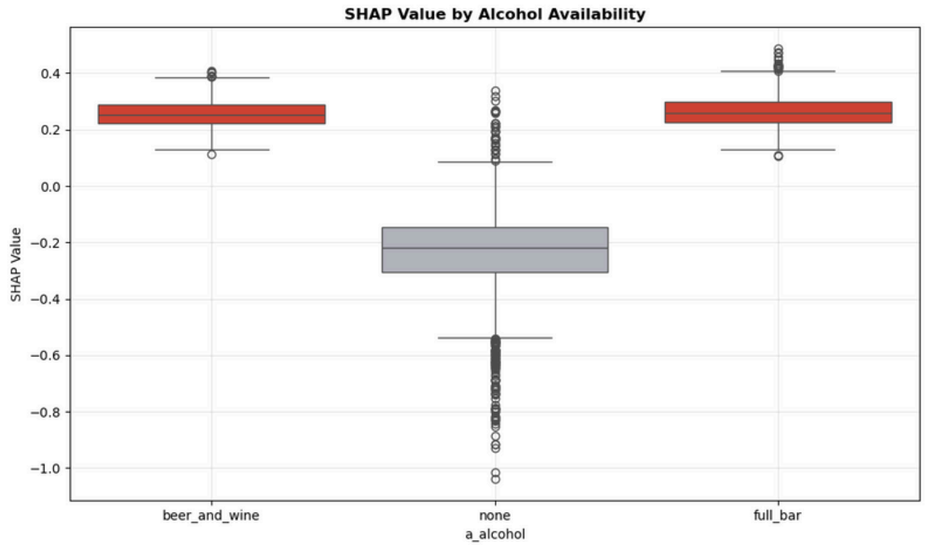


Outdoor seating and group-friendly environments increase repeat visits

Insights Deep Dive

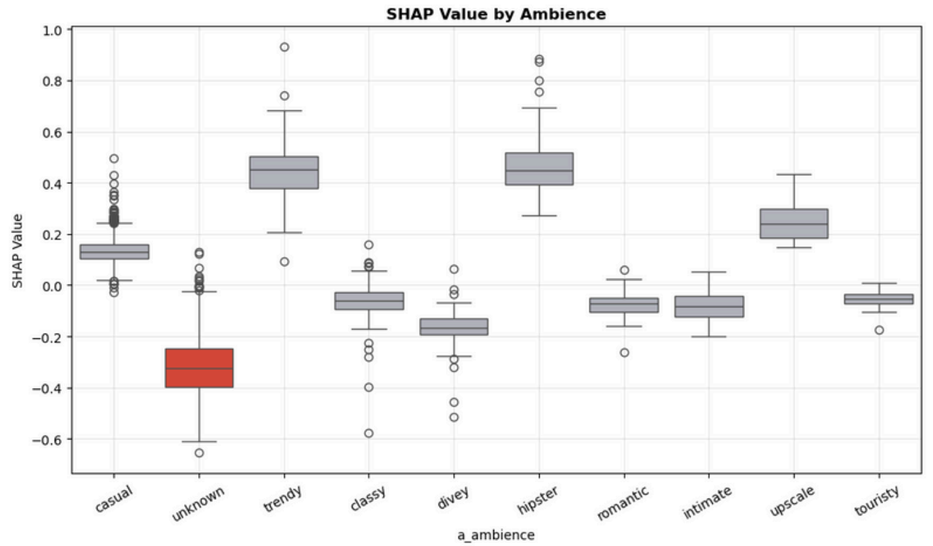
Loyalty Score

Insight ③



- Alcohol service increases loyalty, despite reducing stability

Insight ④

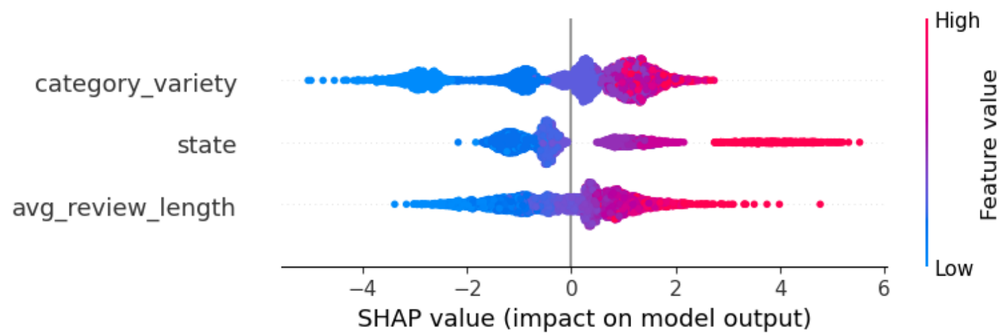


- Unclear ambience correlates with lower loyalty

Insights Deep Dive

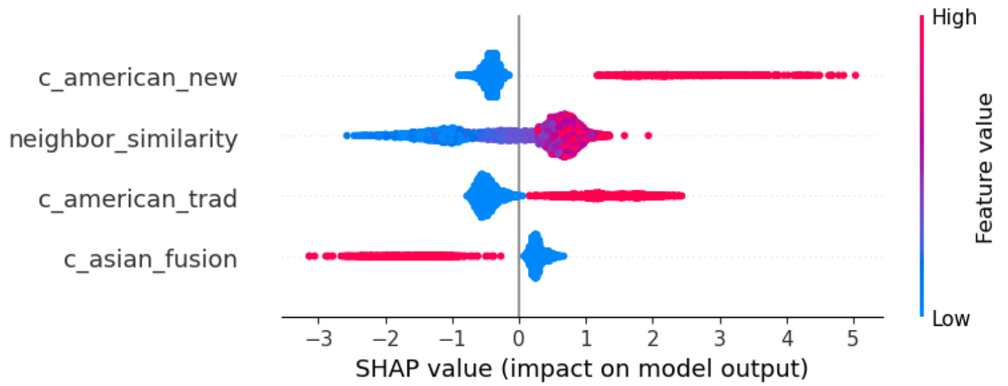
Reliability Score

Insight ①



- Breadth of food categories is the most influential factor

Insight ②

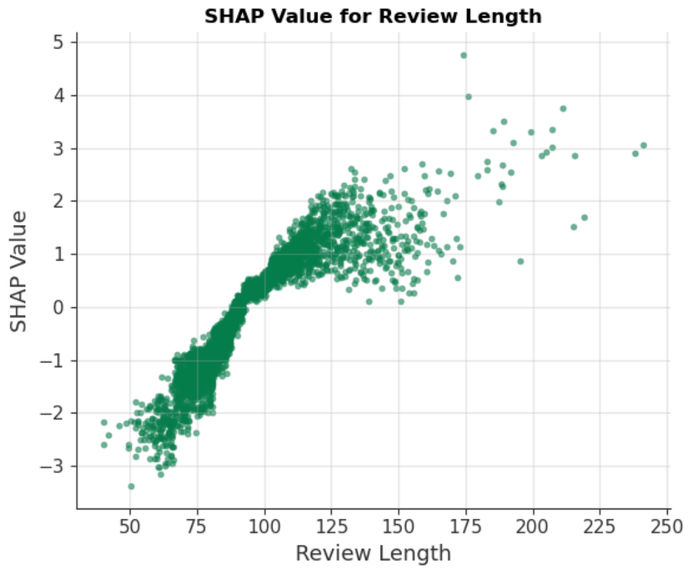


- Restaurants serving familiar, mainstream cuisines achieve higher reliability than niche or experimental concepts

Insights Deep Dive

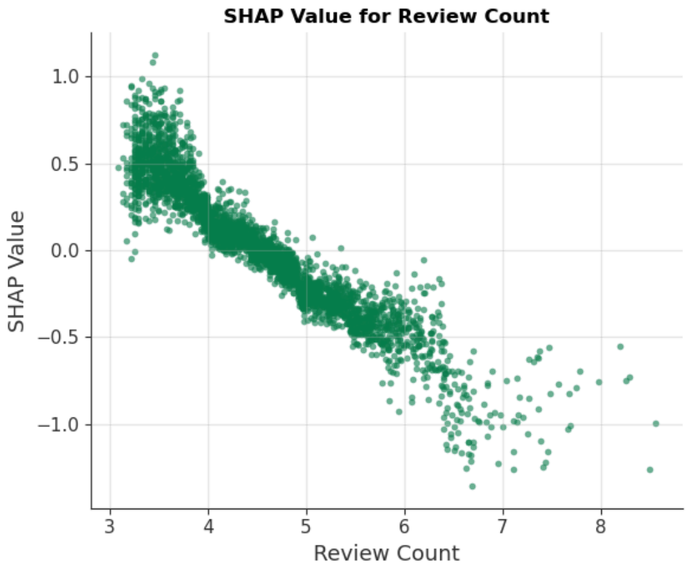
Reliability Score

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)