Unlocking Business Success Patterns in Yelp Data

Only 1 in 20 Yelp Food & Beverage businesses achieve elite success — what separates them?

Why Yelp Dataset?

150k+ Unique Businesses

6.7M+ Unique Reviews

1.4k+ Unique Cities

Perfect for
lever → impact playbook

Successful businesses build trust through reputation signals (reviews, stars); people build it through shared experiences.

Business success factors mirror what makes mentoring valuable: credibility, accessibility, relevance.

Why this matters?

Data Presented

- Yelp public dataset (multiple JSON files: business, user, review, check-in, tip).
- Chose the Business file as primary data source.

Business Goal

- Identify what variables influence high success businesses.
- Success defined as:
  Rating ≥ 4.5 stars
  At least 20 reviews

Questions we want to answer:

Performance Drivers:

- What variables seem most strongly correlated with the outcome of interest?

Operational Recommendations:

- If this dataset were our internal data, what actionable steps would you recommend?
- Where would you suggest collecting more data?

From raw JSON to calibrated, actionable levers

We turn reviews into revenue levers -- doubling success odds can drive double-digit lift in competitive categories.

The approach:

Step 1 – From Raw Data to Usable Table
- Extracted Business JSON from Yelp dataset.
- Converted to a structured dataframe with key variables.
- Cleaned & organized data to enable success analysis.

Step 2 – Coverage & Attribute Parsing
- Validated key fields & distributions (ratings, reviews, categories, hours).
- Fixed gaps with reliable ratings & standardized scores.
- Tested attributes for rating impact.

Step 3 – Modeling & Diagnostics
- Defined target (≥4.5 stars) and standardized features.
- Identified positive and negative drivers
- Built Regularized Logistic Regression, achieving strong performance.

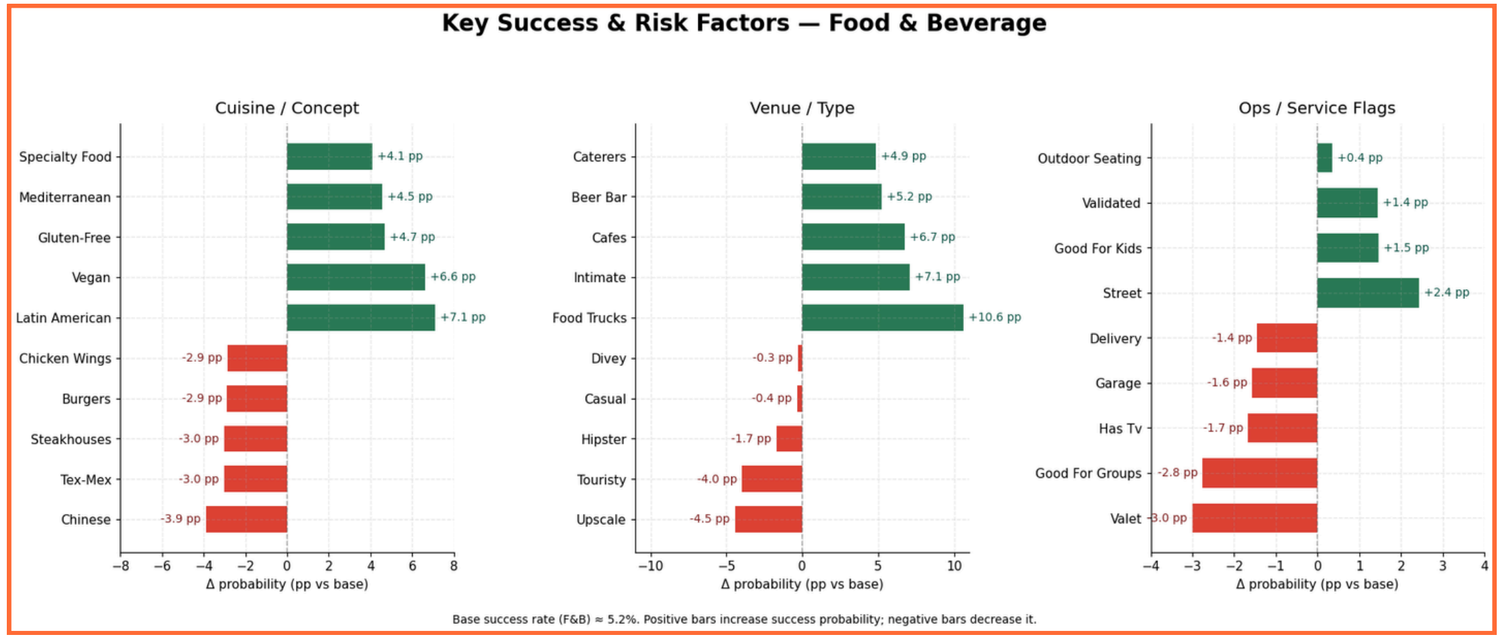
Step 4 – Roll-Up Charts/Tables
- Built driver lift charts to translate coefficients into probability shifts.
- Tested thresholds to balance precision vs recall.
- Aggregated features into roll-up themes (cuisine, venue, operations).

Step 5 – Validation
- Ran sanity checks with tree models
- Applied calibration
- Reviewed thresholds → focused on calibrated probabilities.

Step 6 – Recommendations
- Built an operator playbook mapping model results to actions & metrics.
- Logged data gaps for future fixes (ops, ambience, menu, geo).
- Created playbook charts to visualize top drivers by impact × reach.

Interpreting Success & Risk Factors

What variables seem most strongly correlated with the outcome of interest?



How to Read the Chart

- Green bars = factors that increase chances of high ratings.
- Red bars = factors that reduce chances.
- Numbers show percentage-point changes compared to the base success rate of 5.2% (≈5 in 100 businesses succeed).
- Example: A +3 pp lift means success goes from 5/100 → 8/100.

Summary of Findings

- Only 5.2% of Food & Beverage businesses (≈5 in 100) reach 4.5+ stars with 20+ reviews — this is our base success rate.
- We built a model to test what factors across Cuisine, Venue, and Operations drive businesses to stand out.
- The chart shows the top 5 success drivers and bottom 5 risk factors, measured as probability point (pp) changes.
- These point changes are added or subtracted from the 5.2% base rate to see how a business type would perform.

Examples

- Food Trucks: +10.6 pp → success rises from 5/100 → 16/100. Big lift, but niche.
- Good for Groups: –2.8 pp → success drops from 5/100 → 2–3/100. Negative impact, but common.

How We Prioritize

Combine Impact (lift/drop size) × Reach (how common it is).
High lift + small reach → niche pilot.
Moderate effect + wide reach → priority fix.

Success isn’t random—there’s a repeatable playbook.

If this dataset were your internal data, what actionable steps would I recommend?

Turning Data into a Playbook

We took the model results and translated them into an action-oriented playbook for operators.

Each row in the playbook shows:
Lever (cuisine, venue, or operation)
Whether it is a Success Driver or a Risk Factor

The priority (High / Medium / Low) based on impact × reach

Lever	(Cuisine / Venue / Ops)
Direction	(Success / Risk)
Priority	(High / Medium / Low)
Action	(short description)

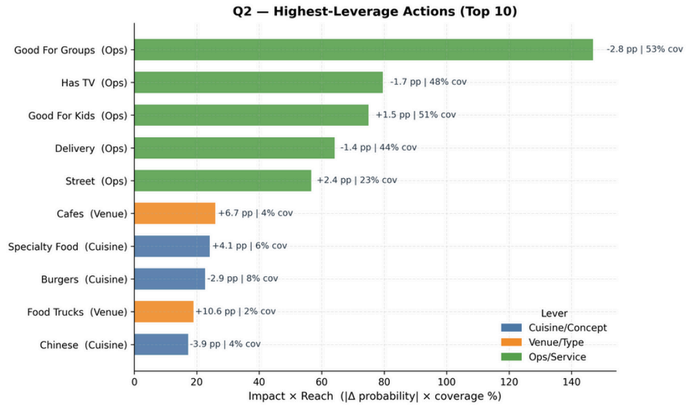
(q2_playbook_top_actions.csv) Priority Matrix Table

How a Business Uses It

If a restaurant wants to refresh its menu, the playbook suggests cuisines with higher odds of success (e.g., Mediterranean, Latin American).

If operations like delivery are dragging ratings, the playbook flags this as a high-risk area, with suggestions like partnering with third-party vendors or fixing packaging.

Each action comes with an experiment design (e.g., A/B testing) so changes can be validated before scaling.



If This Were Your Internal Data - My Recommendations

- Use the Playbook → a prioritized list of success drivers & risk factors (by impact × reach).
- Test Menu Changes → pilot projects before rolling out widely.
- Fix Operational Risks → address high-risk levers, consider redesign, partnerships, or removal.
- Run A/B Experiments → validate changes step by step before scaling.

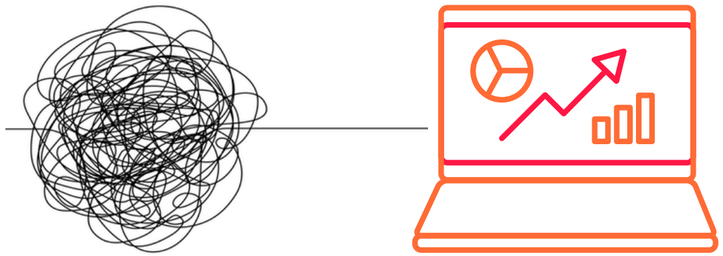
High Impact + High Reach	Priority Fixes
High Impact + Low Reach	Niche Pilots
Low Impact + High Reach	Watchlist
Low Impact + Low Reach	Low priority

Impact x Reach Quadrant

Big data isn’t useful, until you make it usable

Challenges I faced

- Messy data formats → solved with parsing & cleaning.
- Missing & overlapping data (e.g., alcohol vs cocktail bar vs wine) → consolidated to reduce skew.
- Bias in reviews (elite users, skewed ratings) → normalized.
- Precision–recall tradeoff → tuned thresholds, accepted balance.
- Actionability gap → bridged by designing the Playbook CSV.



Where would I suggest collecting more data?

Quick Wins

- Tag reviews by channel (dine-in vs delivery)
- Track promised vs actual ETA & packaging quality
- Add change log (attribute updates)
- Weekly roll-up of key metrics (% reviews > 4.5, review volume, clicks → calls)
- Improve photo coverage (menu, outdoor, recent updates)

Longer-Term Investments

- Capture ambience details (TV zones, noise, lighting, patio/table mix, wait times, reservations)
- Record parking counts by type & policy hours
- Track menu metadata (vegan/gluten-free share, seasonal items)
- Add geo-context controls (foot traffic, income, POI density)