Decontextualizing Ratings: Executive Summary

Julian Gould and Junyu Ma

April 22, 2025

1 Problem and Goal:

Restaurant ratings often reflect more than just food and service quality. We separate two classes of restaurant attributes:

1. Experience attributes: Food quality, service, ambiance, etc.
2. Context attributes: (trendy vs. un-trendy neighborhood), price point, accessibility (hours of operation and proximity to public transit), and social hype.

Our goal is to disentangle these influences to identify restaurants with excellent food, service, and ambiance, regardless of their context. Our product is an **experience score** that evaluates the quality of a dining experience of a restaurant, independent of the contextual attributes

The core challenge is that we cannot measure experience quality independent of context directly. "Experience" is inherently subjective, limiting the kinds of data available. While full-text reviews may provide some information, there is good reason to avoid measuring "experience" from them via natural language processing. Every dining experience happens in a context, potentially biasing any review's evaluation of the experience.

2 Data:

Our analysis used three primary data sources:

1. Google Review data:
   - Overview: We scraped the Google Review data of every restaurant in Philadelphia ($\approx$ 3700 restaurants) using Outscraper.
   - Purpose: This is our primary dataset of restaurant features. Beyond exact location and review distributions (1-5 stars), the data contains information about cuisine type, hours of operations, restaurant website information, social media presence, etc.

- Limitations: We could not get full-text reviews. Google Reviews has anti-scrubbing protections in place, and the full-text data was prohibitively expensive.
- Cleaning: Restaurants were removed from the data set on the following bases: insufficiently many reviews; falling outside the geographic boundaries of the greater Philadelphia area; having too many missing fields; being a food truck.

2. SEPTA Trolley data:

- Overview: The SEPTA trolley is the primary public transit network in Philadelphia. We scraped the locations of SCEPTA trolley stops around the city from the SEPTA API.
- Purpose: Measure how accessible a restaurant is by public transit.
- Limitations: Certain trolley lines were missing from the API.
- Cleaning: Missing lines were patched in by hand.

3. US Census data:

- Overview: The 2020 US Census was used to get demographic information around the city at the census-tract level (smaller than zip codes). Data was scraped through the US Census API
- Purpose: Get access to information about neighborhood character, such as median income.
- Limitations: Some missing data. Unclear how to handle restaurants on the boundary of multiple census tracts.
- Cleaning: Restaurants with missing census tract data was filled in by k-nearest neighbors (geographically).

3 Methodology

We compute experience scores by modeling how much of a restaurant's rating can be predicted from context alone, and treating the leftover, the residual, as a signal of true dining quality.

3.1 Meta Model:

We explain our methodology with the following simplified meta model. Suppose the average rating of a restaurant x is given by the following funcitonal form:

$$R(x) = \beta_0 + \beta_1 E(x) + \beta_2 C(x) + \epsilon$$

where $R(x)$ is the average rating, $E(x)$ is the "experience quality", and $C(x)$ is the "context quality." While $C(x)$ can be measured from the data, $E(x)$ cannot. Our methodology starts by fitting $R(x)$ on $C(x)$ alone, yielding predicted average rating:

$$\hat{R}(x) = \hat{\beta}_0 + \hat{\beta}_2 C(x)$$

The residual, $R(x) - \hat{R}(x) \approx \beta_1 E(x)$ is a scaled approximation of the experience quality $E(x)$. Since $\beta_1 > 0$, the larger the residual, the higher the experience score.

While we don't expect—and ultimately don't find—that a linear model fully captures the relationship between context and ratings, under reasonable assumptions, the residuals from such a model can still serve as ordinal proxies for dining experience: higher residuals suggest better experiences, even if the exact scale is not meaningful.

3.2 Model Assumptions

1. Review Homogeneity: Different cohorts of diners may eat at restaurants in different contexts. If those cohorts leave reviews in different ways, this residual approach would not be reliable. We assume that people review similarly once you control for context.

2. Geospatial Exogeneity: If an aspect of context is itself influenced by experience quality, it should not be used to predict ratings. Otherwise, the residual will "subtract out" part of the very experience we're trying to measure. To avoid this, we assume certain contextual features are exogenous. In particular, we assume that location is exogenous: restaurants with high or low experiential quality are equally likely to appear in any neighborhood.

3. Self-Selection of Diners: We assume that diners tend to visit restaurants that match their tastes, preferences, and expectations. This self-sorting behavior means that a restaurant's reviews are, to some extent, written by a relatively consistent and appropriate audience.

4. Constant Quality Through Time: We assume that restaurant quality is essentially constant, so even old reviews are still accurate reflections of the current quality.

3.3 Endogeneity of Price

As discussed in assumption 2, a key challenge in our model is that some contextual features may be endogenous. Higher prices may reflect better food and service, meaning they are partly determined by the very experience we're trying to isolate. If we naïvely use price to predict rating, we risk "soaking up" part of the experiential signal we hope to recover in the residual. To address this, we use Two-Stage Least Squares to handle endogeneity. In the first stage, we predict price using a set of exogenous instrumental variables that influence price but are assumed to be uncorrelated with experiential quality. In our case, we use features like cuisine type and neighborhood median income, which plausibly affect how much a restaurant charges, but not the direct quality of the dining experience. Specifically, we fit an ordered logit model of `price` (ordinal) on `cuisine_type` (categorical) and `census_tract_median_income` (numeric).

In the second stage, we substitute the predicted price from the first stage into our rating model. This helps us isolate the portion of price variation that is independent of experience. By doing so, we ensure that our experience score (the residual) is not contaminated by experience-driven variation in price itself.

This correction is crucial: without it, we risk underestimating the quality of restaurants that manage to deliver excellent experiences at lower prices, and overestimating those that rely on pricing as a signal of quality.

3.4 Final Model

After exploring a variety of different models, we found that a random forest model performed best. We fit the random forest model on the following regressors:

- `predicted_price` (numerical): the price as predicted from stage 1 of the 2SLS;
- `census_tract_median_income` (numerical): median income in the neighborhood the restaurant is in;
- `census_tract_population` (numerical): population of the neighborhood the restaurant is in;
- `trolly_distance` (numerical): distance to the nearest trolley stop;
- `city_hall_distance` (numerical): distance to city hall (center of Philadelphia downtown);
- `weekly_hours` (numerical): hours open per week;
- `review_count` (numerical): total number of reviews.

We also tried regressing on functions of these regressors (cross terms, squares, and square roots). However, we dropped these terms due to issues of over-fitting and non-explainability.

All seven regressors ended up having reasonably high importance scores. This likely means that we could be doing a better job of predicting average rating from contextual attributes. Future work on feature refinement would be valuable.

4 Output and Evaluation

The resulting residuals offer an interpretable, experience-isolated metric for comparing restaurant experience quality across contexts. Our analysis highlights both the power and the limits of contextual modeling: the residuals are meaningful precisely because the model doesn't explain everything.

We cannot directly evaluate our model, since we cannot measure experience quality directly. While we can assess how well the model predicts ratings, we cannot evaluate the residuals in the usual out-of-sample sense. Our confidence in them rests on the plausibility of our assumptions and the robustness of our modeling choices.

we can perform qualitative evaluation by inspecting the restaurants at the extremes of the residual distribution—those with the highest and lowest experience scores.

For high-scoring restaurants, we ask: Does this place feel underrated for its context? The highest scoring restaurant, *Hot Cluck*, looks to have delicious food, despite being located in a less popular neighborhood. Reading through Google reviews reveals people raving about the food.

Conversely, low-scoring restaurants may appear overrated once context is accounted for. They might have high prices, trendy locations, or lots of foot traffic, but reviews that hint at underwhelming food or service. The lowest scoring restaurant, *Geno's Steaks*, is a notorious tourist trap. Among locals, it is regarded as overpriced and overrated.

These restaurants, as well as others near the top and bottom of the list indicate we have found a valuable signal for hidden gem restaurants. The highest praise we can give this model is that as Philadelphia locals, we genuinely intend on trying the restaurants at the top of the list.