



FROM CONSUMER CONSENSUS TO INDIVIDUAL PREFERENCE: HOW RATING AGGREGATION SHAPES PREDICTIVE ACCURACY IN ONLINE CONSUMER REVIEWS

Author Information

Siyanbola Hakeem Opeyemi

Applied Statistics Department, Yunnan University, Kunming, China

Corresponding author

Email: hakeem.siyانبola@gmail.com

Abstract: Online consumer ratings are widely used to summarize product evaluations, but reported prediction errors depend on the unit of analysis. Using 1,048,340 BeerAdvocate reviews covering 42,716 beers and 28,761 reviewers, this study compares reconstruction of contemporaneously recorded overall ratings from sensory scores under two representations: beer-level averages and individual reviews. Linear Regression, Random Forest, and a validation-weighted ensemble were evaluated using beer-disjoint development and test partitions. The ensemble achieved MAE = 0.194, RMSE = 0.280, and $R^2 = 0.790$ for beer-level aggregates, compared with MAE = 0.305, RMSE = 0.405, and $R^2 = 0.682$ for individual reviews. Across matched model classes, raw-review RMSE was 38.8–45.6% higher. These errors are not like-for-like estimates for an identical prediction target: aggregation changes outcomes, sensory predictors, observation counts, and weighting (one beer versus one review per observation). Permutation importance identified taste, palate, and aroma as the leading predictors. A crossed random-intercept model attributed 28.9% of rating variance to beer identity, 7.0% to reviewer identity, and 64.1% to residual review-level variability. The findings show that product-level consensus is easier to reconstruct in this setting, not that personalized recommendations are more accurate. Retailers and review platforms should evaluate consensus estimation and individual-response reconstruction separately and display rating dispersion alongside product averages.

IndexTerms - Online consumer reviews; consumer preference; rating aggregation; consumer heterogeneity; recommender systems; retail analytics.

1. INTRODUCTION

Online consumer ratings and reviews have become important information resources for consumers, retailers, manufacturers, and digital platforms. Prior retail research demonstrates that online product ratings and multi-attribute reviews provide information relevant to consumer evaluation, satisfaction, and marketplace outcomes [29–31]. Consequently, the manner in which consumer ratings are represented and analyzed has implications not only for predictive modeling but also for the interpretation of consumer feedback and the design of recommendation systems.

A fundamental distinction arises between collective consumer evaluations and individual consumer preferences. An aggregate product rating summarizes evaluations from multiple consumers and can therefore serve as an indicator of market-level consensus. An individual rating, by contrast, reflects the evaluation of a particular consumer and may incorporate personal preferences, expectations, prior experience, and other sources of heterogeneity. Research on recommendation uncertainty and consumer sensory evaluation recognizes that observed evaluations can contain meaningful variation rather than simply measurement error [10,11,15,27]. Although aggregated and individual ratings originate from the same review environment, they therefore represent different analytical constructs.

This distinction has important implications for predictive analytics. Averaging consumer ratings at the product level reduces dispersion by partially smoothing differences among individual evaluations. Predictive models trained and evaluated using aggregated observations may consequently exhibit lower prediction errors than models operating at the individual-review level. Such differences become particularly important when product-level predictive performance is interpreted as evidence that a system can predict individual consumer preferences with comparable precision. Research on uncertainty and trustworthy

recommender systems further emphasizes that predictive accuracy should be interpreted in relation to the uncertainty, heterogeneity, and feedback characteristics present in the underlying recommendation environment [9–16].

Existing recommender-system research has devoted considerable attention to collaborative filtering, graph-based recommendation, multimodal systems, robustness, explainability, uncertainty, and increasingly sophisticated user-item representations [1–14,19–22]. Review-aware systems also exploit textual and rating information to represent users and products more effectively [7,8]. However, comparatively less attention has been directed toward how the analytical representation of consumer ratings themselves affects the interpretation of predictive accuracy. In particular, predictive performance obtained from product-level averages and performance obtained from individual consumer ratings should not automatically be treated as equivalent indicators of a model's ability to represent consumer preferences.

This issue is particularly relevant in sensory evaluation. Consumer assessments of products such as food and beverages involve multiple perceptual attributes, while individual preferences can differ even when consumers evaluate the same underlying product [27]. Beer provides a useful empirical setting because overall evaluations are associated with sensory dimensions including taste, aroma, palate, and appearance, and prior research has demonstrated that machine-learning methods can recover meaningful relationships between beer characteristics and consumer appreciation [23]. At the same time, the availability of repeated ratings across beers and reviewers makes it possible to distinguish stable product-level variation from reviewer-level and residual review-level variability.

This study addresses the issue by treating rating representation as an explicit component of predictive evaluation. Using more than one million BeerAdvocate reviews, the study compares two analytical representations of the same underlying consumer-review information. The first aggregates overall ratings and sensory attributes to the beer level, producing a smoothed representation of collective consumer evaluation. The second retains ratings and attributes at the individual-review level, thereby preserving consumer disagreement and other review-level variability. Importantly, this comparison does not change only the dependent variable: aggregation changes the unit of analysis and the distributions of both the outcome and associated sensory predictors. The empirical question is therefore how predictive performance changes when consumer-review information is represented at the collective product level rather than at the individual-review level.

To provide a controlled comparison, matched Linear Regression, Random Forest, and validation-weighted ensemble models are evaluated using the same sensory attribute set under a beer-disjoint development-test design. The analysis is complemented by bootstrap uncertainty assessment, permutation feature importance, and a crossed random-intercept model that decomposes observed rating variability into beer-level, reviewer-level, and residual review-level components. The use of uncertainty assessment and interpretability analysis is consistent with broader efforts to move recommender-system evaluation beyond predictive accuracy alone [9–11,13,19–22].

Accordingly, the study addresses four research questions:

RQ1: How does predictive performance differ between beer-level aggregated consumer evaluations and raw individual consumer ratings?

RQ2: Is the performance difference consistent across matched linear, tree-based, and ensemble models?

RQ3: Which sensory attributes are most influential under aggregated and individual-review representations, and how stable are these relationships across the two conditions?

RQ4: How much of the observed variation in consumer ratings is attributable to beer identity, reviewer identity, and residual review-level variability?

This study makes three contributions. Conceptually, it distinguishes consensus predictability from individual-preference predictability, highlighting why the two should not be treated as interchangeable measures of consumer understanding. Methodologically, it demonstrates how changing the analytical level through aggregation can materially alter reported predictive performance even when the underlying review source, predictor constructs, model classes, and held-out products remain comparable. Managerially, it shows why product-level prediction and individualized recommendation should be evaluated separately: strong prediction of collective product evaluations does not necessarily imply equally strong prediction of individual consumer responses. This distinction is relevant to the broader development of reliable, interpretable, and consumer-oriented recommendation systems [9,13,19,21,25].

2. LITERATURE REVIEW AND CONCEPTUAL BACKGROUND

2.1 Online Consumer Reviews and Consumer Decision-Making

Online consumer ratings and reviews provide explicit signals about consumers' experiences with products and services and constitute important information for recommender systems and consumer analytics [1,7,8,12]. Unlike purely transactional data, review environments can contain both quantitative ratings and qualitative information reflecting users' perceptions of product characteristics. Review-aware recommender systems consequently use such information to improve representations of consumer preferences and product attributes [7,8].

From a consumer-analytics perspective, ratings should not necessarily be interpreted as direct measurements of objective product quality. Consumer evaluations arise from interactions between product characteristics and individual perceptions, expectations, experiences, and preferences. Different consumers may therefore provide different evaluations of the same product even when its observable characteristics remain unchanged. Research concerned with uncertainty in recommender systems similarly recognizes that prediction must accommodate variation in observed consumer feedback rather than assuming that every rating represents an error-free expression of a fixed underlying preference [10–12].

This distinction is particularly relevant to product categories characterized by subjective sensory evaluation. Food and beverage preferences depend on multiple sensory dimensions and consumer-specific responses, making them useful domains for investigating relationships between observable product evaluations and subjective preference [27]. Beer ratings provide such a setting because consumers commonly evaluate several sensory dimensions alongside an overall assessment, and previous research has demonstrated the potential of machine learning to identify relationships between beer characteristics and consumer appreciation [23].

2.2 Consumer Preference Heterogeneity and Rating Disagreement

Consumers are heterogeneous in their preferences, experiences, evaluation criteria, and responses to product characteristics. Variation in ratings should therefore not automatically be treated as erroneous labels. Instead, observed disagreement may combine meaningful consumer heterogeneity with contextual influences, consumer-product interactions, measurement variation, and other unexplained factors. Research on uncertainty and noisy labels highlights the importance of distinguishing genuine variation from measurement or labeling error when interpreting predictive performance [10,11,15,16].

This issue has direct implications for recommender systems. Personalized recommendation seeks to identify items that are relevant to particular users rather than merely products that receive high average evaluations [1,12]. Consequently, variation between consumers is not necessarily an undesirable property of the data; it may constitute part of the phenomenon that a personalized system is intended to model. Robust and trustworthy recommender-system research similarly emphasizes the need to account for noisy, sparse, biased, uncertain, or distributionally heterogeneous feedback when assessing system performance [9,13].

At the individual level, a consumer rating can therefore be viewed as reflecting both product-related structure and additional reviewer- or context-specific variation. Predictive models based primarily on observable product attributes may successfully capture stable differences among products while leaving substantial individual-level variability unexplained. Quantifying this decomposition is useful for determining whether apparent prediction difficulty primarily reflects differences among products, stable reviewer tendencies, or variation remaining at the individual review level.

2.3 Aggregated Ratings and Individual Consumer Evaluations

Product-level aggregation transforms individual evaluations into a measure of collective tendency but also reduces the disagreement retained in the underlying rating distribution. Consumer-review research shows that rating dispersion provides information beyond the average rating: low dispersion indicates greater agreement, whereas high dispersion reflects heterogeneous consumer evaluations and can influence how consumers interpret product ratings [31,32]. Thus, products with similar mean ratings may represent substantially different patterns of individual consumer response.

This distinction has direct implications for predictive modeling. When individual ratings are averaged, positive and negative deviations partially offset one another, producing a smoother product-level representation that may be easier to predict. Higher accuracy under aggregation should therefore not automatically be interpreted as stronger prediction of individual consumer preferences. Aggregated ratings primarily represent collective consensus, whereas raw ratings preserve consumer disagreement and review-level variability relevant to personalization [10,11,31,32].

2.4 Predictive Analytics and Recommender Systems

Recommender systems employ a wide range of methods to infer preferences and identify items likely to be relevant to individual users. These approaches include collaborative filtering and conventional recommendation methods as well as graph neural networks, review-aware models, multimodal architectures, and other advanced representations [1–8]. Recent research has also increasingly addressed robustness, fairness, uncertainty, and explainability as important dimensions of recommendation-system quality [9–14,19–22].

Recommendation quality depends on more than accuracy: noisy feedback, explanation, and representation also matter [9,15,16,19–22]. Product-level prediction estimates consensus; personalization additionally requires evidence about variation between consumers.

Nevertheless, sophisticated modeling does not remove the need to define the prediction task appropriately. Product-level prediction and consumer-level personalization answer different questions. A model that accurately estimates the expected evaluation of a product can support product comparison or consensus estimation, while a personalized recommendation system must additionally accommodate variation among individual users. Consequently, evaluation should distinguish performance obtained from collective product summaries from performance obtained when individual consumer responses are retained.

2.5 Sensory Ratings, Beer Evaluation, and Predictive Interpretation

Sensory and consumer research provides an additional foundation for examining rating representation. Consumer responses to food and beverage products arise from multiple sensory characteristics, and contemporary sensory research increasingly uses computational methods to analyze complex relationships between product attributes and consumer appreciation [27].

Beer provides a particularly suitable empirical context because consumer reviews frequently contain separate evaluations of taste, aroma, palate, and appearance in addition to overall ratings. Schreurs et al. [23], for example, demonstrated that machine-learning methods can recover meaningful relationships between beer characteristics and consumer appreciation. These findings support the use of beer as a domain for investigating predictive relationships among sensory evaluations.

However, strong relationships between sensory dimensions and overall ratings do not eliminate consumer heterogeneity. Two consumers can evaluate similar sensory characteristics differently when forming an overall judgment. Consequently, predictive performance based on average product evaluations may differ substantially from performance at the individual-review level. Examining whether the relative importance of sensory attributes remains stable across these representations can also help distinguish changes in predictive difficulty from changes in the underlying attribute relationships.

2.6 Research Gap and Contribution

Existing research addresses recommender-system architectures, uncertainty, explainability, and consumer-rating heterogeneity [1–16,19–27], while recent retail research also shows that inconsistency between individual and product-average ratings can affect consumer decision-making [28–33]. However, limited evidence examines how shifting from individual review-level evaluations to product-level aggregation changes measured predictive performance. This study addresses that gap by comparing both representations under a common beer-disjoint evaluation framework. Rather than proposing a new algorithm, it demonstrates how aggregation can alter apparent predictability and, consequently, the interpretation of model performance for product-level consensus versus individual-level personalization.

3. METHODOLOGY

3.1 Dataset Description and analytical sample

The study used the BeerAdvocate review corpus distributed for recommender-system research. The source file contained 1,048,575 observations. After requiring a valid overall rating, beer identifier, and reviewer identifier, 1,048,340 reviews remained for analysis. These observations represented 42,716 unique beers and 28,761 unique reviewers. Table 1 summarizes the principal variables used in this study.

Review frequency was highly right-skewed across beers. The mean was 24.54 reviews per beer, whereas the median was 2 (IQR: 1–8; range: 1–2,999); 35.68% of beers had only one review and 51.63% had two or fewer reviews. Because beer-level means based on sparse feedback may provide limited estimates of collective evaluation, a sensitivity analysis was conducted using beers with at least five eligible reviews.

Table 1. Description of variables used in the analysis.

Variable	Role / description
review_overall	Overall beer rating (prediction target)
review_taste	User-assigned taste score
review_aroma	User-assigned aroma score
review_palate	User-assigned palate/mouthfeel score
review_appearance	User-assigned appearance score
beer_abv	Alcohol by volume (%)
beer_beerid	Beer identifier used for grouping and leakage control
review_profilename	Reviewer identifier used for variance-component analysis

3.2 Data Preprocessing and missing-value handling

The numerical variables review_overall, review_taste, review_aroma, review_palate, review_appearance, and beer_abv were converted to numeric form. Records lacking the target, beer identifier, or reviewer identifier were excluded. Predictor missingness was handled by median imputation using medians estimated from the internal fit partition only; the same training-derived values were then applied to validation, development, and test data. Beer and reviewer identifiers were not used as ordinary numeric predictors. Because the final predictive models were ordinary least squares and tree-based Random Forest, z-score scaling was not required in the final workflow.

3.3 Dual Representation Framework

Two analytical representations of the consumer-review data were evaluated. The raw representation retained each individual review_overall value together with the corresponding review-level sensory predictors. The aggregated representation was constructed by averaging overall ratings within each beer and applying the same beer-level aggregation to the five numerical predictors. Consequently, the aggregated analysis uses beers as the observational units, whereas the raw analysis uses individual reviews as the observational units.

The aggregated representation is therefore interpreted as a smoothed product-level consensus benchmark, while the raw representation preserves individual consumer disagreement and other review-level variability. Importantly, the comparison does not isolate aggregation of the dependent variable alone. Aggregation changes both the outcome and predictor distributions as well as the unit of analysis. The resulting performance difference should therefore be interpreted as the effect of changing the analytical representation of the rating task from individual reviews to product-level summaries, rather than as a pure causal effect of target averaging.

$$\bar{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}$$

where $\bar{R}_i$ represents the aggregated rating of beer i , R_{ij} denotes the rating assigned in review j for beer i , and n_i is the number of ratings received by beer i .

3.4 Leakage-controlled data partitioning

Unique beer identifiers were randomly divided with seed 123 into 80% development beers and 20% held-out test beers. All reviews associated with a given beer were assigned to only one partition. The resulting data contained 850,058 raw development reviews and 198,282 raw test reviews; after beer-level aggregation, there were 34,172 development beers and 8,544 test beers. The held-out test set was not used for fitting or ensemble-weight selection.

Within the development partition, a second beer-disjoint fit-validation split was created using seed 456. This validation set was used to select the weights of the linear-regression/Random-Forest ensemble. Thus, model combination was tuned independently of final test performance.

3.5 Predictive Models

Three matched configurations were evaluated under both analytical representations: Linear Regression, Random Forest, and a weighted Linear Regression–Random Forest ensemble. Linear Regression provides an interpretable additive benchmark, while Random Forest accommodates nonlinear relationships and interactions among structured predictors. Ensemble approaches can combine complementary predictive behavior across component models [17,18]. The predictor set was review_taste, review_aroma, review_palate, review_appearance, and beer_abv.

3.5.1 Linear Regression

Ordinary least squares regression served as the interpretable baseline, modeling the overall rating as an additive function of the five predictors.

3.5.2 Random Forest

Random Forest models were estimated with the ranger implementation. Final models used 500 trees and $mtry = 3$. The minimum node size was 10 for the raw condition and 5 for the aggregated condition. Permutation importance was requested during model fitting. Parallel processing was used where available, and seed 123 was fixed for reproducibility.

3.5.3 Validation-weighted ensemble

The ensemble combined Linear Regression and Random Forest predictions. Candidate Random-Forest weights from 0 to 1 in increments of 0.01 were evaluated on the internal validation set, with the complementary weight assigned to Linear Regression. The weight minimizing validation RMSE was selected separately for each target representation. The final raw ensemble used weights 0.55 (Random Forest) and 0.45 (Linear Regression); the aggregated ensemble used 0.26 and 0.74, respectively.

3.6 Evaluation Metrics

Performance was assessed on the held-out test set using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). Lower MAE and RMSE indicate better predictive accuracy, whereas higher R^2 indicates greater explained variance.

Mean Absolute Error:

$$MAE = \frac{1}{n} \sum_{i=1}^n \hat{y}_i - y_i$$

Root Mean Squared Error:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

Coefficient of Determination:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

These metrics provide complementary perspectives on predictive accuracy, error magnitude, and explanatory power.

3.7 Bootstrap uncertainty analysis

To quantify uncertainty in held-out Random Forest performance, we conducted 1,000 paired nonparametric bootstrap resamples at the beer level. The common set of 8,544 held-out beer IDs was sampled with replacement in each replicate. For the aggregated representation, each selected beer contributed its beer-level observed rating and fixed Random Forest prediction. For the raw representation, each selected beer contributed all its associated held-out review observations and corresponding fixed predictions; a beer selected multiple times contributed its reviews multiple times. We calculated MAE and RMSE separately for each representation, retaining review-level weighting for the raw condition and beer-level weighting for the aggregated condition. For each paired draw, we also calculated the difference between raw and aggregated RMSE. Two-sided 95% confidence intervals were obtained from the 2.5th and 97.5th percentiles of the bootstrap distributions. The procedure resampled test-set predictions without refitting models or repeating model selection; thus, intervals quantify held-out sampling variability conditional on the fitted models rather than total model-training uncertainty.

3.8 Permutation feature importance

Random-Forest feature importance was estimated using permutation importance. This approach evaluates the contribution of a predictor by measuring the deterioration in predictive performance when its values are disrupted while the fitted model is otherwise retained. In the present study, importance values are used as global model-interpretation measures and are interpreted comparatively rather than as causal effects. This distinction is consistent with the broader explainable-AI literature, which emphasizes that post-hoc feature attribution describes model behavior and should not automatically be interpreted as causal explanation [19,21].

3.9 Variance-component analysis

To quantify the sources of subjective variability, a crossed random-intercept model was fitted to a reproducible random sample of 100,000 reviews:

$$R_{ij} = \mu + b_i + u_j + \varepsilon_{ij}$$

where b_i is the beer-specific random intercept, u_j is the reviewer-specific random intercept, and ε_{ij} is residual review-level variation. The model was estimated by restricted maximum likelihood using lme4. Variance shares were computed by dividing each component by total variance. These shares quantify the relative contribution of beer identity, reviewer identity, and residual review-level variation to observed ratings.

3.10 Review-frequency sensitivity analysis

To examine whether sparse product-level feedback influenced the aggregation results, the predictive analysis was repeated among beers with at least five eligible reviews. This restriction retained 998,669 reviews across 14,530 beers. Common beer-disjoint training, validation, and test partitions were applied to the raw and beer-level aggregated representations, with no beer identifiers shared across partitions. Linear Regression, Random Forest, and validation-weighted ensembles were evaluated using RMSE, MAE, and R^2 , with ensemble weights selected on the validation partition before final evaluation on the held-out test set.

As an additional robustness specification, the restricted-sample models used four sensory predictors (taste, aroma, palate, and appearance), excluding alcohol by volume. The analysis therefore assesses whether the difference between individual-review and aggregated prediction persists under a common sensory-only specification and a minimum-review threshold; it is not an exact replication of the five-predictor primary models.

4. RESULTS

4.1 Dataset characteristics and target distributions

Table 2. Analytical sample and leakage-controlled partition sizes.

Measure	Value
Source observations	1,048,575
Analytical raw reviews	1,048,340
Unique beers	42,716
Unique reviewers	28,761
Raw development reviews	850,058
Raw test reviews	198,282
Aggregated development beers	34,172
Aggregated test beers	8,544
Reviews retained in ≥ 5 -review sensitivity analysis	998,669
Beers retained in ≥ 5 -review sensitivity analysis	14,530

The held-out aggregated target had mean 3.667 (SD = 0.611; median = 3.75; IQR = 0.580), whereas the raw target had mean 3.832 (SD = 0.719; median = 4.00; IQR = 1.00). The larger dispersion of raw ratings is visible in Figure 1 and is consistent with the greater heterogeneity retained at the review level.

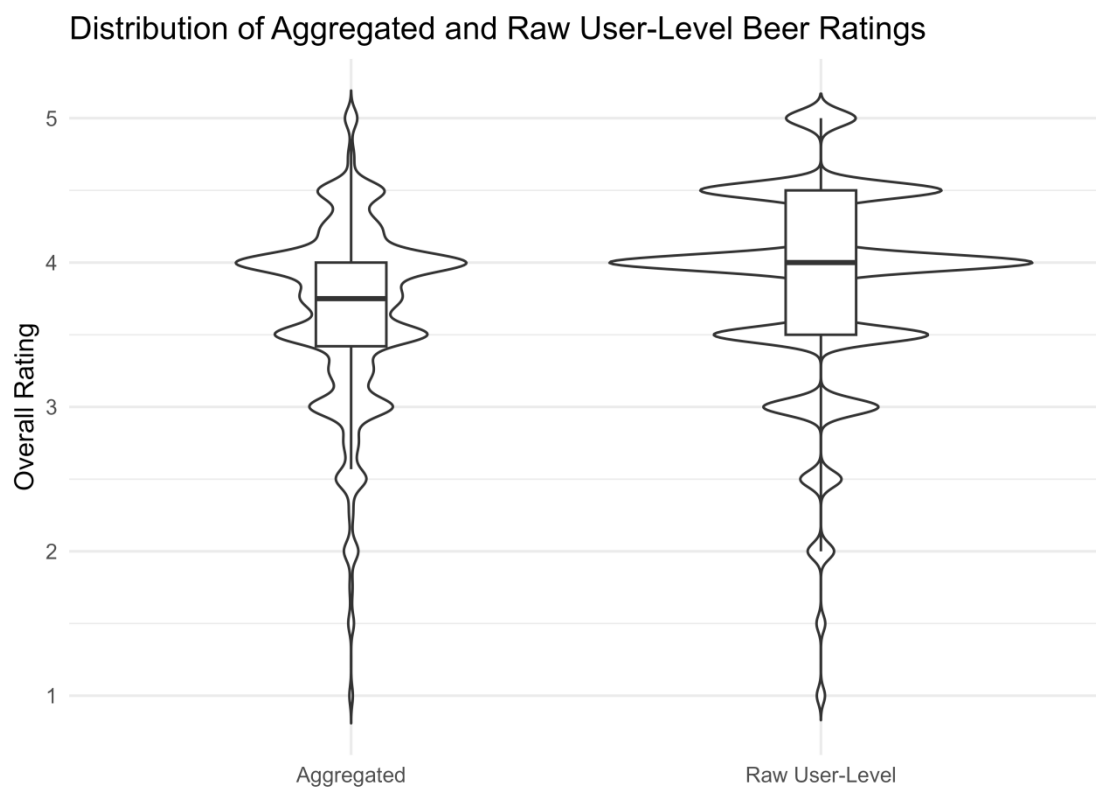


Fig. 1. Distribution of overall ratings under aggregated beer-level and raw review-level representations.

The violin plots and kernel-density estimates illustrate the lower dispersion associated with beer-level aggregation. The aggregated target exhibits substantially lower dispersion than the raw target, demonstrating the variance-reduction effect introduced by rating aggregation.

4.2 Matched-model predictive performance**Table 3. Final held-out performance of matched models under aggregated and raw target representations.**

Condition	Model	MAE	RMSE	R ²
Aggregated	Linear Regression	0.195	0.281	0.789
Aggregated	Random Forest	0.204	0.294	0.769
Aggregated	Weighted LM-RF Ensemble	0.194	0.280	0.790
Raw	Linear Regression	0.307	0.409	0.677
Raw	Random Forest	0.308	0.408	0.678
Raw	Weighted LM-RF Ensemble	0.305	0.405	0.682

The weighted ensemble achieved the best overall performance in both target conditions. Under aggregated ratings, it produced MAE = 0.194, RMSE = 0.280, and R² = 0.790. Under raw ratings, the corresponding values were MAE = 0.305, RMSE = 0.405, and R² = 0.682. The similar ordering across target conditions indicates that the main change in apparent accuracy was associated with target representation rather than a reversal in model ranking.

Linear Regression was the strongest single model under the aggregated target (RMSE = 0.281, R² = 0.789), whereas Random Forest was marginally stronger than Linear Regression on raw RMSE and R² (0.408 and 0.678 versus 0.409 and 0.677). The differences between these individual models were small relative to the gap between aggregated and raw prediction.

4.3 Effect of analytical representation**Table 4. Relative change in predictive performance from aggregated to raw ratings.**

Model	MAE change (%)	RMSE change (%)	R ² change (%)
Linear Regression	57.3	45.6	-14.3
Random Forest	51.1	38.8	-11.8
Weighted LM-RF Ensemble	57.0	44.7	-13.7

For Linear Regression, moving to raw ratings increased MAE by 57.3% and RMSE by 45.6%, while R² decreased by 14.3%. For Random Forest, MAE increased by 51.1%, RMSE by 38.8%, and R² declined by 11.8%. The weighted ensemble showed a 57.0% increase in MAE, a 44.7% increase in RMSE, and a 13.7% decrease in R². Thus, substantial deterioration occurred for every matched model, demonstrating that prediction was more difficult under the raw review-level representation using the available sensory attributes.

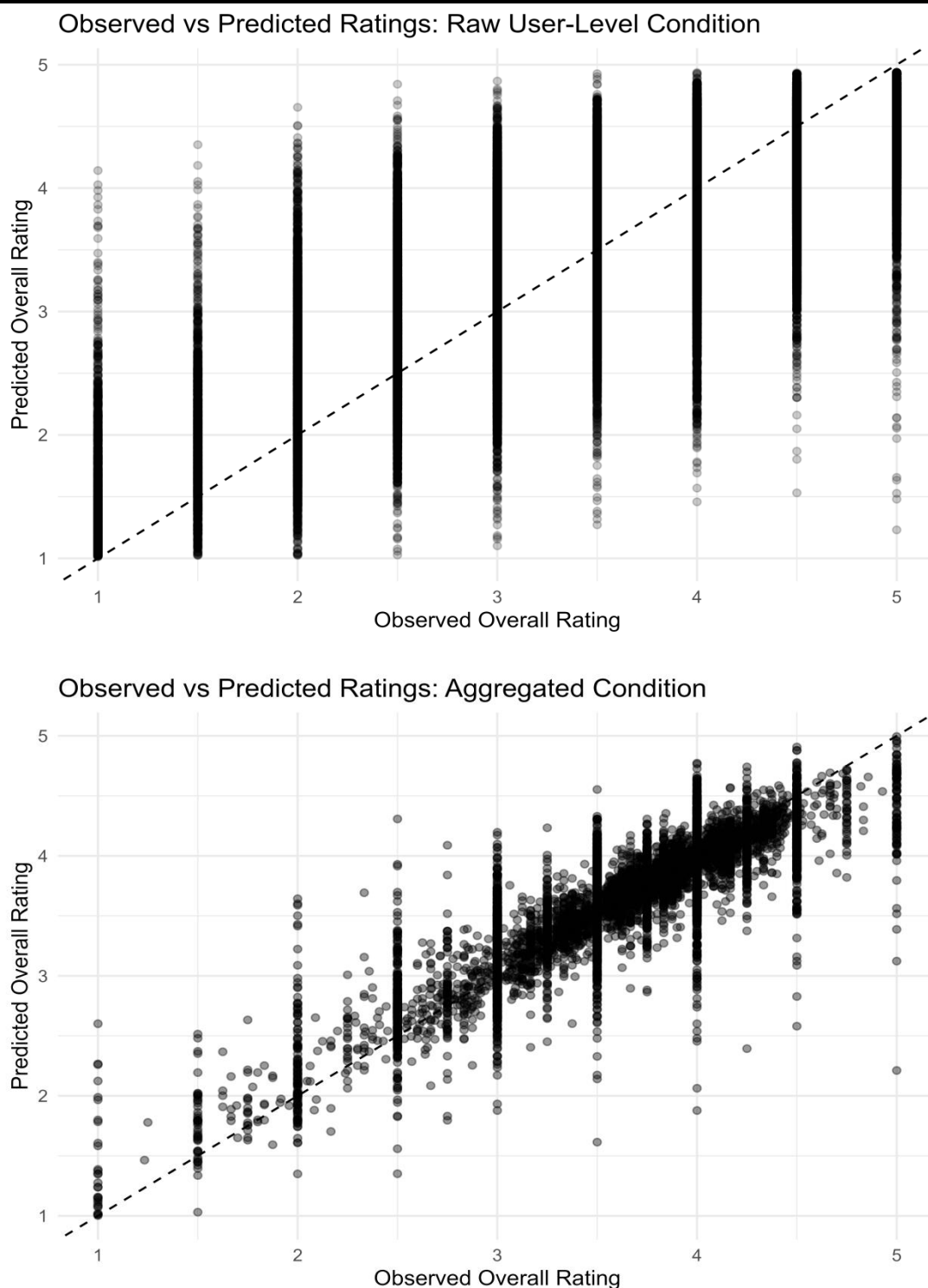


Fig. 2. Observed versus predicted overall ratings for the Random Forest model: (A) raw review-level representation; (B) aggregated beer-level representation.

4.4 Bootstrap robustness

Table 5. Random-Forest bootstrap confidence intervals based on 1,000 resamples.

Condition	Model	Metric	Estimate	95% CI lower	95% CI upper
Aggregated	Random Forest	MAE	0.204	0.199	0.208
Aggregated	Random Forest	RMSE	0.294	0.285	0.302
Raw	Random Forest	MAE	0.308	0.303	0.313
Raw	Random Forest	RMSE	0.408	0.401	0.416

Note. Bootstrap resampling was paired by held-out beer ID. All raw reviews belonging to a sampled beer were retained, including repeated selections. Random Forest predictions were held fixed; models were not retrained. Raw metrics are review-weighted, while aggregated metrics are beer-weighted.

Table 5 reports Random Forest test-set uncertainty under paired beer-cluster resampling. Aggregated RMSE was 0.293661 (95% bootstrap CI: 0.285447–0.302000), compared with raw RMSE of 0.407673 (95% CI: 0.400844–0.415761). Aggregated MAE was 0.203665 (95% CI: 0.199311–0.207951), whereas raw MAE was 0.307646 (95% CI: 0.303439–0.312668). The observed raw-minus-aggregated RMSE difference was 0.114011; its bootstrap mean was 0.114122 (95% CI: 0.103320–

0.125027). All 1,000 paired resamples yielded a positive difference. The findings indicate that the lower error of the aggregated representation is robust to resampling held-out beers, while the contrast remains between outcomes evaluated at different observational levels.

4.5 Permutation feature importance

Table 6. Random-Forest permutation feature importance by target representation.

Feature	Permutation importance	Condition
review_taste	0.4083	Aggregated
review_palate	0.0860	Aggregated
review_aroma	0.0341	Aggregated
review_appearance	0.0130	Aggregated
beer_abv	0.0099	Aggregated
review_taste	0.4432	Raw
review_palate	0.1018	Raw
review_aroma	0.0253	Raw
beer_abv	0.0219	Raw
review_appearance	0.0131	Raw

Taste was the dominant predictor in both target conditions, with importance 0.4083 for aggregated ratings and 0.4432 for raw ratings. Palate ranked second (0.0860 and 0.1018), followed by aroma. Appearance and alcohol by volume contributed substantially less. The broad ranking was therefore stable despite the pronounced change in predictive difficulty, suggesting that target representation altered error magnitude more strongly than it altered the principal sensory drivers identified by the model.

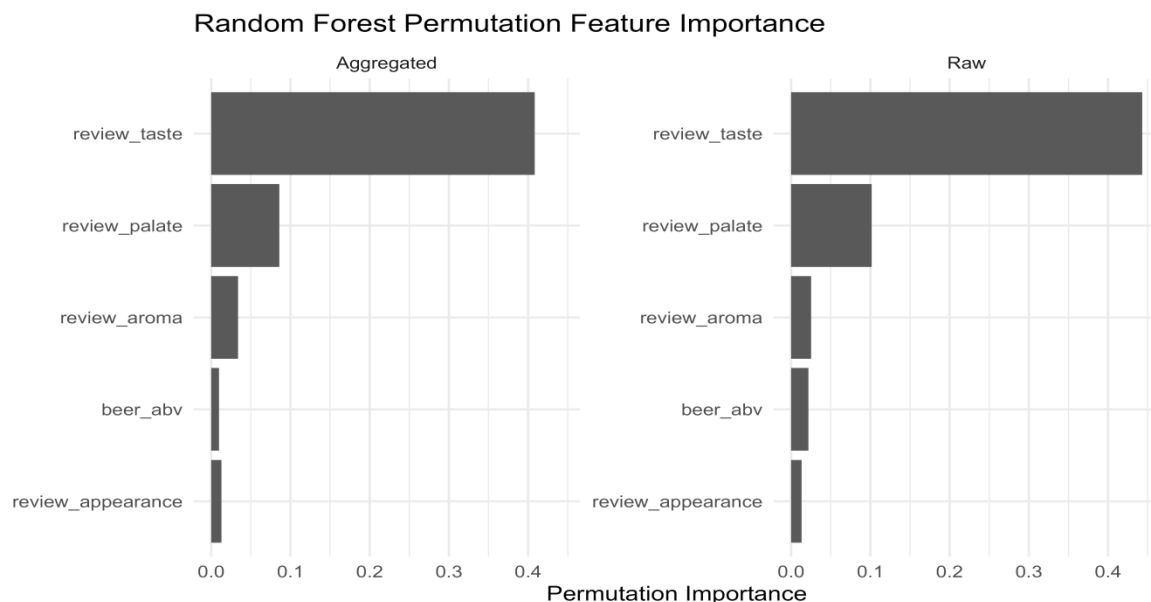


Fig. 3. Random-Forest permutation feature importance under aggregated beer-level and raw review-level representations.

4.6 Residual diagnostics

Random-Forest residuals under the raw condition were concentrated around zero, indicating limited systematic location bias in the central part of the distribution. The Q-Q plot, however, showed departures from normality in both tails. This pattern is compatible with heterogeneous review-level errors and indicates that a simple Gaussian residual assumption would not fully describe the predictive error distribution. These diagnostics are descriptive because Random Forest does not require normally distributed residuals for estimation.

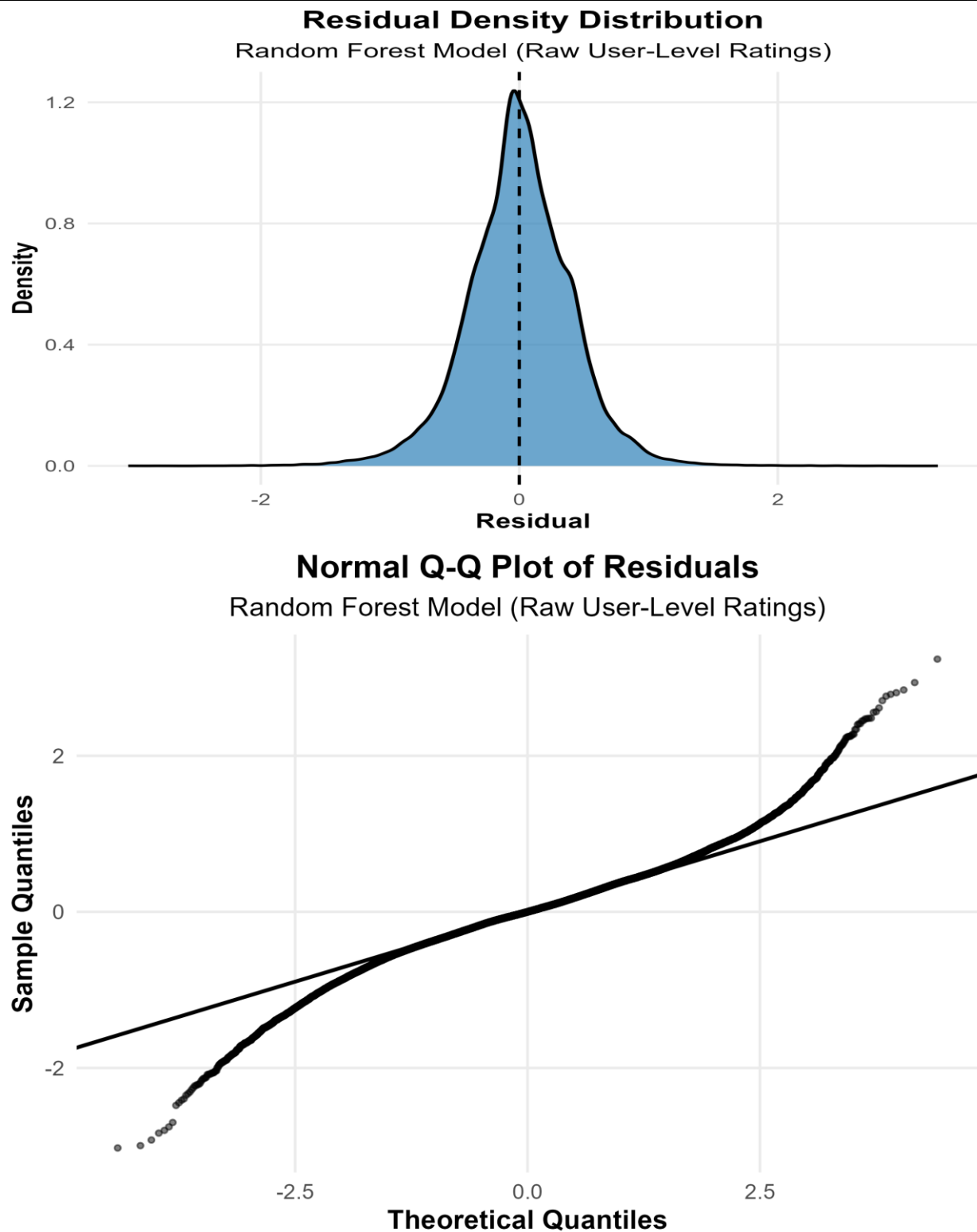


Fig. 4. Random-Forest residual diagnostics under the raw review-level representation: (A) residual distribution; (B) normal Q–Q plot.

(A) Residual density distribution illustrating prediction error concentration around zero.

(B) Normal Q–Q plot assessing residual distribution and prediction uncertainty. Together, these diagnostics provide evidence regarding model fit and the variability inherent in subjective user ratings.

4.7 Variance-component analysis

Table 7. Variance components from the crossed random-intercept model (n = 100,000 reviews).

Component	Variance	Proportion of total	Percent of total
Beer	0.1520	0.289	28.9
Reviewer	0.0370	0.070	7.0
Residual	0.3368	0.641	64.1

Beer identity accounted for 28.9% of total rating variance (variance = 0.1520; beer ICC = 0.289), while reviewer identity accounted for 7.0% (variance = 0.0370; reviewer ICC = 0.070). The residual component was largest at 64.1% (variance = 0.3368). This residual share represents within-beer, within-reviewer, contextual, interaction, and otherwise unmodeled review-

level variability. The results show that reviewer-specific differences are measurable, but most unexplained variability lies beyond stable beer and reviewer intercepts.

4.8 Review-frequency sensitivity analysis

Restricting the analysis to beers with at least five reviews retained 998,669 reviews (95.26% of the analytical sample) across 14,530 beers. Under the common beer-disjoint evaluation, the raw validation-weighted ensemble achieved $RMSE = 0.4071$, $MAE = 0.3078$, and $R^2 = 0.6563$, whereas the corresponding beer-level aggregated ensemble achieved $RMSE = 0.1551$, $MAE = 0.1166$, and $R^2 = 0.8906$. Thus, substantially lower prediction error under aggregation persisted after sparsely reviewed beers were excluded, indicating that the observed aggregation effect was not attributable solely to products represented by very few reviews.

5. DISCUSSION

The central finding is that changing the representation of consumer ratings changes the measured difficulty of contemporaneous overall-rating reconstruction. The best ensemble yielded $RMSE = 0.280$ and $R^2 = 0.790$ for aggregated beer-level observations, compared with $RMSE = 0.405$ and $R^2 = 0.682$ for individual reviews. Across matched model classes, raw-review $RMSE$ was 38.8–45.6% higher. This is a contrast between distinct prediction tasks, not a like-for-like comparison on one target: aggregation alters the overall-rating outcome, sensory predictors, sample size, and weighting convention. Aggregated errors weight each beer once, whereas raw errors weight each review. The consistent direction across Linear Regression, Random Forest, and the ensemble suggests that the difference is not specific to one algorithm.

The findings distinguish beer-level consensus estimation from individual-response reconstruction. Averaging overall and sensory ratings reduces disagreement and produces a smoother product-level task. Because sensory predictors and overall ratings are recorded within the same review, neither analysis constitutes a prospective test of personalized recommendations, future consumer choices, or user-specific ranking quality [10–13].

Beer identity explained 28.9% of rating variance, reviewer identity 7.0%, and the residual 64.1%. The residual may include consumer–product interactions, context, temporal effects, and measurement variation; it must not be interpreted wholly as meaningless noise [10,11,15,16].

Linear Regression slightly outperformed Random Forest for aggregated ratings, while Random Forest had a marginal advantage for raw reviews. These model differences were small relative to the representation gap, highlighting the importance of evaluation design rather than any single algorithm.

Taste dominated permutation importance in both representations, followed by palate and aroma. Similar predictor rankings alongside different error magnitudes suggest that the main distinction is the predictability of the respective evaluation tasks, not a wholesale change in influential sensory attributes [23,27,28].

Paired resampling of held-out beer clusters supported the stability of the Random Forest performance contrast: the observed raw-minus-aggregated $RMSE$ difference was 0.1140 (95% bootstrap CI: 0.1033–0.1250). Because the raw and aggregated targets represent different observational units, this comparison reflects a representation-level predictive contrast, not a paired difference between individual review outcomes.

The review-frequency sensitivity analysis provides complementary robustness evidence. After excluding beers with fewer than five reviews, substantially lower prediction error persisted under the aggregated representation, indicating that the observed aggregation effect was not driven solely by sparsely reviewed products. This remains an internal sensitivity analysis and does not establish a universal minimum review count for reliable consensus estimation.

The results reinforce a methodological point for retail analytics: model accuracy is interpretable only in relation to the underlying prediction target and decision context. Product averages can inform benchmarking, whereas individualized services require validation against individual outcomes [9–16].

5.1 Managerial and Practical Implications

The findings have several implications for retailers, digital platforms, recommendation-system developers, and consumer-analytics teams.

First, retailers should match their evaluation unit to the decision: beer-level means support product benchmarking and assortment comparisons, whereas individual-review outcomes are needed when assessing consumer-specific predictions. Platform interfaces could display review counts and rating dispersion alongside means so that apparent consensus is not mistaken for uniform preference. These are design implications, not interventions evaluated in this study.

Second, platforms should not promote product-level accuracy as evidence of personalization. The study reconstructs overall ratings from sensory scores recorded in the same review; it does not predict future purchases or demonstrate user-specific recommendation quality.

Third, review platforms should distinguish mean scores from disagreement. Two products with similar averages may have different rating distributions; displaying dispersion or segment-specific summaries could make that difference visible [31,32]. The present analysis motivates such designs but does not test their effects on consumer decisions.

Fourth, retail analytics reports should label the outcome level and weighting convention. Raw $RMSE$ is review-weighted, whereas aggregated $RMSE$ is beer-weighted; their contrast therefore describes different prediction tasks rather than a causal effect of averaging. This distinction matters when comparing vendors, ranking products, or claiming personalization gains.

Finally, taste, palate, and aroma remain useful descriptive signals at both levels [23,27,28]. Their predictive importance does not establish causal influence or eliminate the need for user history and context in personalized systems.

5.2 Future Research Directions

Future research should test the distinction between consensus and individual-preference predictability across additional retail and service contexts. Consumer sensory research indicates that preference formation can differ across product categories and evaluation settings [27,28]; cross-category evidence would therefore help determine whether the magnitude of the aggregation effect observed here generalizes beyond beer ratings.

Methodologically, future studies could incorporate user-disjoint evaluation to assess prediction for previously unseen consumers and compare the present structured models with matrix factorization, graph-based, deep, review-aware, and multimodal recommender systems [1–8]. Incorporating review text, consumer histories, temporal information, and user-item interactions may help explain additional individual-level variability. Review-based and multimodal recommendation research provides particularly relevant foundations for these extensions [7,8,24].

Future work should also investigate uncertainty-aware and distributional predictions rather than representing consumer preference exclusively through point estimates. Existing research on recommendation uncertainty provides methodological foundations for quantifying uncertainty around rating predictions and recommendations [10,11]. From a consumer-analytics perspective, an additional extension would be to examine whether rating distributions, personalized review summaries, or consumer-segment-specific evaluations provide better decision support than conventional aggregate ratings.

6. LIMITATIONS

Several limitations should be considered when interpreting the findings. First, the study examines one online-review domain and one platform; consequently, the magnitude of the observed aggregation effect may differ across retail categories, services, platforms, and consumer populations. The findings establish the phenomenon within the BeerAdvocate environment but do not constitute external validation across consumer markets. Although the ≥ 5 -review sensitivity analysis reduces concern that the principal result is driven by extremely sparse product evaluations, the five-review threshold is a robustness criterion rather than an empirically established minimum for reliable consensus estimation.

Second, the aggregated and raw conditions differ not only in representation of the outcome but also in observational granularity and predictor aggregation. The comparison therefore estimates the performance consequences of moving from an individual review-level representation to a beer-level aggregated representation; it should not be interpreted as isolating the causal effect of averaging the dependent variable alone.

Third, the predictive models use structured sensory ratings and alcohol content but exclude review text, beer style, temporal context, consumer history, purchase information, and consumer-product interaction features that may explain additional individual-level variation. Such information has been incorporated into more advanced review-aware and multimodal recommender systems [7,8].

Fourth, the development-test partition is beer-disjoint but not reviewer-disjoint. A reviewer may therefore contribute ratings to both development and test partitions even though the same beer cannot appear in both. The reported raw-rating performance consequently represents prediction for previously unseen beers within the same review platform rather than a strict cold-start evaluation for previously unseen consumers.

Fifth, the bootstrap analysis uses fixed predictions from the original held-out test set. Consequently, the reported intervals capture variability associated with resampling beers but do not account for uncertainty from refitting the predictive models, tuning hyperparameters, or selecting ensemble weights. Moreover, the raw and aggregated outcomes differ in their units of analysis and weighting, so differences in prediction error should not be interpreted as causal effects of aggregation.

Finally, the crossed random-intercept model was estimated on a reproducible 100,000-review subsample for computational tractability and uses random intercepts only. The 64.1% residual component should not be interpreted solely as consumer heterogeneity; it may contain consumer-product interactions, contextual effects, temporal variation, measurement variation, and other unmodeled sources.

7. CONCLUSION

This study shows that the analytical representation of online ratings substantially affects measured performance in reconstructing contemporaneous overall evaluations. Across more than one million BeerAdvocate reviews and beer-disjoint held-out evaluation, the best ensemble achieved $RMSE = 0.280$ and $R^2 = 0.790$ for aggregated beer-level observations, versus $RMSE = 0.405$ and $R^2 = 0.682$ for raw reviews. These RMSE values describe different targets and observation units rather than performance on the same outcome: aggregation changes the ratings and sensory predictors, the number of observations, and the relative weighting of beers and reviews. The direction of the contrast persisted in a sensitivity analysis restricted to beers with at least five reviews.

The evidence supports a distinction between reconstructing collective product evaluations and reconstructing individual review responses; it does not directly test personalized recommendation performance. Taste, palate, and aroma remained the principal predictors in both representations. Variance-component analysis attributed 28.9% of observed rating variance to beer identity, 7.0% to reviewer identity, and 64.1% to residual review-level variation, indicating substantial variability beyond stable beer and reviewer effects.

For retail analytics, beer-level averages can support consensus estimation, product benchmarking, and assortment decisions, while review-level analyses retain disagreement relevant to understanding individual responses. Neither result establishes accuracy for future purchases, user-specific rankings, or personalized recommendations. Platforms should report the evaluation unit and weighting convention, communicate rating dispersion alongside averages, and validate personalization claims using genuinely user-specific prospective outcomes.

DECLARATIONS

Ethics statement: This study analyzed an existing publicly accessible online-review dataset and involved no direct interaction with reviewers. Reviewer profile identifiers are present in the source schema and were used only for analytical grouping and variance-component estimation; the study did not attempt to identify individual reviewers.

Competing interests: The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability: The analytical scripts, model evaluation outputs, robustness-analysis results, and supporting reproducibility materials associated with this study are publicly archived on Zenodo (Version 1.0.0; <https://doi.org/10.5281/zenodo.23238507>). The associated development repository is available at <https://github.com/sahlworkskrd-create/Consumer-Consensus-to-Individual-Preference-Repository>. The original BeerAdvocate review dataset is third-party material and can be obtained from the UCSD Recommender Systems Dataset Repository (<https://cseweb.ucsd.edu/~jmcauley/datasets.html>). The repository documentation identifies any original analytical materials or partition identifiers that are not included in the archive.

Code availability: A reproducible Python script for the revised paired beer-cluster bootstrap, the held-out Random Forest prediction files, and the 1,000-replicate outputs are available in the accompanying submission materials. The script reproduces all archived bootstrap replicates to numerical tolerance 1e-12 using NumPy default_rng(20261011).

Reproducibility statement: The analysis comprised 1,048,340 reviews, 42,716 beers, and 28,761 reviewers. Development/test partitions were beer-disjoint (seed 123), with internal beer-disjoint validation (seed 456). Revised confidence intervals use 1,000 paired resamples of 8,544 held-out beers and fixed Random Forest predictions. The supplied Python verification script uses NumPy default_rng(20261011) and reproduces the archived replicate results; an R implementation with the same numeric seed is not replicate-identical.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process: During preparation of this manuscript, the author used ChatGPT (OpenAI) for language refinement, structural editing, and presentation of methodological descriptions. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

REFERENCES

- [1] Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.), 2011. Recommender Systems Handbook, first ed. Springer. <https://doi.org/10.1007/978-0-387-85820-3>.
- [2] Raza, S., Rahman, M., Kamawal, S., Toroghi, A., Raval, A., Navah, F., Kazemeini, A., 2026. A Comprehensive Review of Recommender Systems: Transitioning from Theory to Practice. Computer Science Review 59, 100849. <https://doi.org/10.1016/j.cosrev.2025.100849>
- [3] Priyansha, U., Nizar Banu, P. K., 2026. Graph Neural Networks in Recommendation Systems for Superior User Experiences. <https://doi.org/10.1002/9781394351572.ch6>
- [4] Anand, V., & Maurya, A. K., 2024. A Survey on Recommender Systems Using Graph Neural Network. ACM Transactions on Information Systems, 43, 1–49. <https://doi.org/10.1145/3694784>.
- [5] Gao, C., Zheng, Y., Li, N., Li, Y., Qin, Y., Piao, J., Quan, Y., Chang, J., Jin, D., He, X., Li, Y., 2023. A survey of graph neural networks for recommender systems: Challenges, methods, and directions, ACM Trans. Recomm. Syst. 1, 1–51. <https://doi.org/10.1145/3568022>.
- [6] Wu, S., Sun, F., Zhang, W., Xie, X., Cui, B., 2022. Graph neural networks in recommender systems: A survey, ACM Comput. Surv. 55 (5), 1–37. <https://doi.org/10.1145/3535101>
- [7] Hasan, E., Rahman, M., Ding, C., Huang, J.X., Raza, S., 2026. Review-based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives. ACM Computing Surveys. 58(1). <https://doi.org/10.1145/3742421>
- [8] Liu, Q., Hu, J., Xiao, Y., Zhao, X., Gao, J., Wang, W., Li, Q., Tang, J., 2025. Multimodal recommender systems: A survey, ACM Comput. Surv. 57 (2) Article 26, <https://doi.org/10.1145/3695461>.
- [9] Zhang, K., Cao, Q., Sun, F., Wu, Y., Tao, S., Shen, H., Cheng, X., 2026. Robust Recommender System: A Survey and Future Directions. ACM Computing Surveys. 58(1). <https://doi.org/10.1145/3757057>
- [10] Coscrato, V., Bridge, D., 2023. Estimating and evaluating the uncertainty of rating predictions and top-n recommendations in recommender systems. ACM Trans Recomm Syst. 1(2):1–34. <https://doi.org/10.1145/3584021>
- [11] Paliwal, C., Majumder, A., Kaveri, S., 2024. Predictive Relevance Uncertainty for Recommendation Systems, in: Proceedings of the ACM Web Conference (WWW '24), ACM, New York, NY, USA, 2024, pp. 3900–3909, <https://doi.org/10.1145/3589334.3645689>.
- [12] Ibrahim, O. A. S., Younis, E. M. G., Mohamed, E. A., & Ismail W. N., 2025. Revisiting recommender systems: An investigative survey. Neural Computing and Applications, 37, 2145–2173. <https://doi.org/10.1007/s00521-024-10828-5>
- [13] Ge, Y., Liu, S., Fu, Z., Tan, J., Li, Z., Xu, S., Li, Y., Xian, Y., & Zhang, Y. 2024. A Survey on Trustworthy Recommender Systems. ACM Transactions on Recommender Systems, 3(2), Article 13, 1–68. <https://doi.org/10.1145/3652891>
- [14] Zhao, Y., Wang, Y., Liu, Y., Cheng, X., Aggarwal, C. C., & Derr, T. 2025. Fairness and Diversity in Recommender Systems: A Survey. ACM Transactions on Intelligent Systems and Technology, 16(1), 1–28. <https://doi.org/10.1145/3664928>
- [15] Northcutt, C. G., Jiang, L., Chuang, I.L., 2021. Confident Learning: Estimating Uncertainty in Dataset Labels. Journal of Artificial Intelligence Research, 70, 1373–1411. <https://doi.org/10.1613/jair.1.12125>
- [16] Song, H., Kim, M., Park, D., Shin, Y., Lee, J.-G., 2023. Learning From Noisy Labels With Deep Neural Networks: A Survey. IEEE Transactions on Neural Networks and Learning Systems, 34(11), 8135–8153. <https://doi.org/10.1109/TNNLS.2022.3152527>
- [17] Soni, D., Singh, K., Choudhary, A., Pathak, D., 2025. Machine Learning-Based Credit Risk Prediction: A Systematic Review of Techniques, Challenges, and Future Directions. International Journal of Research and Review in Applied Science, Humanities, and Technology. <https://doi.org/10.71143/4086m281>
- [18] Choi, J.W., Hidayat, M.S., Cho, S.B., Hwang, W.H., Lee, H., Cho, B.K., Kim, M.S., Baek, I., Kim, G., 2025. Recent Trends in Machine Learning, Deep Learning, Ensemble Learning, and Explainable Artificial Intelligence Techniques for Evaluating Crop Yields Under Abnormal Climate Conditions. Plants, 14(18), 2841. <https://doi.org/10.3390/plants14182841>
- [19] Zhang, Y., Chen, X., 2020. Explainable recommendation: A survey and new perspectives. Foundations and Trends in Information Retrieval, 14(1), 1–101. <https://doi.org/10.1561/15000000066>
- [20] Marconi, L., Matamoros Aragon, R. A., & Epifania, F., 2022. Discovering the Unknown Suggestion: a Short Review on Explainability for Recommender Systems, in: Proceedings of the 2nd Italian Workshop on Artificial Intelligence and

- Applications for Business and Industries (AIABI 2022), CEUR-WS, pp. 1–15. Available from: <https://ceur-ws.org/Vol-3463/paper3.pdf>.
- [21] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [22] Guo, S., Zhang, S., Sun, W., Ren, P., Chen, Z., Ren, Z., 2023. Towards Explainable Conversational Recommender Systems, in: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*, ACM, New York, NY, USA, pp. 2786–2795. <https://doi.org/10.1145/3539618.3591884>
- [23] Schreurs, M., Piampongsant, S., Roncoroni, M., Cool, L., Herrera-Malaver, B., Vanderaa C., Theßeling F.A., Kreft L., Botzki A., Malcorps P., Daenen L., Wenseleers T., Verstrepn K. J., 2024. Predicting and improving complex beer flavor through machine learning. *Nature Communications*, 15, 2368. <https://doi.org/10.1038/s41467-024-46346-0>
- [24] Jiang, H., Duan, J., Qu, Z., Wang, J., 2024. MARE: Multi-aspect rationale extractor on unsupervised rationale extraction. *Proceedings of EMNLP*, 11734–11745. <https://doi.org/10.18653/v1/2024.emnlp-main.655>
- [25] Jannach, D., Said, A., Tkalcic, M., Zanker, M., 2025. Recommender Systems for Good (RS4Good): Survey of Use Cases and a Call to Action for Research that Matters. *ACM Trans. Recomm. Syst.* <https://doi.org/10.1145/3746648>
- [26] Kim, K., Kim, S., Lee, G., Jung, J., Shin, K., 2026. A comprehensive survey on multi-behavior recommender systems: extended taxonomy and recent advances. *International Journal of Data Science and Analytics* 22 (40). <https://doi.org/10.1007/s41060-025-01019-z>
- [27] Zhu, Y., Liu, J., & Liu, Y. 2023. Understanding the relationship between umami taste sensitivity and genetics, food-related behavior, and nutrition. *Current Opinion in Food Science*, 50, 100980. <https://doi.org/10.1016/j.cofs.2022.100980>
- [28] Engler, T. H., Winter, P., Schulz, M., 2015. Understanding online product ratings: A customer satisfaction model. *Journal of Retailing and Consumer Services*, 27, 113–120. <https://doi.org/10.1016/j.jretconser.2015.07.010>.
- [29] Yi, J., Oh, Y. K., 2022. The informational value of multi-attribute online consumer reviews: A text mining approach. *Journal of Retailing and Consumer Services*, 65, 102519. <https://doi.org/10.1016/j.jretconser.2021.102519>.
- [30] Li, K., Chen, Y., Zhang, L., 2020. Exploring the influence of online reviews and motivating factors on sales: A meta-analytic study and the moderating role of product category. *Journal of Retailing and Consumer Services*, 55, 102107. <https://doi.org/10.1016/j.jretconser.2020.102107>
- [31] Lee, S., Baek, H., 2021. Does the dispersion of online review ratings affect review helpfulness? *Computers in Human Behavior* 117, 106670. <https://doi.org/10.1016/j.chb.2020.106670>.
- [32] Langan, R., Besharat, A., Varki, S., 2017. The effect of review valence and variance on product evaluations: An examination of intrinsic and extrinsic cues. *International Journal of Research in Marketing* 34(2), 414–429. <https://doi.org/10.1016/j.ijresmar.2016.10.004>.
- [33] Liu, Y., Zhang, Q., Liu, X., 2026. Decoding inconsistent online reviews in retail: Dual cues, product type moderation, and implications for consumer decision-making. *Journal of Retailing and Consumer Services* 93, 104948. <https://doi.org/10.1016/j.jretconser.2026.104948>.

SUPPLEMENTARY MATERIAL

This supplementary material provides additional experimental configurations, model specifications, and sensitivity-analysis results supporting the analyses reported in the main manuscript.

Table S1. Experimental configuration.

Parameter	Value
Source observations	1,048,575
Analytical raw observations	1,048,340
Unique beers	42,716
Unique reviewers	28,761
Development/test split	80% / 20%
Split unit	Beer identifier (beer-disjoint)
Internal validation	Beer-disjoint fit/validation within development data
Random seed	123 (primary); 456 (internal validation)
Bootstrap iterations	1,000
Predictors	Taste, aroma, palate, appearance, ABV
Aggregated target	Mean overall rating within beer
Raw target	Individual review_overall rating
Variance-component sample	100,000 reviews

Table S2. Final model and ensemble settings.

Model	Final configuration
Linear Regression	Ordinary least squares
Random Forest - raw	500 trees; mtry = 3; min.node.size = 10; permutation importance
Random Forest - aggregated	500 trees; mtry = 3; min.node.size = 5; permutation importance
Weighted ensemble - raw	RF weight = 0.55; LM weight = 0.45
Weighted ensemble - aggregated	RF weight = 0.26; LM weight = 0.74

Ensemble weights were selected on the internal beer-disjoint validation partition by minimizing RMSE over candidate Random-Forest weights from 0 to 1 in increments of 0.01. The held-out test set was not used to determine ensemble weights.

Table S3. Sensitivity analysis restricted to beers with at least five reviews

Representation	Model	RMSE	MAE	R ²
Individual reviews	Linear regression	0.4134	0.3085	0.6455
Individual reviews	Random forest	0.4072	0.3086	0.6562
Individual reviews	Validation-weighted ensemble	0.4071	0.3078	0.6563
Beer-level aggregation	Linear regression	0.1564	0.1178	0.8887
Beer-level aggregation	Random forest	0.1601	0.1199	0.8834
Beer-level aggregation	Validation-weighted ensemble	0.1551	0.1166	0.8906

Note: The sensitivity analysis retained 998,669 reviews from 14,530 beers with at least five eligible reviews. Beer identifiers were mutually exclusive across training (9,300 beers), validation (2,324 beers), and test (2,906 beers) partitions. Ensemble weights were selected using validation RMSE. The selected RF/LR weights were 0.82/0.18 for the individual-review representation and 0.36/0.64 for the aggregated representation. Random Forest models used 500 trees, mtry = 3, and minimum node sizes of 10 and 5 for the individual-review and aggregated representations, respectively.

The primary analysis used five predictors: taste, aroma, palate, appearance, and alcohol by volume (ABV). The ≥ 5 -review sensitivity analysis used four sensory predictors (taste, aroma, palate, and appearance), excluding ABV. Consequently, the sensitivity analysis evaluates the persistence of the representation-level performance difference under a restricted sample and sensory-only specification, rather than exactly replicating the primary five-predictor analysis.

Table S4. Paired beer-cluster bootstrap contrast in Random Forest RMSE (raw minus aggregated).

Observed difference	Bootstrap mean	95% percentile CI	Positive resamples
0.114011	0.114122	0.103320–0.125027	1,000/1,000

Note. The same sampled beer IDs were used for both representations in every replicate. The observed difference and bootstrap mean are distinct quantities. The 1,000/1,000 proportion is a descriptive bootstrap result, not an exact p-value.

S5. Reproducibility of the paired beer-cluster bootstrap

The verification uses the common 8,544 held-out beer IDs, fixed Random Forest predictions, and 1,000 paired resamples with replacement. For each sampled beer, the aggregated metric uses one beer-level prediction and the raw metric uses all associated review-level predictions, including repeated contributions when a beer is sampled multiple times. Raw error is review-weighted and aggregated error is beer-weighted. The archived bootstrap replicates were generated using NumPy `default_rng(20261011)`. The companion Python script reproduces all 1,000 archived replicates to absolute tolerance $1e-12$; R `set.seed(20261011)` produces different random draws. No models are refitted.