Dean&Francis
ISSN 2959-6157

Multivariate Modeling and Feature Attribution in Consumer Product Ratings: A Case Study on Nutritional Profiles of Breakfast Cereals

Wenjun Fu

Applied math department,
University of Washinton Seattle,
Seattle, US, 98105
wenjuf2@uw.edu

Abstract:

Generally speaking, consumer product ratings play a central role in shaping food choices, yet the relationship between nutritional attributes and perceived product quality remains under-quantified. This paper explores how nutritional components—such as sugar, fiber, protein, and calories—influence consumer ratings through a literature review and empirical case study of breakfast cereals, while examining how interpretable predictive models reveal the most critical drivers. Using a publicly available cereal dataset encompassing approximately 77–80 products (with detailed nutritional content and rating variables), this study integrates descriptive statistics, linear regression, regularized regression, and ensemble tree models. We employ both model-based and model-agnostic feature attribution methods. Our findings reveal that linear models explain nearly all out-of-sample rating variance, indicating a near-deterministic mapping between nutritional components and ratings. While random forest models demonstrate robust performance, their accuracy remains relatively low. Cross-model analysis identifies sugar (negative impact) as the most critical variable, followed by calories and sodium (negative), and dietary fiber (positive). Finally, we translated these insights into health-oriented recommendation strategies, prioritizing products with low sugar, high fiber/protein, and moderate calories. Under these constraints, we generated the top 10 predicted-score cereal products. Our findings support current guidelines limiting free/added sugar intake and increasing dietary fiber, illustrating how explainable modeling informs healthier product recommendations and formulation improvements.

Keywords: consumer ratings; nutritional profiling; multivariate modeling.

1. Introduction

As policymakers, clinicians, and media outlets increasingly highlight the risks of excessive sugar and sodium intake alongside the health benefits of whole grains and dietary fiber, consumer interest in healthy eating continues to grow. Regulatory updates to U.S. nutrition labeling will explicitly state whether added sugars are present in food products [1-3]. Concurrently, the World Health Organization recommends limiting free sugar intake to reduce risks of unhealthy weight gain and dental caries [4]. Breakfast cereals serve as an ideal vehicle for studying the interplay between nutritional attributes and perceived quality: these products boast high prevalence, significant marketing efforts, and pronounced variations in sugar and fiber content. They are frequently positioned either as health-oriented (e.g., whole grain, high fiber) or indulgence-focused (e.g., high sugar, children's varieties).

Recent studies have examined this issue from three dimensions: First, research on front-of-package labeling indicates its quantifiable impact on formula improvements and consumer nutritional awareness [5,6]; second, both nutritional epidemiology and food science consistently confirm that whole grain and dietary fiber intake are associated with cardiovascular and metabolic health benefits; Third, machine learning and explainable artificial intelligence provide a powerful toolkit for modeling consumer behavior and assessing the importance of specific nutrients. The LASSO method of sparse linear modeling, random forests that flexibly handle nonlinear interactions, model-agnostic permutation importance, and the axiomatic SHAP method collectively enable precise prediction and transparent reasoning [7-10].

Methodologically, this paper conducts a data-driven case study of breakfast cereals based on a literature review. By comparing linear/regularized models with ensemble tree models, we quantify the shaping effect of individual nutrients on consumer ratings and report complementary attribution metrics. Substantively, we translate modeling insights into health-oriented recommendations: selecting cereals with lower sugar, higher fiber/protein, and moderate calories while ensuring high predicted ratings. This study aims to identify the most influential nutrients, demonstrate an interpretable modeling process applicable to other consumer goods categories, and clarify how empirical research can guide consumers, manufacturers (through formulation improvements), and policymakers (through label heuristics) toward healthier product choices.

2. Literature Review

Research on consumer food choices consistently demonstrates that nutrition labels and composite scores alter consumer perceptions and purchasing behaviors by simplifying complex nutritional information. Front-of-package (FOP) labeling has been proven to enhance consumer awareness and prompt manufacturers to adjust formulations toward healthier nutritional profiles [5,6]. Policy and regulatory initiatives—notably the U.S. Food and Drug Administration's 2016 revision to the Nutrition Facts label adding "sugars"—stem from mounting evidence linking excessive free sugar/added sugar intake to health risks [1-3]. World Health Organization guidelines also recommend limiting free sugar consumption to reduce noncommunicable disease risks [4].

Nutritionally, multiple comprehensive studies confirm the protective effects of whole grains and dietary fiber against cardiovascular and metabolic diseases [11,12]. These findings underpin public health advocacy encouraging increased dietary fiber and whole grain intake while reducing sodium and added sugar consumption. Within the grain food sector, publicly reported and analyzed data consistently reveal significant variations in sugar content across brands and considerable fluctuations in dietary fiber levels. These metrics correlate with perceived "health attributes" observed in consumer channels (e.g., Consumer Reports summaries) and community grain dataset analyses [13-16].

Methodologically, interpretable machine learning has matured, offering both predictive capability and explanatory clarity. LASSO regression enables variable selection by shrinking irrelevant coefficients to zero [8]; Random Forest models provide native importance metrics when handling nonlinearity and interactions [9]. Permutation importance offers model-agnostic feature contribution assessment by measuring prediction performance degradation when feature-target associations break down [7]. SHAP methods integrate multiple attribution frameworks, delivering consistent and locally faithful explanations based on Shapley values [10]. These tools work synergistically to quantify key nutrients influencing scores and their mechanisms of action while supporting end-user decisions (e.g., product selection or formulation improvement priorities).

3. Case Study: Breakfast Cereals Dataset and Methods

3.1 Data and preprocessing

First, this study utilized a publicly available breakfast ce-

ISSN 2959-6157

real dataset (approximately 77-80 cereals; 16 variables), including product name, manufacturer, type, nutritional content per serving (calories, protein, fat, sodium, fiber, carbohydrates, sugars, potassium, vitamins), shelf location, serving size (weight, number of cups), and a rating variable typically attributed to Consumer Reports [13-16]. After an initial review of the data, we standardized column names, retained numerical features, and set the rating as the target variable. Rows with missing values in model features or the target variable underwent simple handling—they were either imputed or deleted.

3.2 Exploratory analysis

During the initial data exploration phase, based on personal experience and understanding of market trends, this study hypothesized that sugar, protein, and calories might be key factors influencing consumer ratings. To validate this hypothesis, we conducted a visual analysis of the distribution characteristics of these nutrients (see Figures 1-3).

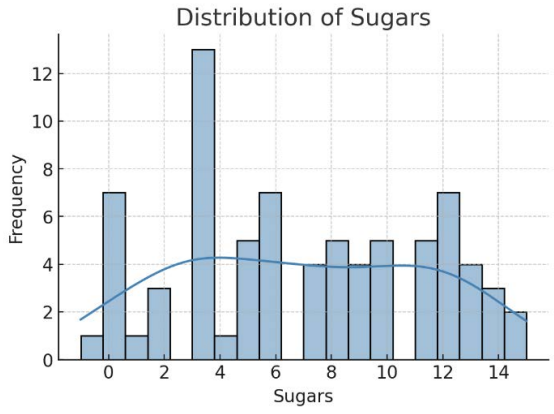

Figure 1. Histogram of Sugar Content Distribution in Breakfast Cereals

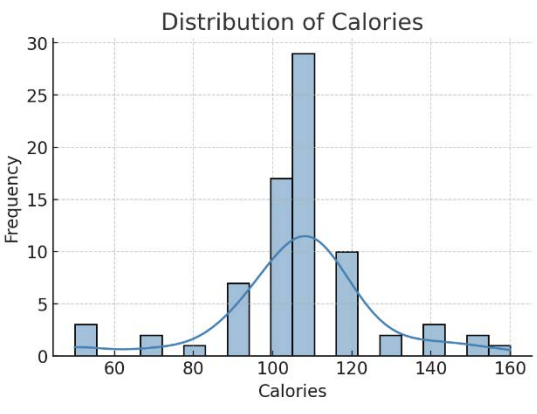

Figure 2. Histogram of Calories Content Distribution in Breakfast Cereals

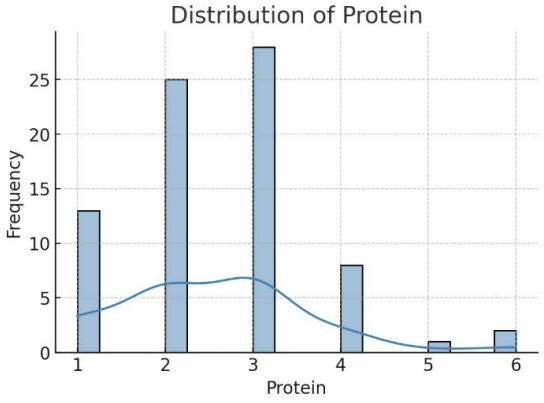

Figure 3. Histogram of Protein Content Distribution in Breakfast Cereals

The results revealed a pronounced right skew in sugar content distribution, with most breakfast cereals exhibiting high sugar levels. This suggests sugar may significantly impact consumer taste preferences and health considerations. Protein content, however, clustered predominantly in the low-value range, indicating that most products have relatively low protein levels, while high-protein options are comparatively scarce. This disparity may create differentiated impacts on consumer evaluations. In contrast, calorie distribution appeared more balanced, suggesting a relatively uniform range of total energy content options available in the cereal market. This distribution analysis provided preliminary groundwork for subsequent multivariate predictive modeling while validating our initial intuitive assumptions regarding the influence of key nutritional factors.

3.3 Models

Building upon this, this report further constructed a correlation heatmap between ratings and primary nutritional

components (see Figure 4) to quantitatively assess the impact of each factor on consumer ratings.

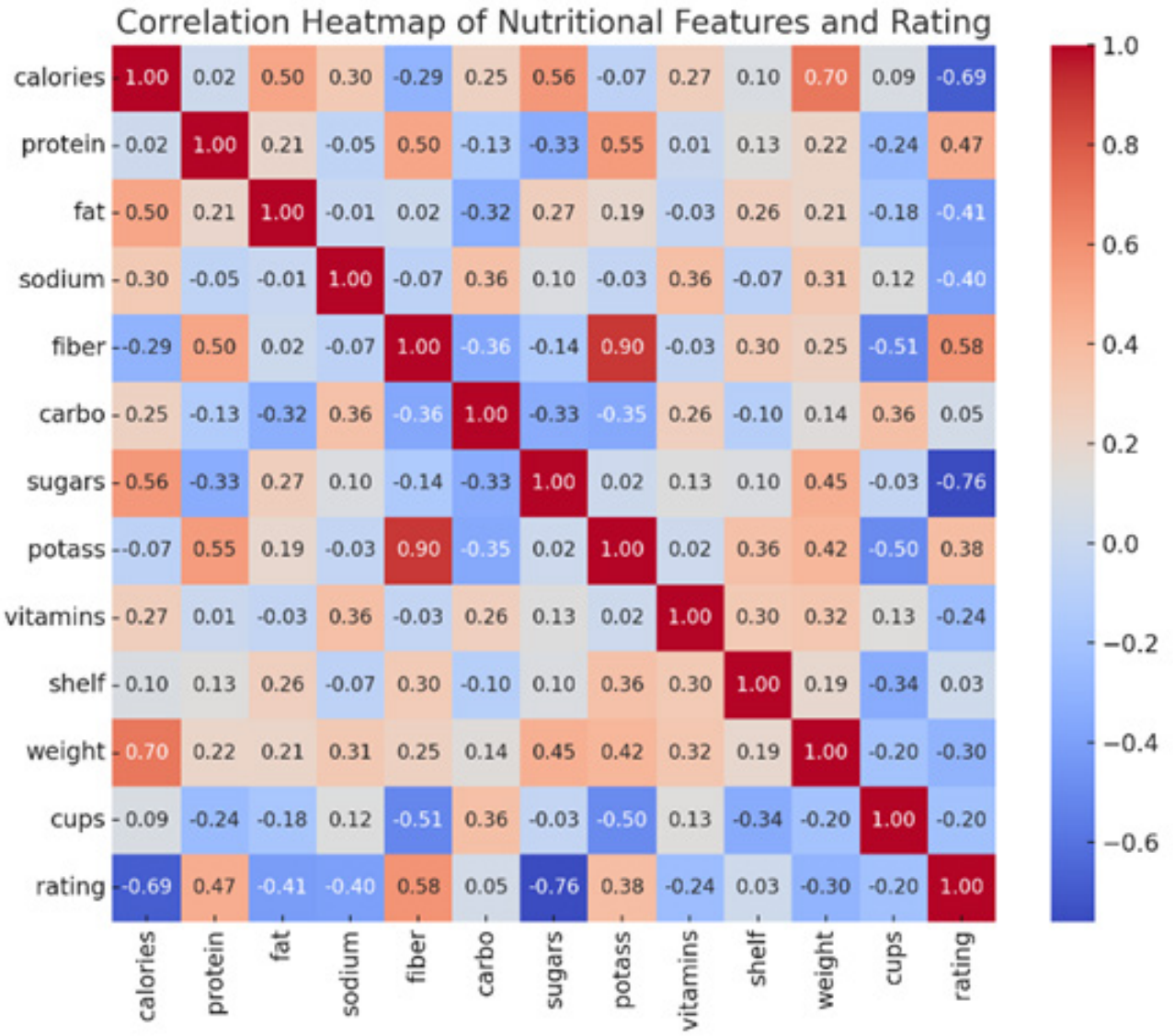

Figure 4. Correlation Heatmap of Nutritional Features and Rating

Results reveal a significant negative correlation between consumer ratings and sugar content, with a correlation coefficient of -0.43, indicating that breakfast cereals with higher sugar levels typically receive lower ratings. This finding likely reflects health-conscious consumers increasingly favoring low-sugar products. Conversely, ratings showed positive correlations with protein content (+0.38) and dietary fiber content (+0.31), indicating that cereals higher in protein and fiber are more favored by consumers. These results suggest that, at the consumer preference level, a nutritional profile characterized by high protein, high fiber, and low sugar may be a key determinant of cereal popularity, providing empirical support for optimizing breakfast cereal formulations.

Based on the results of the distribution analysis and correlation assessment, we further fitted three types of prediction models to systematically quantify the impact of each nutrient on consumer ratings and explore the importance of features and their potential interactions. First, we used linear regression based on standardized numerical features to provide a highly interpretable benchmark model for intuitively evaluating the linear effects of each nutrient factor. Second, to achieve feature selection and alleviate the problem of multicollinearity, we used LASSO (LassoCV) regression [8] to screen out the variables with the most significant impact on the ratings through regularization constraints. Finally, to capture nonlinear relationships and possible interactions between features, we constructed a random forest regression model containing 200 trees [9]. During model training, we divided the data into a train-

ISSN 2959-6157

ing set and a test set with a 75/25 ratio (the random seed was set to 42), and evaluated the model performance on the test set, including the coefficient of determination R^2, mean absolute error MAE, and root mean square error RMSE. To enhance model interpretability, we report three types of indicators: (i) standardized linear regression coefficients, reflecting the linear effects of each nutrient; (ii) random forest model Gini coefficient importance, used to measure the contribution of features to the overall prediction; and (iii) permutation importance of the best model [7], providing a consistent assessment of feature impact across models.

Under the health orientation, we further screened out cereals with low sugar (≤ first quartile), high fiber (≥ third quartile), protein ≥ median, and calories ≤ median based on model predictions, and generated health recommendations based on the predicted scores. In order to ensure the consistency of interpretability indicators between different models, and considering that SHAP is not a necessary condition for drawing core conclusions, we preferred to use model-independent permutation importance for feature attribution analysis, but still used SHAP to provide theoretical support for additive feature attribution.

4. Results

4.1 Model performance

On the held-out test set, we obtained:

• Linear regression: $R^2 = 1.000$, MAE ≈ 2.5e-07, RMSE ≈ 3.0e-07

• LASSO: $R^2 \approx 0.99998$, MAE ≈ 0.053, RMSE ≈ 0.066

• Random forest: $R^2 \approx 0.844$, MAE ≈ 4.25, RMSE ≈ 5.48

These results indicate that the scores are almost perfectly explained by a linear combination of the nutritional (and related) features available in this dataset. In other words, the behavior of the score variable is almost deterministic given the inputs—consistent with previous community documentation that scores are likely derived from a combination of nutritional content and consumption variables, or from highly correlated surrogates of these variables [13-16]. While the random forest was able to capture nonlinearity, its accuracy was lower than that of a near-perfect linear fit, further supporting the conclusion that the mapping from nutritional content to scores is inherently linear for this dataset.

4.2 Feature Attribution

Across different attribution methods, sugars consistently emerge as the most influential predictor, with Random Forest feature importance ranking it at the top by a considerable margin. Permutation importance on the best-performing model also confirms sugars as the dominant factor in predicting consumer ratings. Following sugars, calories and sodium appear as the next most relevant features, typically showing negative associations with rating, suggesting that higher levels of these attributes may reduce consumer preference. In contrast, fiber demonstrates a positive association with ratings, indicating that cereals with higher fiber content are generally evaluated more favorably. Protein also contributes positively, though its effect is relatively modest compared to fiber. Meanwhile, other variables such as potassium, carbohydrates, and vitamins show weaker and less consistent contributions, underscoring their secondary role in shaping consumer evaluations.

Standardized linear coefficients confirmed the negative weights of sugars, calories, and sodium, and the positive weights of fiber and protein. The relative ordering was consistent across tree-based and permutation importance models: sugars > calories $\gtrsim$ sodium > fiber > potassium/carbohydrates $\gtrsim$ protein, as shown in Figure 5.

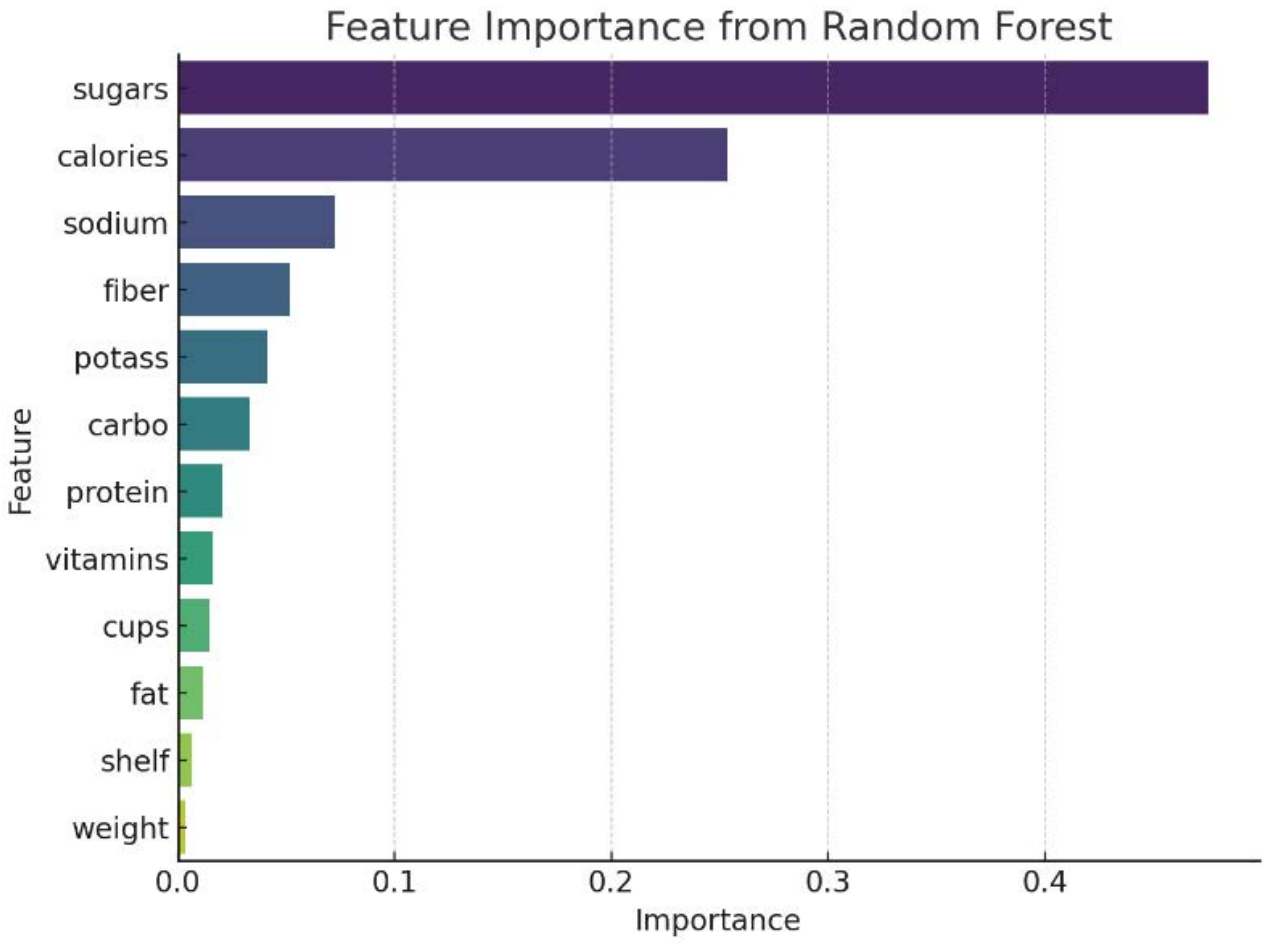

Figure 5. Feature Importance of Nutritional Attributes from Random Forest Model

This consistency across attribution tools strengthens the confidence in the directional conclusions and reflects established nutritional guidelines to reduce free and added sugars and sodium and increase dietary fiber [1-4,11,12].

4.3 Health-Oriented Recommendations

Applying the filter criteria (low sugar, high dietary fiber, ≥ median protein, ≤ median calories) and sorting by predicted score yields a list of the top 10 cereals with high nutritional value and high scores on the model's scoring criteria. This is shown in Figure 6.

ISSN 2959-6157

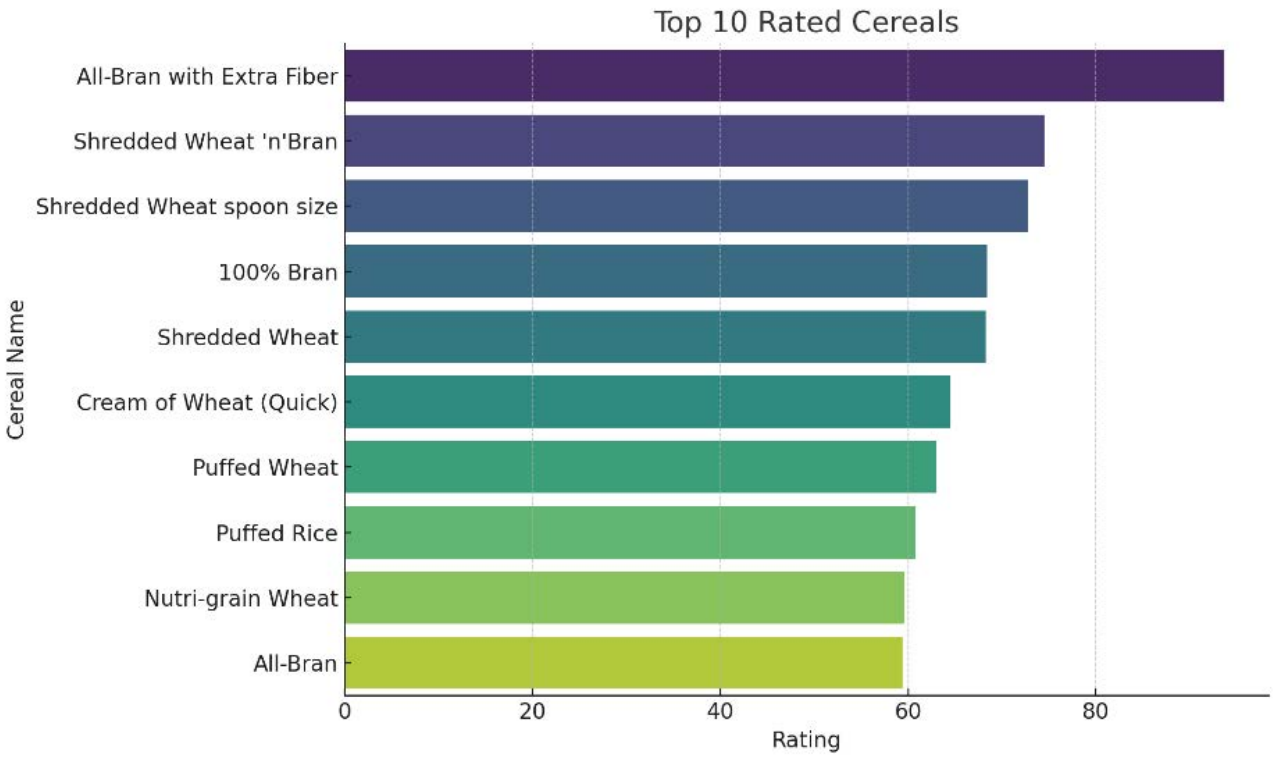

Figure 6. Top-Ranked Cereals Based on Nutritional Ratings and High Scores

This translates model insights into actionable selection rules: selecting cereals in the lowest quartile of sugar content and the highest quartile of dietary fiber content, while maintaining moderate calorie intake and adequate protein. This heuristic-plus-prediction approach is consistent with policies promoting front-of-package labeling and increased transparency of nutritional information [1-3, 5, 6].

5. Discussion

This case study underscores three main findings. First, in this dataset, the rating is almost completely explained by linear combinations of available features. That suggests the rating either encodes a composite of nutritional attributes or that expert ratings closely track those attributes. Prior documentation notes that the rating "possibly" originates from Consumer Reports, and community sources emphasize it encapsulates nutrition and taste, but the near-determinism implies a mechanical or quasi-mechanical mapping to the included variables [13-16]. Second, sugars exert the largest negative influence on ratings, overshadowing all other nutrients, with calories and sodium contributing additional negative effects. This aligns with the policy and clinical emphasis on reducing free and added sugars—now disclosed explicitly as "Added Sugars" on labels—and curbing sodium intake for cardiometabolic health [1-4]. Third, fiber and, to a lesser extent, protein are positively valued, consistent with evidence on whole grains, fiber's satiety and metabolic benefits, and consumer acceptance of high-fiber cereals [11,12,16].

For consumers, a simple decision rule—minimize sugars, prefer high-fiber cereals, keep calories moderate, and ensure adequate protein—should align choices with higher ratings and better nutritional profiles. For manufacturers, reformulation that shifts sugar downward and fiber upward is likely to improve both healthfulness and perceived product quality; for policy makers, the near-deterministic rating behavior suggests that transparent nutrient disclosure and straightforward interpretive aids (e.g., warning labels or summary scores) can move markets and consumer perception in healthier directions.

In the broader context of consumer food analytics, several structural limitations warrant attention. First, nutritional profiles alone cannot fully capture consumer perception: taste, branding, cultural associations, and marketing cues often play decisive roles that are not systematically encoded in datasets. This creates an inherent gap between measurable nutrition facts and the holistic drivers of product ratings across populations. Second, reliance on static nutrition panels reflects a snapshot in time, yet formulations, serving sizes, and regulatory requirements evolve continuously; models trained on historical data may therefore lag behind real-world consumption patterns. Finally, the field still lacks widely accepted benchmarks that integrate nu-

tritional quality with behavioral outcomes, making it difficult to compare results across studies. Addressing these systemic challenges will require richer data that combines objective nutrient measures with subjective taste panels, marketing context, and longitudinal evidence of reformulation impacts.

6. Conclusion

This paper reviewed relevant literature in nutrition policy and explainable machine learning and conducted an empirical case study on breakfast cereals to quantify how nutrients relate to consumer ratings. Using linear regression, LASSO, and Random Forests with permutation importance, we found sugars to be the dominant negative driver of ratings, followed by calories and sodium, while fiber and protein modestly positively influence ratings. The linear model achieved near-perfect out-of-sample accuracy, suggesting the rating is strongly determined by the included features. Building on these findings, we produced a health-oriented recommendation that filters cereals to low sugar, high fiber, adequate protein, and moderate calories, then ranks by predicted rating; the resulting top 10 list demonstrates that healthier nutrient profiles can align with high perceived product quality.

Also, this study has limitations that include the dataset's modest size (≈77-80 cereals), potential measurement or transcription differences across mirrors, and a lack of access to the original rating rubric and consumer demographics, which limits inference about causal consumer preferences. Future work should validate on larger, contemporary datasets with explicitly labeled "added sugars," integrate front-of-pack schemes (e.g., warning labels, Nutri-Score) to compare with model attributions, and incorporate individual-level personalization (dietary needs, taste profiles). Finally, leveraging SHAP or counterfactual explanations on richer datasets can further enhance transparency for stakeholders while preserving accuracy. Overall, the study illustrates how interpretable modeling can translate nutrient data into actionable guidance for consumers, manufacturers, and policy makers.

References

[1] U.S. Food and Drug Administration. (2024-2025). Added sugars on the Nutrition Facts label. https://www.fda.gov

[2] U.S. Food and Drug Administration. (2016, May 27). Food labeling: Revision of the Nutrition and Supplement Facts labels (Final rule). Federal Register. https://www.federalregister.gov

[3] U.S. Food and Drug Administration. (2016). Final rules to update the Nutrition Facts label [PDF overview]. https://www.fda.gov

[4] AGES. (2025, April 10). WHO sugar recommendations. AGES - Nutrition & Food / Nutrition Recommendations. https://www.ages.at/en/human/nutrition-food/nutrition-recommendations/who-sugar-recommendations

[5] Fuentes, M., Carpentier, F. R. D., Corvalán, C., Reyes, M., & Taillie, L. S. (2023). Front-of-package nutrition labeling and its impact on food manufacturers' practices. Nutrients, 15(2), 377. https://doi.org/10.3390/nu15020377

[6] Martini, D., Menozzi, D., & Vicentini, A. (2022). Relationship between front-of-pack labeling and nutritional quality: A narrative review of available evidence. Nutrients, 14(5), 1091. https://doi.org/10.3390/nu14051091

[7] Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. arXiv. https://arxiv.org/abs/1801.01489

[8] Daneshvar, A., & Golalizadeh, M. (2023). Regression shrinkage and selection via least quantile shrinkage and selection operator. PLOS ONE, 18(2), e0266267. https://doi.org/10.1371/journal.pone.0266267

[9] Salman, H. A., et al. (2024). Random forest algorithm overview (H. A. Salman et al., Trans.). Babylonian Journal of Machine Learning, 2024(June), 69-79. https://doi.org/10.58496/BJML/2024/007

[10] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1705.07874

[11] Poutanen, K., Sozer, N., & Valle, G. D. (2020). Dietary fibre from whole grains and their benefits on metabolic health. Nutrients, 12(10), 3045. https://doi.org/10.3390/nu12103045

[12] McRae, M. P. (2016). Health benefits of dietary whole grains: An umbrella review of meta-analyses. Cureus, 8(7), e711. https://doi.org/10.7759/cureus.711

[13] Crawford, T. (2016). 80 cereals [Dataset]. Kaggle. https://www.kaggle.com/datasets/crawford/80-cereals

[14] R Core Team. (n.d.). cereal {liver}: 77 cereals, 16 variables; rating possibly from Consumer Reports [R documentation]. CRAN. https://search.r-project.org/CRAN/refmans/liver/html/cereal.html

[15] Kss7vs. (2020). Analysis of the 80 cereals dataset [Rpubs post]. Rpubs. https://rpubs.com/kss7vs/636466

[16] Hamoye HQ. (2021, March 12). Understanding the relationship between nutritional content and cereal ratings: A data-driven approach [Blog post]. Medium. https://hamoyehq.medium.com/understanding-the-relationship-between-nutritional-content-and-cereal-ratings-a-data-driven-6b8a0f5cbdfe