

Launch Date: August 03, 2026

Study Name: (synthetic data) Energy Drink Round 2 User Features

LINESOL

DATA ANALYSIS REPORT

Linesol, Inc. | hello@Linesol.online



SUMMARY

This report presents the findings of a market research study to assess consumer preferences and behaviours towards the specified Yes/No question: Would you buy this energy drink? The study aimed to identify key factors associated with the likelihood of positive responses to the question posed.

Using the Linesol mobile app, a group of 333 participants was asked to respond 'Yes' or 'No' to a question presented on a card, which included parameters such as Titles, Images, Descriptions, and Prices, providing additional context. Alongside each response, the participant's question card features and user characteristics were collected, resulting in a rich dataset that was analysed with advanced statistical methods to provide valuable insights.

Key Findings Include:

1. **Impact of User Characteristics:** The analysis examined how various user characteristics influence the likelihood of positive responses.
2. **Suggested Improvements:** Detailed diagnostic results to refine feature selection, improving model stability, interpretability, and predictive accuracy.

This analysis provides decision-ready evidence for the tested configuration and sample. These findings can inform strategy design and follow-up testing when combined with broader commercial, operational, or research context.



INTRODUCTION

OUR APPROACH

Understanding human behaviour and preference patterns is crucial for driving innovation. This study focused on several key question card features — Titles, Images, Descriptions, and Prices — each forming a distinct card configuration presented to participants. By analysing responses across these combinations, we identified trends and patterns that demonstrated consistent strength across the participant group.

Machine learning techniques were applied to evaluate the combined predictive relationships between question card attributes, user characteristics, and recorded responses. The objective was to determine which factors most strongly predicted a positive response.

User characteristics were an integral part to understand how demographic factors impact positive responses. Only users with traits that were specified in the study were permitted to participate, ensuring the data accurately reflects the target consumer base.

Participants used the Linesol mobile app, responding Yes/No to the posed question displayed via a question card with randomized question card features. They were required to spend at least 10 seconds considering the question presented, ensuring thoughtful engagement. Financial compensation (\$0.05) was provided upon submission of each response.

This method fosters a motivational framework for authentic responses, ensures participation from a diverse group, and yields reliable data for actionable insights.

ADDRESSING LIMITATIONS

- **Diverse Representation:** While efforts were made to ensure a diverse participant pool, achieving complete equality in demographic representation may not always be possible. Before the study begins, we provide monthly active user feature distribution data serving as a metric for identifying potential representation imbalances. During the study we restrict the inclusion of individuals whose features are disproportionately prevalent in the dataset effectively enhancing diversity for the study.
- **Sample Size and Machine Learning Model:** The accuracy of our insights depends on selecting an appropriate sample size and machine learning model. While we provide guidelines (e.g., a minimum of 200 participants and 20 per feature), the ideal number varies by study complexity and goals. To manage this trade-off, model performance is evaluated against observed response patterns, and study outcomes are monitored over time. These results inform scheduled updates to sampling strategies, demographic coverage, and model configuration, allowing future studies to be delivered with improved accuracy, broader representation, and faster turnaround.
- **Study Scope and Future Research:** This study focuses on specific features associated with the likelihood of positive responses, though it may not account for all contributing factors. The analysis includes detailed multicollinearity results to aid in refining subsequent studies, guiding future research to avoid correlated features, replacing them with more informative variables. By narrowing the focus to exclude less impactful elements and exploring new ones, future investigations can uncover additional key factors, paving the way for models that more effectively capture the complexities of human behavior and decision-making.

FEATURE IMPORTANCE

To perform our analysis, we modelled the dataset to predict a participant's response (Yes/No) based on their user features. Both a Ridge Regression and a Random Forest model were fitted to the dataset, and their performances were compared based on their accuracy scores, which represent the percentage of correct predictions out of the total number of predictions made. The better-performing model was selected, and its feature importance rankings were presented.

To enhance our analysis, we also calculated the linear correlation direction (positive/negative) of each variable with respect to the positive response rate to determine whether these features are associated with higher or lower likelihood of a positive response rather than just their importance scores. This was done using the Pearson correlation coefficient, which measures the linear correlation between two sets of data.

The question posed to participants was: Would you buy this energy drink?
Number of participants: 333.

Below is a table showing a list of the user features ranked in terms of their importance in predicting a positive response as modelled by the Ridge Regression model, which achieved an accuracy score of 74.4%.

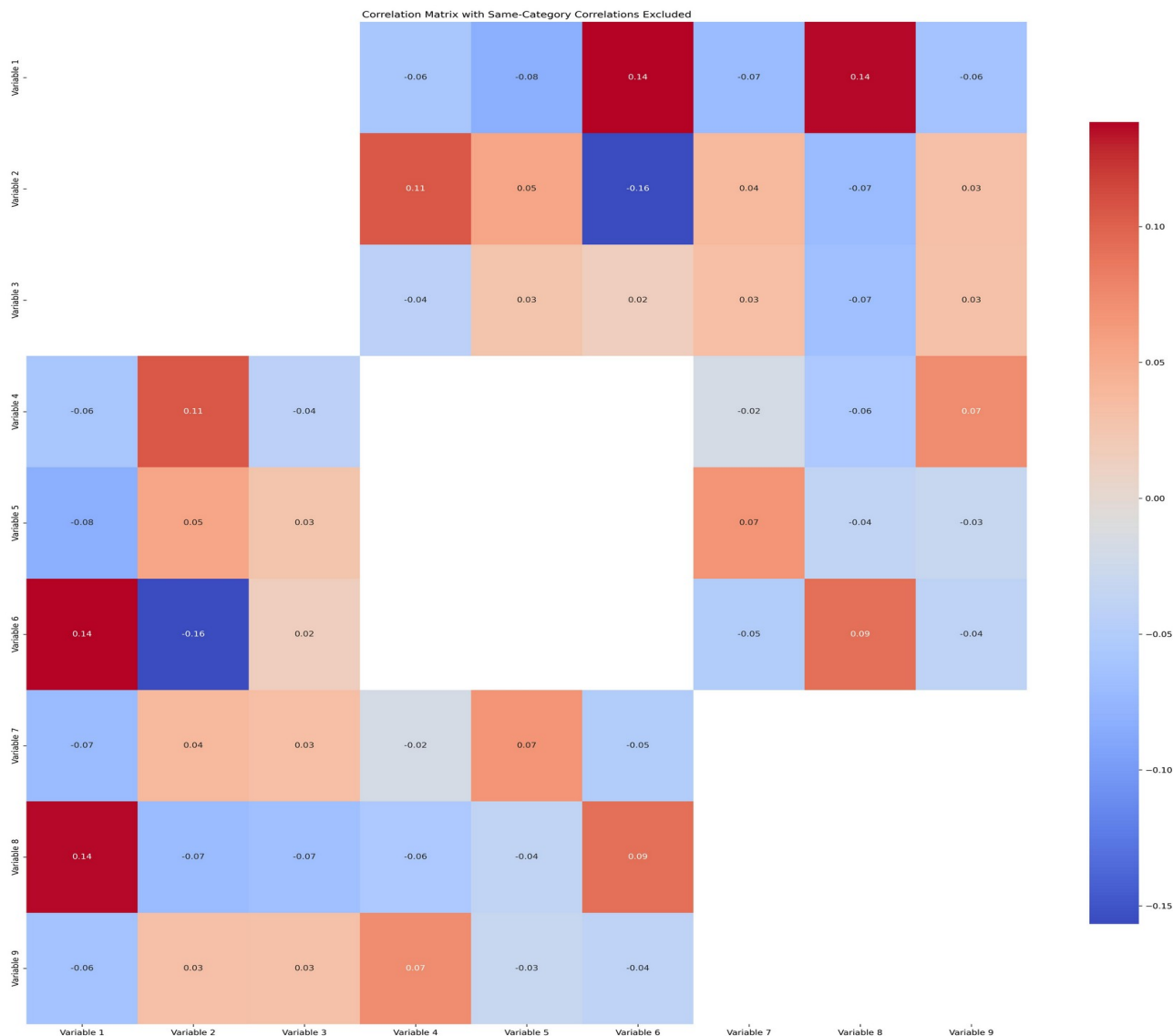
Feature	Feature Importance Score	Correlation Direction
Caffeinated drinks per day (width 2): 4-5	1.00	positive
Gym workout visits per week (width 3): 6-8	0.75	positive
Age years (width 10, start 18): 18-27	0.51	positive
Gym workout visits per week (width 3): 3-5	0.07	positive
Age years (width 10, start 18): 28-37	0.00	negative
Age years (width 10, start 18): 38-47	0.00	negative
Caffeinated drinks per day (width 2): 2-3	0.10	negative
Gym workout visits per week (width 3): 0-2	0.27	negative
Caffeinated drinks per day (width 2): 0-1	0.96	negative



SUGGESTED IMPROVEMENTS

To enhance the effectiveness of future studies, it is essential to refine feature selection by prioritizing variables with the most significant impact while also exploring new factors that could influence decision-making. A key challenge in this process is multicollinearity, where high correlations between predictor variables can obscure their individual effects, making it difficult to determine which features hold real predictive importance. A thorough assessment of multicollinearity, alongside feature importance evaluations, allows for the refinement of feature selection in the study. This approach helps mitigate issues such as overfitting and misinterpretation of results, leading to more reliable and actionable results.

To detect and quantify these relationships, a correlation matrix is a useful tool, as it displays both the strength and direction of correlations among the features. The visual representation of the correlation matrix below highlights the most significant correlations between variables across various categories, offering insight into potential multicollinearity issues. Placeholder variable names referenced in the matrix can be found in the appendix.



Below is a table listing the pairs of variables with the strongest correlation scores. A positive correlation value close to 1 indicates a strong direct relationship, where an increase in one variable is associated with an increase in another. Conversely, a negative correlation value close to -1 suggests a strong inverse relationship, where an increase in one variable corresponds with a decrease in another. A correlation value near 0 indicates no linear relationship between the variables.

Variable Pair	Correlation Score
Caffeinated drinks per day (width 2): 4-5 & Age years (width 10, start 18): 28-37	-0.157
Caffeinated drinks per day (width 2): 4-5 & Age years (width 10, start 18): 18-27	0.138
Gym workout visits per week (width 3): 3-5 & Age years (width 10, start 18): 18-27	0.137
Caffeinated drinks per day (width 2): 0-1 & Age years (width 10, start 18): 28-37	0.105
Gym workout visits per week (width 3): 3-5 & Caffeinated drinks per day (width 2): 4-5	0.0912

The correlation matrix and accompanying multicollinearity table include correlations across all features to provide a comprehensive diagnostic view. These encompass integrity-check correlations (expected to be minimal) and potential multicollinearity signals (where correlations may indicate overlapping predictors). The matrix is presented as a single integrated view to enable comparative analysis:

- Question card permutation balance: These correlations check that question card permutations are distributed evenly and independently across participants. Since each participant sees only one permutation and they're allocated evenly, these should be near zero. Material correlations here signal data encoding or allocation issues, not behavioral patterns.
- Question card permutation distribution across demographic base: These verify that question card exposure hasn't accidentally correlated with user demographics due to class imbalances or random allocation. They should also be low but may show slight variance from incidental associations, serving as a secondary integrity check to ensure demographic balance.
- User feature interdependencies: These are the main multicollinearity signals, where high correlations indicate overlapping demographics that may represent latent constructs (e.g., socioeconomic status bundled into age, income, and education). This is where composite variables can consolidate related predictors into a single, more precise feature.

Presenting all correlations together allows readers to benchmark user feature interdependencies against the baseline "noise" from allocation mechanics. If user feature correlations are notably larger than the integrity checks, it flags real overlaps that could dilute model precision and suggests opportunities for composite variable development.

Multicollinearity can obscure the contribution of correlated variables, leading to unstable feature importance estimates, reduced model accuracy, and less actionable insights (related diagnostics, including class distribution and imbalance, are reported in the technical details section). To mitigate these effects, appropriate machine learning models were selected. Ridge Regression applies regularization to stabilize coefficient estimates by penalizing large coefficients, while Random Forest models manage multicollinearity through inherent feature selection during tree splits, as only a subset of variables is considered at each split. Although these models improve robustness and provide a clearer understanding of which variables are most predictive of outcomes, they do not fully resolve the challenges posed by highly correlated predictors. Additional feature refinement is therefore required to ensure accurate and

interpretable results.

Refining feature selection is essential for addressing multicollinearity and improving model performance. This involves both removing non-influential features and replacing highly correlated predictors with more informative variables. By combining multicollinearity diagnostics with feature importance results, it is possible to identify redundant features for removal and select alternative predictors that capture unobserved or composite effects. This approach reduces noise, mitigates overfitting, and enhances model interpretability, leading to more reliable insights and a clearer understanding of the features holding real predictive importance.

Multiple Correspondence Analysis (MCA) for Composite Variable Development

Multiple Correspondence Analysis (MCA) is a dimensionality-reduction technique designed for categorical data. It identifies latent dimensions by capturing shared variance (inertia) across related categorical variables, allowing multiple correlated predictors to be consolidated into a single composite variable representing an underlying construct.

Because MCA requires analyst decisions—such as variable selection, dimensionality retention, and interpretation of resulting dimensions—the analysis is performed externally by the study creator rather than automatically within the Linesol pipeline.

Raw one-hot encoded feature data is available through the dashboard and API, enabling researchers to conduct their own MCA analysis and define composite variables derived from correlated predictors. Approved composite variables can be submitted to **hello@linesol.online** for review and, if validated, may be added as selectable user features for future studies.

Strategies for Improving Future Studies:

To enhance the quality of future studies and minimize the impact of multicollinearity, the following strategies are recommended:

- **Introduce New Features:** Develop and test new predictors that capture underlying psychological, behavioral, or contextual predictors of the outcome. This can help replace highly correlated variables and support the development of a more nuanced and informative model.
- **Replace Highly Correlated Features:** Consolidate clusters of highly correlated predictors into a single feature or composite variable grounded in theoretical or empirical justification.
- **Remove Redundant Features:** Identifying and eliminating variables that contribute to multicollinearity without enhancing predictive power improves model efficiency and interpretability.
- **Hierarchical Strategy:** Use a hierarchical approach by entering predictors in stages across multiple studies, beginning with variables that are theoretically critical to the outcome.
- **Expand Sample Size:** Increasing dataset size provides more stable coefficient estimates and reduces the influence of multicollinearity.

These strategies provide a framework for refining feature selection and improving model performance, ensuring future studies capture the complexities of human behaviour and decision-making more accurately.



CONCLUSION

This market research report presents the findings of a study that examined how different question card designs and user characteristics affect the likelihood of positive responses to the question: Would you buy this energy drink? Conducted with a sample of 333 participants, modelled by the Ridge Regression model, which achieved an accuracy score of 74.4%.

The analysis provides insights into consumer behavior across various question card variations and demographic groups, supporting strategic, data-driven decision-making. Key findings include:

- **Consumer Characteristics and Preferences:** User features were analysed to provide insights into how various consumer characteristics and preferences are associated with the likelihood of positive responses, helping refine marketing approaches and resource allocation.

The consumer characteristics most associated with higher positive response rates include:

1. Caffeinated drinks per day (width 2): 4-5
2. Gym workout visits per week (width 3): 6-8
3. Age years (width 10, start 18): 18-27

The consumer characteristics most associated with lower positive response rates include:

1. Caffeinated drinks per day (width 2): 0-1
2. Gym workout visits per week (width 3): 0-2
3. Caffeinated drinks per day (width 2): 2-3

These insights provide evidence for strategy decisions related to this study's configuration and sample, and can contribute to growth when combined with broader commercial data and follow up research.

TECHNICAL DETAILS

AI Disclosure (per ICC/ESOMAR 2025 Code): This report's analysis and insights are generated through a fully automated, machine learning-integrated data analysis pipeline with no per-study human review or intervention. System-level design and validation ensure methodological integrity. All outputs use anonymized, aggregated participant data only.

OVERVIEW OF THE DATASET

The dataset begins as an anonymous collection of Yes/No responses accompanied by the participant's user features. This dataset is then one-hot encoded and dummy variables are applied such that the dataset is comprised only of binary (True/False) values in each column, this prepares the dataset for machine learning analysis.

The study was defined by selecting a variety of categories and specifying the features in each category to assess their association with the likelihood of positive responses. The following categories and features were included in the study:

The question that was posed to participants was: Would you buy this energy drink?

Question Card Features: Each user is only able to participate once in the study so only sees one possible combination of question card features.

Category	Feature
Not Available	N/A

User Features: The user features that are specified in the study restrict the demographic of participants in the study, since only users with compatible features are permitted to participate.

Category	Feature
Age years (width 10, start 18)	38-47
Age years (width 10, start 18)	18-27
Age years (width 10, start 18)	28-37
Caffeinated drinks per day (width 2)	0-1
Caffeinated drinks per day (width 2)	4-5
Caffeinated drinks per day (width 2)	2-3
Gym workout visits per week (width 3)	0-2
Gym workout visits per week (width 3)	3-5
Gym workout visits per week (width 3)	6-8

Additionally, by specifying only one feature within a particular user feature category, the demographic profile of participants in the study is refined. However, the association of these types of features with response likelihood cannot be investigated since they must exist for all participants. Similarly, by specifying single feature question card categories,

additional context can be provided to the question being presented.

The following single-feature categories were specified:

Category	Feature
Image	 voltshift_clean_performance_ can.png
Title	VoltShift Energy
Price	2.99
Description	16 fl oz (473 ml), zero sugar energy drink with vitamins B6 & B12. Mango flavour.
Country of residence	USA

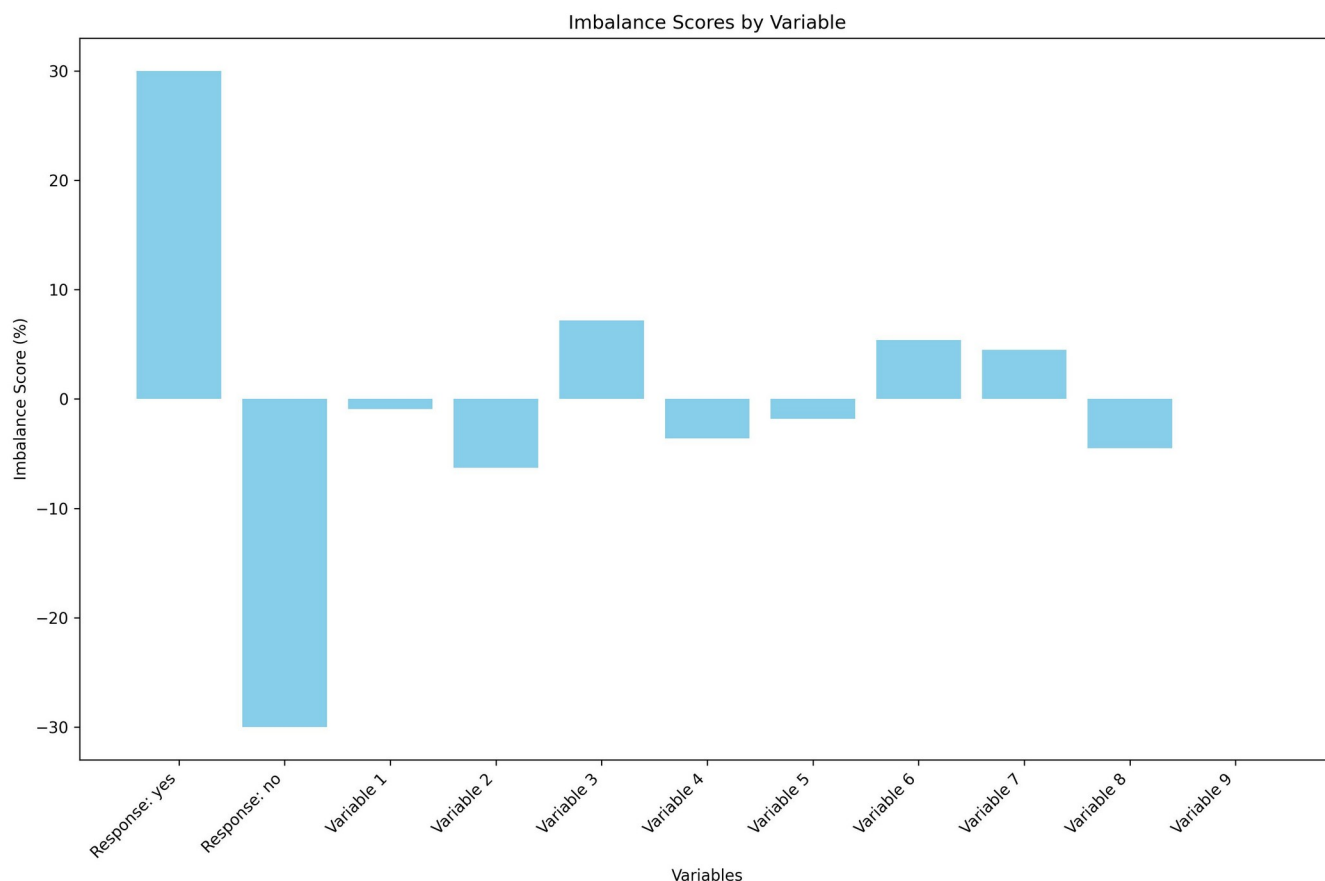
CLASS DISTRIBUTION AND IMBALANCE

Class distribution refers to how evenly participants are divided among the possible options within a category. For instance, if the category has four possible options, the ideal class distribution would have 25% of the participants in each group. Striving for an even class distribution is important because it helps ensure that the model does not become biased toward one class over others which would lead to skewed predictions and reduced overall performance.

When a particular option is either underrepresented or overrepresented, this results in class imbalance. We measure this imbalance by calculating the percentage by which its actual representation in the dataset is higher or lower than the ideal distribution.

Number of participants: 333

Below is a bar chart describing the class imbalance per variable described as a % deviation from the ideal distribution.



MACHINE LEARNING DETAILS AND OPTIMIZATION

When fitting machine learning models to the dataset, the goal is to build a predictive algorithm that accurately forecasts a participant's Yes/No response based on question card features (e.g., titles, images, descriptions, prices) and user characteristics.

For this analysis, we applied an ensemble approach using two complementary models:

- **Ridge Regression** — A regularized linear model that estimates feature impacts through coefficients, with L2 regularization to handle multicollinearity and stabilize results by penalizing large coefficients.
- **Random Forest** — A non-linear ensemble of decision trees that captures interactions and complex patterns by measuring impurity reduction (Gini) across trees, with built-in robustness to multicollinearity via random feature subsetting at each split.

To deliver reliable, generalizable feature importance scores:

- We systematically explored a broad range of model configurations to identify consistently high-performing settings, mitigating risks of overfitting (capturing noise) or underfitting (missing patterns).
- Performance was assessed via stratified 5-fold cross-validation (maintaining Yes/No class balance in each fold) for realistic accuracy estimates.
- Accuracy (proportion of correct predictions) served as the key metric.
- Feature importance scores aggregate results from strong configurations across both models, prioritizing consistency to minimize variability and enhance robustness. Final scores are normalized to a 0–1 scale for interpretability.

This fully automated pipeline balances linear stability (Ridge) with non-linear detection (Random Forest), producing

credible insights into factors driving positive responses. All processing uses anonymized, aggregated data only.



APPENDIX

PLACEHOLDER NAMES FOR VARIABLES:

When presenting the multicollinearity results, measures were taken to ensure the readability of the correlation matrix. This included replacing long feature names with placeholders. The original feature names can be found below:

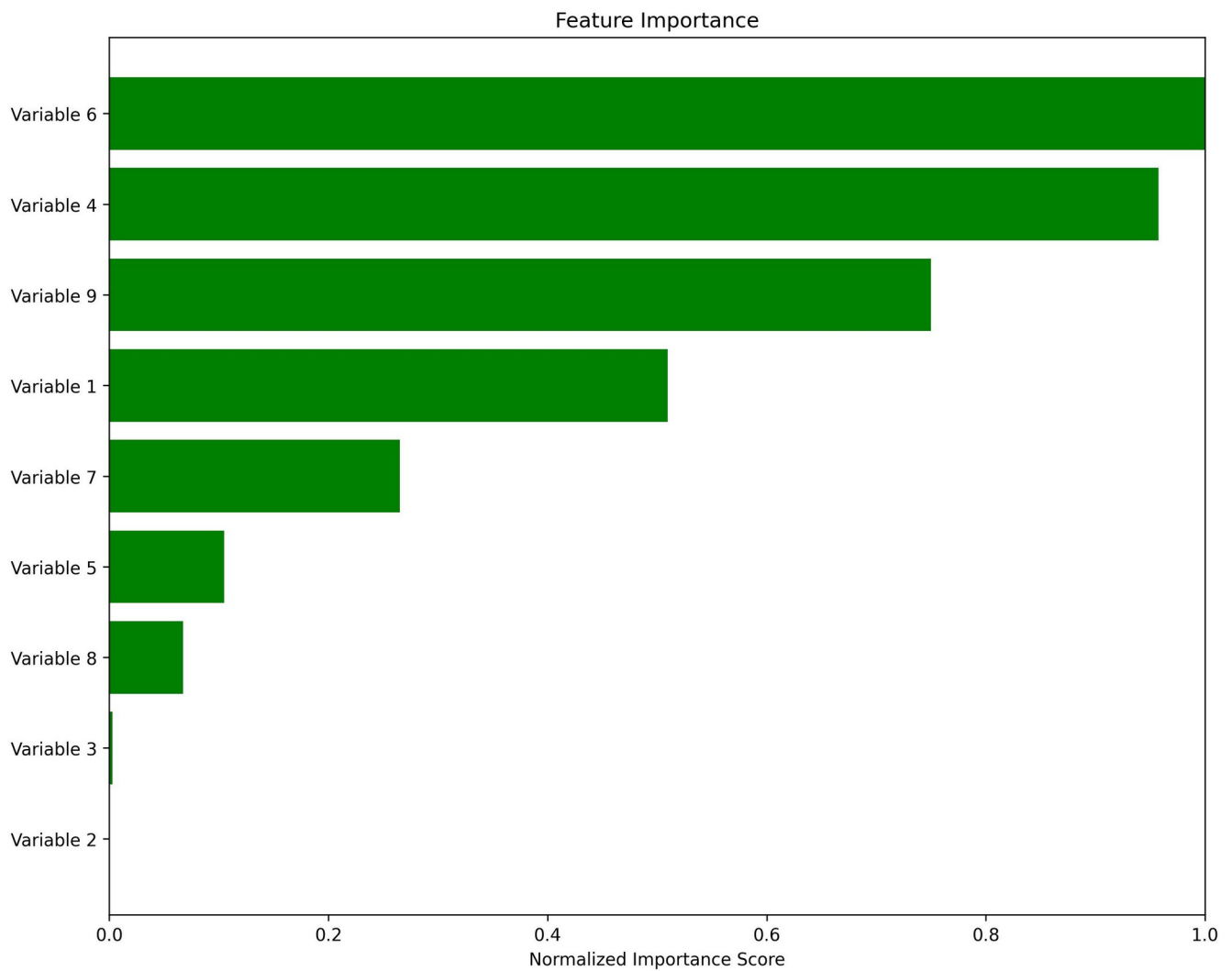
Original Variable Name	Placeholder Variable Name
Age years (width 10, start 18): 18-27	Variable 1
Age years (width 10, start 18): 28-37	Variable 2
Age years (width 10, start 18): 38-47	Variable 3
Caffeinated drinks per day (width 2): 0-1	Variable 4
Caffeinated drinks per day (width 2): 2-3	Variable 5
Caffeinated drinks per day (width 2): 4-5	Variable 6
Gym workout visits per week (width 3): 0-2	Variable 7
Gym workout visits per week (width 3): 3-5	Variable 8
Gym workout visits per week (width 3): 6-8	Variable 9

SECONDARY MODEL RESULTS:

This appendix includes the results of the secondary model for transparency and completeness, providing additional context to the primary analysis.

Ridge Regression Feature Importance Bar Chart:

The following bar chart represents the feature importance scores evaluated with the Ridge Regression model which attained an accuracy score of 74.4%. The feature importance scores are normalized between 0 and 1, and do not indicate the direction of the correlation:



Random Forest Feature Importance Bar Chart:

The following bar chart represents the feature importance scores evaluated with the Random Forest model which attained an accuracy score of 73.2%. The feature importance scores are normalized between 0 and 1, and do not indicate the direction of the correlation:

