GETTING TO KNOW THE *NUTRI*-SCORE

PREDICTING THE NUTRI-SCORES OF CONSUMER SHOPPING BASKETS USING MACHINE LEARNING

LESLEY HAERKENS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2026727

COMMITTEE

dr. Čule
dr. Collin

EXTERNAL SUPERVISORS

prof.dr.ir. Bronnenberg (Tilburg University, Marketing Department)
dr. Pachali (Tilburg University, Marketing Department)

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 15, 2024

WORD COUNT

8,787

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Dr. Boris Čule, for his valuable feedback, comments, and insights, which significantly contributed to the process of writing my thesis. I extend my thanks to my 'external' supervisors from the Marketing Department at Tilburg University, Prof.dr.ir. Bart Bronnenberg and Dr. Max Pachali, for their time, help, and valuable input. I am hopeful that this thesis marks the beginning of great future collaborations. Finally, I would like to thank my family, boyfriend, and friends for their support and motivation throughout my journey at Tilburg University so far, and especially during the thesis period.

GETTING TO KNOW THE *NUTRI*-SCORE

PREDICTING THE NUTRI-SCORES OF CONSUMER SHOPPING BASKETS USING MACHINE LEARNING

LESLEY HAERKENS

Abstract

The increasing prevalence of obesity and related chronic diseases underscores the need for tools guiding consumers toward healthier food choices. This research delves into the Nutri-Score, a label designed to evaluate the overall nutritional quality of food and beverages. By investigating factors that impact consumer nutritional decisions, including features related to shopping baskets and household demographics, valuable insights are gained to contribute to ongoing initiatives promoting healthier nutritional choices and addressing health concerns linked to dietary habits. To predict the average Nutri-Score of shopping baskets, various machine learning algorithms are employed. The study reveals modest improvements compared to a blind baseline model that predicts the overall average for each instance, with XGBoost Regressor emerging as the top-performing model, achieving a Mean Absolute Error of 0.419 and a Root Mean Square Error of 0.553. Nonetheless, the discussion highlights suboptimal performance, particularly for extreme values, indicating challenges in predicting shopping baskets with very low or high average Nutri-Scores. Another aspect of this study focuses on improving model interpretability by employing eXplainable AI techniques. These techniques shed light on features such as the influence of private label products on health-conscious shopping. Additionally, association rule mining uncovers patterns, such as the presence of seemingly healthy products like 'milk' and 'yogurt' in unhealthy baskets, underscoring the multifaceted nature of shopping choices. Lastly, the research suggests future avenues for exploration to refine the understanding of shopping basket healthiness. This aligns with the overarching goal of "*getting to know the* Nutri-*Score*", *i.e.*, obtaining a comprehensive understanding of consumers' nutritional decisions during grocery shopping.

CONTENTS

1 INTRODUCTION

With the rising prevalence of obesity globally, there is an urgent need for effective tools to assist consumers in selecting healthier food and beverage options (De Temmerman et al., 2021). Obesity rates have increased in European countries, including the Netherlands, where over half of adults were overweight in 2021, and approximately 14 percent had obesity (Bergmann et al., 2022; Tarrahi, 2023). Promoting healthier nutritional choices is a vital public health goal worldwide. Governments globally have implemented interventions to address nutrition-related health concerns, and the Nutri-Score, a labeling system on product packaging, emerges as a valuable tool to assess and communicate the overall nutritional quality of food and beverages (Karpyn et al., 2020; Ter Borg et al., 2021).

The Nutri-Score, with its five-colored categories, provides a quick and easy way for consumers to evaluate nutritional quality based on various factors (Clark et al., 2022; Julia & Hercberg, 2018). This thesis delves into understanding how consumers make nutritional choices during supermarket shopping, investigating factors such as promotions, demographics, and other elements that influence these choices. Addressing the critical issues of obesity and poor dietary choices, this research contributes valuable insights to ongoing efforts aimed at promoting healthier nutritional choices and combating health concerns related to dietary habits.

1.1 *Research Strategy & Research Questions*

Firstly, this thesis explores the capability of Machine Learning (**ML**) models to accurately predict the average Nutri-Score of shopping baskets. The study compares a blind baseline model, predicting the overall average for each instance, with more complex ML models, including Decision Tree (**DT**), Random Forest (**RF**), Gradient Boosting Regressor (**GBR**), and XGBoost Regressor (**XGBoost**). Model performance is evaluated using Mean Absolute Error (**MAE**), Mean Squared Error (**MSE**), and Root Mean Square Error (**RMSE**). Accordingly, this thesis considers the following research question:

RQ1. *To what extent can ML algorithms, specifically DT, RF, GBR and XGBoost, effectively predict the average Nutri-Score of consumer shopping baskets based on household panel features, with a focus on assessing predictive accuracy using MAE, MSE, and RMSE metrics?*

Secondly, eXplainable AI (**XAI**) techniques are a crucial component of this study. Tree-based models, such as RF and gradient boosted trees, while accurate, often act as black boxes, making it challenging for humans to understand their decision-making processes (S. Lundberg et al., 2020). Examining the predictive capabilities of ML models for average Nutri-Scores and identifying the factors influencing these predictions underscores the crucial role of XAI as a link between data-driven insights and decision-making. This aligns with the goal of understanding and promoting healthier nutritional choices during grocery shopping. Leveraging the study findings, XAI becomes valuable for marketing experts to gain insights into consumer behavior, enabling tailored strategies. Public health professionals can use XAI for evidence-based interventions, customizing programs to enhance public health. The thesis addresses the following research question:

RQ2. *How can XAI techniques, such as feature importance analysis, partial dependence plots, and SHAP values, be effectively applied to enhance the interpretability of ML models in explaining the factors influencing nutritional choices in shopping settings?*

Lastly, Association Rule Mining (**ARM**) is a powerful data mining technique that identifies patterns and correlations between co-purchased items. Involving the mining of frequent itemsets and association rules, the Apriori algorithm is commonly used, favored for its simplicity and efficiency with large datasets. In this study, ARM is used to reveal patterns within both healthy and unhealthy shopping baskets, addressing the following research question:

RQ3. *What insights can be derived from the association rules extracted through the Apriori algorithm, revealing patterns of co-occurrence and relationships between specific food and beverage items or categories in both healthy and unhealthy shopping baskets?*

In summary, this thesis investigates the predictive capabilities of ML in determining the average Nutri-Score of consumer shopping baskets. Through a comprehensive exploration, this study aims to uncover insights into the factors influencing nutritional choices. It seeks to bridge the gap between predictive accuracy, model interpretability using XAI techniques, and the intricate patterns within shopping baskets. Considering this context, this thesis considers the following overarching research question:

RQ. *To what extent can ML algorithms predict the average Nutri-Score of consumer shopping baskets based on household panel features, and what insights can be gained regarding the factors influencing nutritional choices, including the exploration of ARM to uncover patterns within shopping baskets?*

1.2 *Motivation & Relevance*

This study addresses the pressing issue of unhealthy dietary decisions by investigating factors influencing consumers' nutritional choices during grocery shopping. The emphasis on understanding elements such as promotions and demographic information, along with uncovering patterns in healthy and unhealthy shopping baskets, is crucial for promoting healthier nutritional choices. Unlike existing literature that primarily focuses on the Nutri-Score's effectiveness or predicting Nutri-Scores at the product level, this study uniquely employs ML to predict the average Nutri-Score of shopping baskets. Moreover, it demonstrates the utility of XAI methods in interpreting ML results and offering actionable insights. Additionally, the application of ARM extends beyond conventional sales-boosting purposes, providing deeper insights into consumers' decision-making regarding the nutritional quality of their shopping baskets.

1.3 *Ethics, Data & Technology Statement*

The data for this study is obtained from Aimark and GfK[1]. Tilburg University maintains ownership of the data, notified about its use in this thesis. Consumers voluntarily scan their purchases, and data collection is compensated. While this involves human participants, this thesis uses the end product created by GfK, not the data collection process. Nutri-Scores at the category level, publicly published by Clark et al. (2022), complement the GfK data.

The figures are owned by the author. Figure 12 is adapted from Clark et al. (2022) under their license. Additionally, R is used for data cleaning, while Exploratory Data Analysis (**EDA**) and modeling are carried out in Python. The list of packages and versions is provided in Appendix A. The code is not publicly available. The use of AI-powered technologies is restricted to ChatGPT for code debugging and rephrasing, without using other typesetting tools or services (OpenAI, 2023).

1 GfK is a global market research company, providing insights into consumer behavior and market trends, and Aimark facilitates collaboration between academics and businesses to generate and apply novel insights.

2 RELATED WORK

2.1 *Predicting Nutri-Scores*

Previous research on Nutri-Scores primarily delves into the effectiveness of the Nutri-Score as a Front-of-Pack labeling system. The Nutri-Score, on a scale from -15 (healthiest) to 40 (least healthy), is represented by a color-coded system using letters A (healthiest) to E (least healthy) and corresponding colors from dark green to red (De Temmerman et al., 2021). Originating from the National Epidemiology Research Institute in France, the Nutri-Score gained official status in France in 2017 and has since been adopted by several European countries, including Belgium, Spain, Germany, Switzerland, Luxembourg, and the Netherlands (Ter Borg et al., 2021). After a trial period of almost three years, the Nutri-Score gained official status in the Netherlands as of January 1st, 2024 (NOS, 2024). The Nutri-Score aims to guide consumers toward healthier choices, and for producers it serves as an incentive to improve nutritional content, such as reducing salt or sugar and increasing fiber (Ter Borg et al., 2021).

Despite skepticism about the Nutri-Score's effectiveness, especially regarding category differentiation, where a relatively healthy frozen pizza receives a green A-score and fatty smoked salmon receives a D-score (Van Benthem, 2022), academic research consistently supports its efficacy. This includes aiding consumers in ranking food and beverage products by nutritional quality (Dréano-Trécant et al., 2020), enhancing nutritional intake (Julia et al., 2021) and promoting the purchase of healthier options (Dubois et al., 2020; Fialon et al., 2020). Regarding Nutri-Score prediction, research focuses on using ML models, such as decision trees and gradient-boosting machines, to accurately predict Nutri-Scores for plant-based foods based on nutrient content, potentially enhancing label accuracy and aiding informed consumer decisions (Tachie et al., 2023).

The approach by Tachie et al. (2023) differs from the one in this study, which aims to predict the average Nutri-Score of shopping baskets by considering features related to shopping baskets and household demographics. The decision to incorporate demographic and basket-level features, rather than individual purchase data, aligns with the primary goal of understanding the factors influencing nutritional choices. While retail data is valuable for public health research, obtaining individual purchase data is often challenging due to privacy concerns and the absence of longitudinal data for individual customers (Egan et al., 2014). Opting for the strategy adopted in this study ensures broader relevance and extends the findings' utility beyond specific and often limited data contexts. Additionally, past research has explored socioeconomic and marketing factors influencing consumers' nutritional choices, with age impacting food preferences (Monterrosa et al., 2020), higher-income individuals favoring organic and locally sourced foods (Ogundijo et al., 2021), and effective marketing strategies promoting healthier, lesser-known, food choices (Glanz et al., 2012; Melovic et al., 2020).

Additionally, Tachie et al. (2023) distinguish between predicting the precise Nutri-Score (a numerical range from -15 to 40) and the Nutri-Score grade (ranging from A to E) displayed on product labels. The Nutri-Score assesses products based on seven key aspects, penalizing high levels of energy, saturated fat, sugars, and sodium, while rewarding favorable compositions of protein, fiber, and the proportion of fruits, vegetables, nuts, and certain oils. This study adopts a 1-5 scale transformation, aligning with Clark et al. (2022), who converted the A-E scale for nutritional impact comparison. This transformation assigns 1 to the best and 5 to the poorest nutritional composition, allowing linear mapping of Nutri-Score values to a 1-5 scale, supporting nutritional impact comparisons across diverse foods. As the target variable spans 1 to 5, this study approaches it as a supervised regression task (Bonaccorso, 2017).

In another study, Hossain et al. (2019) explored ML methods, including RF and XGBoost, to predict calorie-based indicators of food security, assessing household food consumption sufficiency in terms of calorie intake. The study focused on indicators such as "calorie poor" (a binary feature with values yes = 1) and actual calorie intake as a continuous variable (Hossain et al., 2019). They assessed the performance of ML methods against non-ML techniques such as logistic regression and ordinary least squares, considering various household-level features, including education, age, gender of the household head and the household size, to predict food security indicators (Hossain et al., 2019).

2.2 *eXplainable AI*

Tree-based ML methods such as RF and gradient boosted trees are widely employed for their accurate predictions based on input features (S. Lundberg et al., 2020). In contexts where interpretability is crucial, understanding how these models use input features becomes essential, despite their often black-box nature. Obtaining explanations for ML models is vital for enhancing reliability, enabling human assessment of causality, and aligning machine reasoning with existing mental models (Haag et al., 2022). With the increasing interest in XAI, various methods have emerged to explain ML model predictions, categorized by interpretability level, applicability, and the necessity of procedures to enhance interpretability (Haag et al., 2022). Post-hoc methods, such as feature permutation, explain already trained models, while intrinsic approaches such as linear regression are inherently interpretable (Adadi & Berrada, 2018; Haag et al., 2022).

This study places its emphasis on model-agnostic post-hoc methods for explaining feature attributions in predicting the average Nutri-Score. Two widely used methods are Local Interpretable Model-agnostic Explanations (**LIME**), which perturbs data and fits an interpretable model, and Shapley Additive Explanations (**SHAP**), using cooperative game theory to determine feature impacts (Haag et al., 2022; Tallón-Ballesteros & Chen, 2020). While LIME has faced criticism for stability issues, SHAP, which avoids randomness, is considered a more stable method (Haag et al., 2022). SHAP transcends merely demonstrating the impact of individual features, extending to reveal its contribution across all possible combinations of features that collectively determine the model's output (Tallón-Ballesteros & Chen, 2020). Additionally, this study incorporates Partial Dependence Plots (**PDPs**), a conventional method introduced by Friedman (2001), illustrating the relationship between the target response and specific input features while considering the average values of all other input features (Pedregosa et al., 2011).

XAI provides interpretable and human-understandable explanations for AI decisions, proving valuable across various sectors, including healthcare, finance (Öztoprak & Orman, 2022), legal and government decisions, and applications such as autonomous vehicles and image analysis (Schlegel et al., 2019; Tiwari, 2023). In marketing, XAI unveils the role of specific customer experience features and aids companies in decision-making (Rallis et al., 2022). In a broader context, XAI promotes transparency in complex models, addresses evidence-based policy needs and generates actionable insights (de Carvalho & da Silva, 2021). Therefore, XAI is useful in understanding consumers' nutritional choices during grocery shopping by making intricate models transparent. Interpretable models contribute to global health by suggesting opportunities and actionable insights for policymakers and marketing professionals. Without XAI, conveying insights beyond predictive accuracy, such as the MAE and related metrics, would be challenging.

2.3 *Association Rule Mining*

ARM is a data mining technique focused on discovering patterns and correlations in large transaction databases, such as frequent co-purchased items (R. Agrawal & Srikant, 1998). It identifies relationships by mining frequent itemsets and generating association rules, which evaluate subsets and calculate support and confidence for each rule (R. Agrawal & Srikant, 1998). ARM, widely applied in domains such as analyzing consumer purchase data (Hemalatha, 2020; Raorane & Kulkarni, 2011; Sani et al., 2022), proves valuable for business decision support, cross-selling, and customer relationship management (Shah et al., 2016). The Apriori algorithm, a popular method for mining frequent itemsets and association rules, is chosen for this study due to its simplicity and efficiency in handling substantial datasets, making it advantageous over other methods such as Frequent Pattern Growth (Han et al., 2011; Kumarr et al., 2019; Shah et al., 2016; Wicaksono et al., 2020).

Furthermore, within the context of promoting healthy decision-making, the exploration of using ARM to establish rules for (un)healthy shopping behavior remains limited in existing literature. For instance, Li (2009) used the Apriori algorithm to find dietary patterns focusing on health, but their approach involved extracting rules based on associations among nutrient information within products. One rule they found, for example, shows a close link between total fat content and magnesium for healthy products. Additionally, Sharma (2017) applied ARM to tackle rising obesity rates by revealing hidden relationships between health-related factors, emphasizing the benefits of regular physical activity. While related to health, ARM serves a different purpose in this study, where it uncovers patterns within healthy and unhealthy shopping baskets, offering insights into consumer preferences for nutritious choices.

3 METHODOLOGY

3.1 *Dataset Description*

3.1.1 *GfK Household Panel*

The raw data files, obtained from Aimark and GfK, comprises individual-level datasets from Dutch households, documenting consumer packaged goods purchases. Categorized into purchase, barcode, shopcode, and panelist sections, the purchase files include details such as purchase date and retailer, barcode files provide product information, shopcode files contain retailer details, and panelist files feature household demographics. The purchase data spans from 2012 to 2022, with varying sizes, totaling 149,461,450 rows when merged. Each row corresponds to a product purchased by a panelist, scanned on a specific day, from a particular store. The shopcode file involves 241 stores, categorized as discount, mid-range, service supermarkets, or other locations. This study encompasses 31,328 distinct panelists, varying in household size, age, income, and education levels of the household head. Appendix B provides the codebooks for these raw data files.

3.1.2 *Nutri-Scores*

Clark et al. (2022) developed an algorithm to estimate the environmental impact and nutritional quality of food and beverage products using publicly accessible data. Their approach leverages government-mandated regulations in the UK and Ireland, where ingredient listings must be arranged in descending order of abundance, and packaging should include the percentage composition of characterizing ingredients (*e.g.*, the proportion of beef in beef lasagna). The algorithm extrapolates information for 89.6% of ingredients lacking percentage composition data based on insights from the remaining 10.4%. Appendix C illustrates this approach. For nutritional assessment, Clark et al. (2022) employed the Nutri-Score, relying on provided information for energy, (saturated) fat, sugars, salt/sodium, carbohydrates, protein, fiber, and the proportion of fruits, vegetables, nuts, and certain oils. In cases of missing data, they estimated composition using ingredient proportions and nutrient data across 52 categories. As indicated previously, to facilitate comparisons, Clark et al. (2022) standardized the A-E scale from 1 to 5, allowing for averaging nutritional impacts across retail categories despite conversion threshold variations. Appendix D provides an example calculation.

3.1.3 *Linking Purchases to Nutri-Scores*

Products were systematically categorized into departments, aisles, and shelves to align with the common practice of grouping items with similar Nutri-Scores together in supermarket aisles, ensuring the nutritional representation of each aisle corresponding to the products it contains (Clark et al., 2022). For brevity, these organizational tiers are referred to as 'categories', with a list provided in Appendix E. Introducing a new category, '50% Beef and 50% Pork', this study calculates the average Nutri-Scores by combining those of 'Beef and Lamb' and 'Meat' to offer a more precise estimation for products containing a mix of beef and pork. Aligning Nutri-Scores with individual household purchases in the GfK panel is crucial for the analysis, accomplished by matching GfK barcode file categories with those established by Clark et al. (2022). Over 95% of the unmatched barcodes (42,739 out of 450,482) are alcoholic beverages, omitted due to the absence of Nutri-Score estimates by Clark et al. (2022). Non-food and non-beverage products lacking Nutri-Score information are also excluded from the analysis.

3.2 *Data Cleaning & Exploratory Data Analysis*

3.2.1 *Shopping Baskets*

This thesis focuses on shopping baskets, aggregating individual purchases made by a single panelist on a specific day at a particular store[2]. The study centers on the retail sector, specifically 'discount supermarkets', 'mid-range supermarkets', and 'service supermarkets'. Table 1 offers an overview of the variables used in this study. The variables are derived as follows: (i) 'Nutri-Score' is calculated by averaging Nutri-Scores across all items in the basket[3] , (ii) 'Items per Basket' is determined by summing the items within the basket, (iii) 'Promotion Ratio' and 'Private Label Ratio' are computed as ratios of items on promotion or private label, respectively, and (iv) the remaining categorical variables are directly extracted from individual purchases since they remain constant when aggregating to shopping baskets.

---

2 In instances where consumers make multiple shopping trips on the same day at the same store using the same purchase method, their purchases are merged into a single basket.

3 In this context, averaging the Nutri-Score entails calculating the mean of individual product scores in the basket, without considering their weights. The exclusion of weights prevents multi-packs from disproportionately influencing the overall average score of a basket.

Table 1: Overview of Variables

Variable (Type)	Description
Nutri-Score (num)	The Nutri-Score of the basket
Items per Basket (num)	The number of items per basket
Promotion Ratio (num)	The ratio of items on promotion per basket
Private Label Ratio (num)	The ratio of private label items per basket
Purchase Method (cat)	The channel where the purchase was made
Store Type (cat)	The retail store type
Household Size (cat)	The size of the panelists' household
Income Group (cat)	The panelists' income group
Age Group (cat)	The panelists' age group
Social Group (cat)	The panelists' social (or educational) group
Year (cat)	The year in which the purchase took place

Notes. 'num' indicates 'numerical'; 'cat' indicates 'categorical'

3.2.2 *Handling Missing Values*

Table 2 provides an overview of where data is missing across different variables in the dataset. Dealing with missing values commonly involves: (i) dropping the feature with missing values, (ii) removing instances with missing values, (iii) imputing missing values with inferred data, and (iv) estimating missing values using data-driven methods (Aggarwal, 2015).

Table 2: Missing Values

Variable	Missing Values
Purchase Method	7,806,462
Social Group	99,960
Income Group	54,476
Age Group	54,476
Household Size	54,476
Private Label	20,175

Notably, 'Purchase Method' contains a lot of missing values, and Figure 1 highlights that the majority of purchases are made in physical stores ('winkelaankoop'), with very few using online delivery ('online-bezorgen') or pickup ('online-ophalen') methods. The large imbalance in classes and the substantial amount of missing data could introduce noise in the model's predictions (Radwan, 2017). Therefore, the feature 'Purchase Method' will be removed.

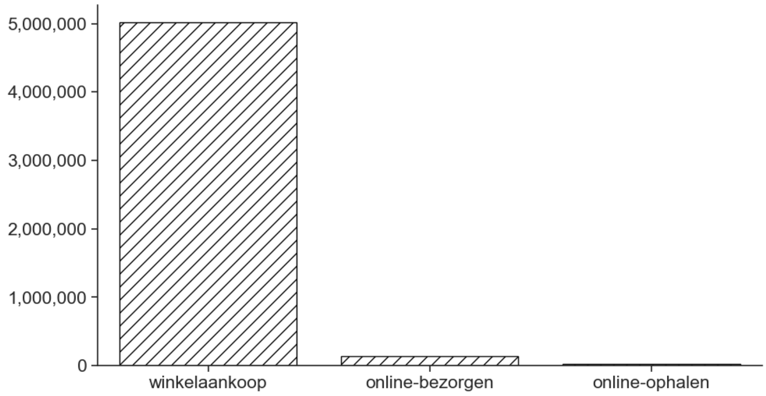


Figure 1: Distribution of Purchase Method (Excluding Missing Values)

Furthermore, 'Social Group' has the second-highest number of missing values, posing challenges for imputation due to limited methods for categorical data. Given the absence of this feature in only around 0.6% of shopping baskets, instances with missing 'Social Group' values are excluded, simultaneously leading to the removal of corresponding missing values in 'Income Group', 'Age Group', and 'Household Size'. These instances, marked by multiple missing features, contribute less information for predicting the average Nutri-Score and are thus eliminated. An even smaller fraction, about 0.1%, of shopping baskets lacks completeness in terms of the ratio of products in the basket that are private label. The instances with missing values for this variable are also removed from the dataset.

Additionally, this study assesses the impact of removing missing values in 'Social Group' (and consequently, in the remaining socioeconomic variables) and 'Private Label' on the distributions of the variables included in this study. Appendix F contains the distributions of the features before and after removing the missing values in 'Social Group' and 'Private Label'. Visual inspection shows a remarkable similarity between the distributions before and after the removal of missing values, supporting the removal in 'Social Group' and 'Private Label' due to the slightly smaller dataset effectively representing the larger one.

3.2.3 *Handling Duplicates & Outliers*

In this study, the presence of duplicates is limited to the panelist file, where each panelist's record appeared for every month they participated. Nevertheless, since each panelist is assigned a unique identifier once during their lifetime, it is permissible to eliminate any duplicate entries from this file. This process ultimately yields the previously mentioned count of 31,328 distinct panelists.

Furthermore, boxplots displaying the presence of outliers in the numerical variables 'Nutri-Score', 'Items per Basket', 'Promotion Ratio', and 'Private Label' can be found in Appendix G. The boxplot for 'Items per Basket' suggests that many shopping baskets contain a relatively low number of items, with only a few exceptions having more items. This observation aligns with typical supermarket transaction patterns where a majority of shopping baskets consist of a small number of items (Martin et al., 2020). Regarding the boxplot of 'Promotion Ratio', the outliers indicate shopping baskets where a high proportion of items are on promotion, potentially corresponding to smaller shopping baskets with few items that are on promotion. Given that these outliers are conceptually valid, they will not be treated as anomalies but rather considered as relevant insights for the analysis.

3.3 *Preprocessing*

3.3.1 *Encoding, Scaling & Sampling*

In ML, transforming categorical variables into numerical representations is a crucial preprocessing step to effectively use them in modeling techniques. Categorical data can be classified as either nominal or ordinal (Verma, 2021). Nominal data includes variables with names but no inherent numerical values. Ordinal data, on the other hand, involves categories with an inherent order or scale (Verma, 2021). Accordingly, 'Store Type' is considered nominal and the variables 'Household Size', 'Income Group', 'Age Group', 'Social Group' and 'Year' are classified as ordinal due to their inherent order[4]. As a result, one-hot encoding is applied to 'Store Type', while label encoding is used for the ordinal variables. Given the normal distribution of the target variable, the average Nutri-Score, and sufficient coverage in categorical variables, no additional preprocessing steps, such as scaling, undersampling, or oversampling are considered necessary.

[4] *E.g.*, in the case of 'Household Size', the order is defined as 1 person < 2 persons < 3 persons, and so forth. Similarly, for 'Income Group', the order is structured as 'below 700 euros' < '700-900 euros' < '900-1100 euros', and so on.

3.3.2 *Feature Importance*

To gain an initial understanding of the importance of the features used in this study, the default *feature_importances_* scikit function of DT was applied. The importance of a feature, or Gini importance, is calculated by measuring how much it contributes to reducing the overall criterion (impurity) in a normalized way (Pedregosa et al., 2011). Figure 2 displays the relative importance scores for each feature, with 'Items per Basket' holding the highest importance, signifying its significant role in Nutri-Score predictions. 'Year' and 'Income Group' also make substantial contributions. Notably, due to one-hot encoding for store types, the model omits one level ('Discount Supermarkets') to avoid the dummy variable trap[5]. As a result, only the importance of the other two store types can be inferred.

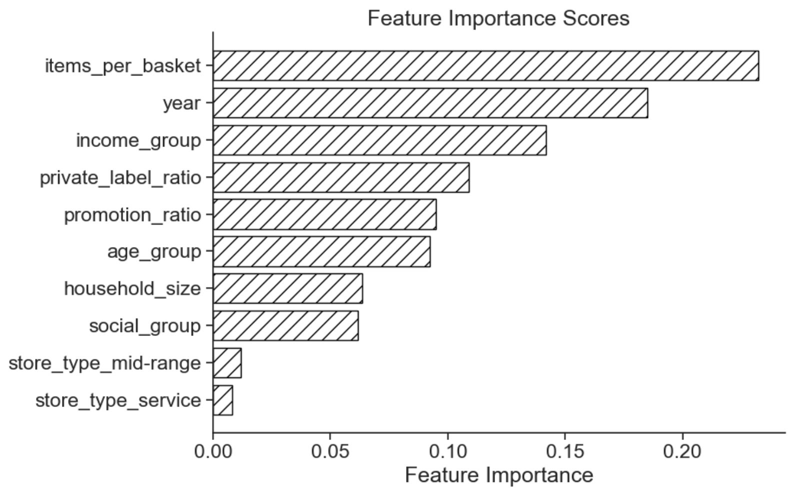


Figure 2: Feature Importance Scores

The significance of 'Income Group' and 'Items per Basket' as important features is not unexpected, as prior studies have highlighted their impact on consumer behavior and food choices. Puddephatt et al. (2020) demonstrate that income is a key influencer of eating behaviors, with higher income often facilitating access to a variety of food choices. Additionally, Prasad et al. (2008) show that households with higher incomes tend to be more health-conscious. Moreover, while not directly tied to the healthiness of shopping baskets, extensive research has explored the dynamics of basket sizes, examining factors such as store variety, basket profitability,

5 The dummy variable trap occurs in regression analysis with one-hot encoding, where the creation of k dummy variables for a categorical variable results in multicollinearity. To prevent this issue, it is recommended to produce k-1 dummy variables by excluding one category (Pramoditha, 2023).

shopping trip types, and even consumer-related variables such as consumer embarrassment (Martin et al., 2020; Nichols et al., 2014; Walters & Jamil, 2003). This highlights that the number of items in a basket is not a random number but plays a multifaceted role in various contexts, shedding light on its importance in consumer decision-making.

The unexpected prominence of 'Year' as a significant factor needs exploration, as it differs from the other variables directly linked to shopping basket composition or household characteristics. Figure 3 shows the average Nutri-Score of shopping baskets over time, indicating a subtle decrease with peaks and lows in December and January, possibly due to holiday indulgence and New Year's resolutions. According to Kamiński et al. (2021), December sees common purchases of goods such as processed fish, food fats and wine, while January experiences a decrease in sales indexes for most consumer goods and increased interest in healthy lifestyle topics based on Google searches. To address December and January seasonality, two extra features are introduced, taking the value of 1 for purchases in those months and 0 otherwise. The 'Year' feature remains to capture subtle, long-term variations without unnecessary temporal complexity.

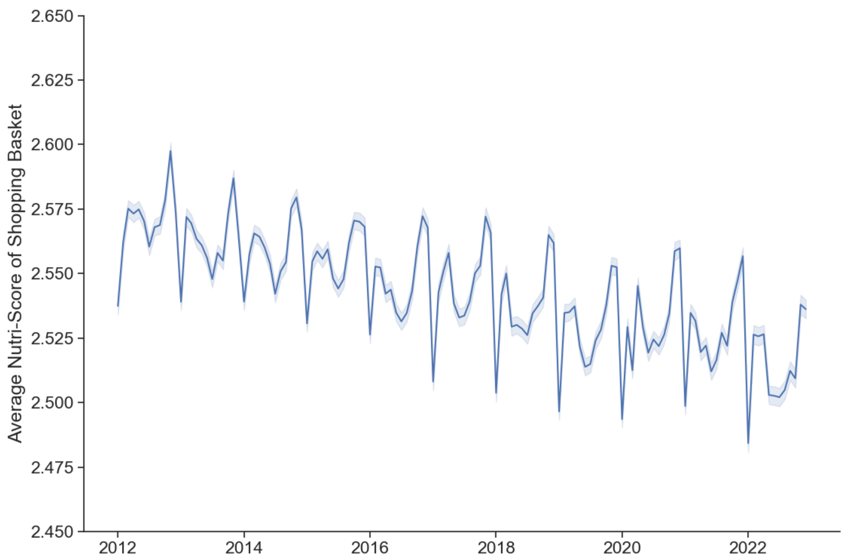


Figure 3: Average Nutri-Score Over Time

3.3.3 *Association Rule Mining*

To prepare the dataset for the Apriori algorithm, additional preprocessing involves adding a basket ID to the purchase-level dataset, creating an individual purchase dataset with two new columns for the Basket ID and the average Nutri-Score. To analyze purchase patterns in relatively healthy and unhealthy shopping baskets, the dataset is split into two subsets: one with average Nutri-Scores less than or equal to 2 and the other with scores greater than or equal to 3.

The decision to split the dataset into relatively healthy baskets (average Nutri-Score ≤ 2) and relatively unhealthy baskets (average Nutri-Score ≥ 3) is rooted in the normal distribution of the average Nutri-Score. With the majority of baskets exhibiting an average Nutri-Score around 2.5, this split focuses on discerning association rules in the tails of the distribution, excluding those closer to the mean. This approach ensures that the association rule exploration is tailored to the extremes of the average Nutri-Score distribution, providing insights into factors influencing both healthier and less healthy shopping choices.

3.4 *Algorithms & Hyperparameter Tuning*

3.4.1 *Predicting Nutri-Scores*

Blind Baseline
The blind baseline model used in this study adopts a straightforward approach by predicting the average Nutri-Score value observed in the training dataset, which is equal to 2.544, for all instances in the test dataset. This simplistic model assumes that the mean value reasonably represents the target variable and serves as a basic benchmark for evaluating the performance of more complex models.

Decision Tree
A DT is a hierarchical structure that recursively divides the data using selected features and split points, determined by criteria such as information gain (Tachie et al., 2023). The process continues until a stopping criterion, such as maximum tree depth or minimum samples per leaf, is met. The final prediction is made by traversing the tree from the root node to a leaf node based on input feature values (Agarwal, 2013; Tachie et al., 2023).

In ML algorithms, tuning hyperparameters is one of the important aspects in building efficient models. One of the most straightforward methods to determine the optimal set of hyperparameters is to train the dataset using every possible combination. This method is commonly known as grid search and offers a highly reliable way to identify the best hyperparameters (T. Agrawal, 2020). The dataset is divided into a training set (70%) and a test set (30%). To ensure the reliability of the results and mitigate computational expenses, a representative sample of the training set, comprising approximately 10,000 rows or 0.1% of the training data, is used. The distributions of the selected sample, as shown in Appendix

H, closely resemble those of the original training set. Additionally, 5-fold cross-validation was employed for the hyperparameter tuning process.

The hyperparameters explored for tuning the DT include 'max_features' (*i.e.*, the number of features to consider for the best split), 'max_depth' (*i.e.*, the maximum depth of the tree), 'min_samples_split' (*i.e.*, the minimum number of samples required to split an internal node), 'min_samples_leaf' (*i.e.*, the minimum number of samples required to be a leaf node), and 'splitter' (*i.e.*, the strategy used to choose the split at each node during the tree-building process). The candidate values are displayed in Table 3. In optimizing each ML model, the tuning process relies on the MAE as the scoring metric. The selection of hyperparameter values is informed by both reasoned choices and, where applicable, insights gained from literature.

Table 3: Hyperparameter Tuning - DT

Hyperparameter	Candidate Values
max_features	auto, sqrt
max_depth	None, 5, 10, 15, 20
min_samples_split	2, 3, 4, 5, 6, 7, 8, 9, 10
min_samples_leaf	1, 2, 3, 4
splitter	best, random

Random Forest

RF is an ensemble method that enhances predictive accuracy by combining multiple decision trees. It effectively handles diverse data types and demonstrates robust performance on large datasets. Random Forest builds individual decision trees on bootstrapped samples, incorporating random feature subsets for each split, and the final prediction is determined by averaging the predictions of these trees (Khan et al., 2022; Tachie et al., 2023). The hyperparameters explored for tuning the RF model include 'n_estimators' (*i.e.*, the number of trees in the forest), 'max_features', 'max_depth', 'min_samples_split', and 'min_samples_leaf'. The candidate values are displayed in Table 4.

Table 4: Hyperparameter Tuning - RF

Hyperparameter	Candidate Values
n_estimators	50, 100, 150, 200, 250
max_features	auto, sqrt
max_depth	None, 5, 10, 15, 20
min_samples_split	2, 3, 4, 5, 6, 7, 8, 9, 10
min_samples_leaf	1, 2, 3, 4

Gradient Boosting Regressor
GBR in scikit-learn is a versatile implementation of the gradient boosting algorithm. In gradient boosting, an ensemble of decision trees is built sequentially, with each tree aiming to correct the errors of the previous one. The model's predictions are the cumulative sum of the individual tree predictions, providing a powerful approach for regression tasks. The hyperparameters explored for tuning the GBR model include 'n_estimators', 'learning_rate' (*i.e.*, the step size for updating the model's parameters during training, impacting algorithm convergence and optimization), 'max_features', 'max_depth', 'min_samples_split', and 'min_samples_leaf'. The candidate values are displayed in Table 5.

Table 5: Hyperparameter Tuning - GBR

Hyperparameter	Candidate Values
n_estimators	50, 100, 150, 200, 250
learning_rate	0.01, 0.05, 0.1
max_depth	None, 10, 20
min_samples_split	2, 4, 8, 10
min_samples_leaf	2, 4, 8, 10

XGBoost Regressor
XGBoost is a specialized and highly optimized implementation of the gradient boosting algorithm. XGBoost adheres to the core principles of gradient boosting but introduces several improvements, such as regularization terms, to enhance both efficiency and performance. Notably, XGBoost is recognized for its speed, scalability, and superior performance, making it a popular choice in ML and a broad range of applications (Hossain et al., 2019). In essence, XGBoost can be considered a specific and advanced implementation of the gradient boosting algorithm. The hyperparameters explored for tuning the XGBoost model include 'n_estimators', 'learning_rate', 'max_depth', and 'min_samples_split'. The candidate values are displayed in Table 6.

Table 6: Hyperparameter Tuning - XGBoost

Hyperparameter	Candidate Values
n_estimators	50, 100, 150, 200
learning_rate	0.01, 0.05, 0.1
max_depth	3, 4, 5
min_samples_split	1, 2, 3

3.4.2 *Association Rule Mining: Apriori*

Apriori, an algorithm for mining frequent itemsets and association rules in transaction databases, operates through iterative candidate itemset generation and pruning based on a minimum support threshold (R. Agrawal & Srikant, 1998). The key insight behind the Apriori algorithm is the Apriori principle, which states that any subset of a frequent itemset must also be frequent. This allows the algorithm to avoid generating and counting the support of all possible itemsets, which would be computationally infeasible for large databases (R. Agrawal & Srikant, 1998). The algorithm comprises two main phases: (i) the frequent itemset generation phase, where it scans the database, counts item support, and iteratively generates candidate itemsets, and (ii) the rule generation phase, where all possible association rules are generated from frequent itemsets, producing rules of the form $X \rightarrow Y$ to signify strong relationships between items in X and Y (R. Agrawal & Srikant, 1998). The pseudocode outlining this process is provided below (Shah et al., 2016).

Pseudocode Apriori Algorithm

C_k: Candidate Itemset of size k
L_k: Frequent Itemset of size k
$L_1 = \{\text{Frequent Items}\}$
for $k = 1; L_k \neq \emptyset; k++$ **do**
  C_{k+1}: Candidates generated from L_k
  for each transaction t in database **do**
    Increment the count of all candidates in C_{k+1} that are contained in t
  end for
  L_{k+1}: Candidates in C_{k+1} with min_support
end for
return $\bigcup_k L_k$

3.5 *Evaluation Metrics*

3.5.1 *Predicting Nutri-Scores: MAE, MSE & RMSE*

The evaluation of the predictive performance of ML models relies on several key metrics, including the MAE, MSE, and RMSE. MAE measures the average absolute differences between predicted and actual values, while MSE computes the average of squared differences, giving higher weight to larger errors. Furthermore, RMSE is calculated by taking the square root of the average of squared differences between predicted ($\hat{y}_i$) and actual (y_i) values. The formula for RMSE is denoted as:

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

Here, n is the number of observations in the dataset.

3.5.2 *Association Rule Mining: Support, Confidence & Lift*

The strength of the relationship is measured by the support and confidence of the rule, which are defined in terms of the frequency of occurrence of the itemsets in the database (R. Agrawal & Srikant, 1998). The support of an association rule is the proportion of transactions that contain both the antecedent (X) and consequent (Y) of the rule (R. Agrawal & Srikant, 1998). This formula is mathematically expressed as:

$$\text{Support}(X \to Y) = P(X \cap Y) \tag{2}$$

The confidence is the proportion of transactions that contain the consequent (Y), given that they also contain the antecedent (X) (R. Agrawal & Srikant, 1998). This formula is mathematically formulated as:

$$\text{Confidence}(X \to Y) = P(Y|X) \tag{3}$$

In addition, Lift measures the likelihood of an itemset appearing together compared to their individual probabilities of occurrence. It indicates whether the occurrence of one item affects the likelihood of the occurrence of another item in the same transaction. Higher lift values suggest stronger associations between items. This formula is mathematically expressed as:

$$\text{Lift}(X \to Y) = \frac{P(X \cap Y)}{P(X) \cdot P(Y)} \tag{4}$$

3.6 *Workflow*

Figure 4 illustrates the workflow used in this study, emphasizing the integration of ML, XAI, and ARM. The process starts with data collection, preparation, and cleaning. The ML models provide predictive insights, while XAI methods aid in comprehending these predictions. Simultaneously, the Apriori algorithm uncovers association rules, unveiling patterns and relationships in the data. This cohesive approach enables exploration of both predictive patterns and understandable associations within shopping basket data.

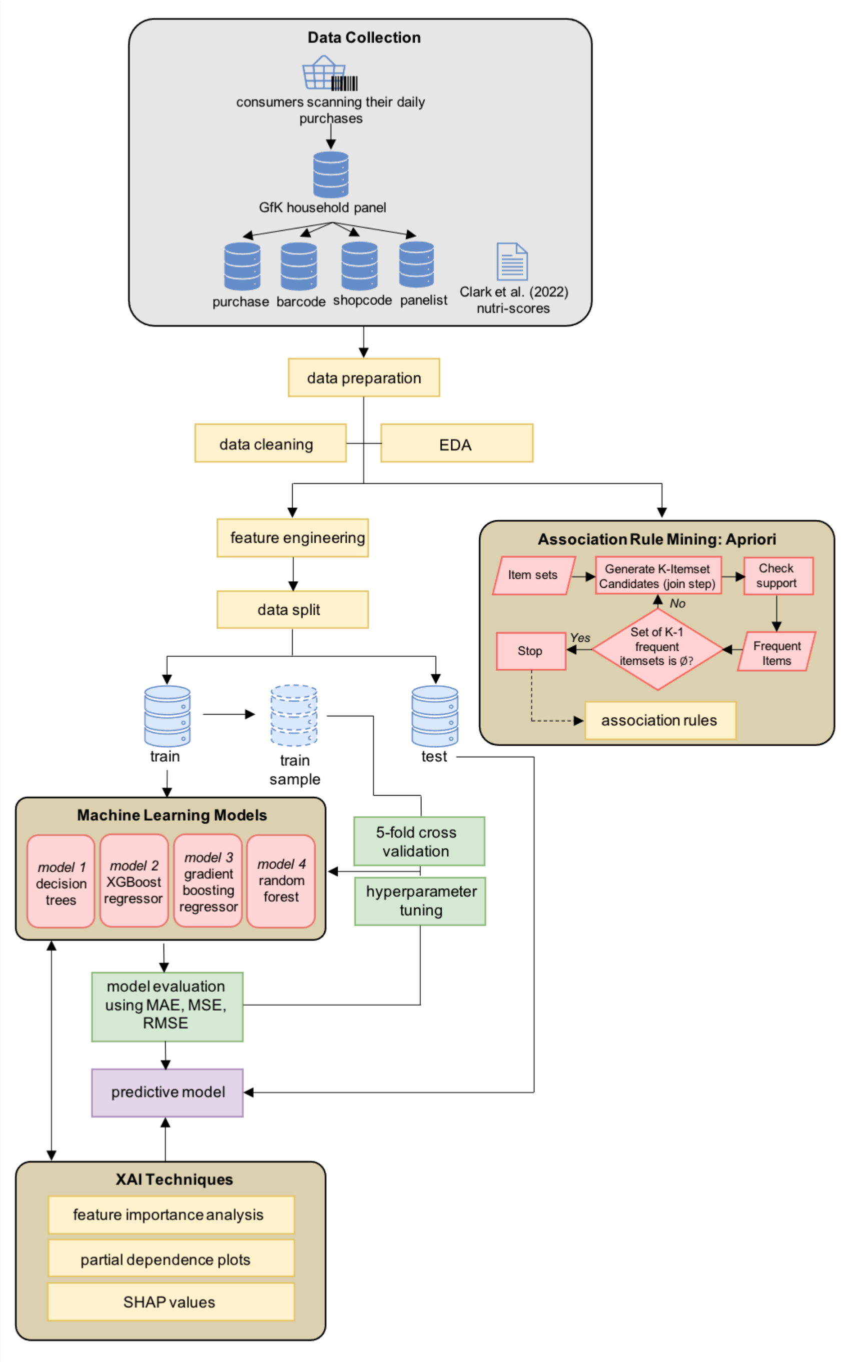


Figure 4: Workflow

4 RESULTS

4.1 *Predicting Nutri-Scores*

The Mean Baseline model, which serves as a simple benchmark by predicting the overall average for each instance, demonstrates a MAE of 0.431, MSE of 0.320, and RMSE of 0.566. The performance of this blind baseline model is compared against the performance of more advanced models. The latter models are tuned, and the best sets of hyperparameters are shown in Table 7.

Table 7: Hyperparameter Tuning - Optimal Values

Model	Hyperparameter and Optimal Value
Decision Tree	*splitter* = 'best', *max_features* = 'auto', *max_depth* = 5, *min_samples_split* = 6, *min_samples_leaf* = 2
XGBoost Regressor	*n_estimators* = 150, *learning_rate* = 0.1, *max_depth* = 3, *min_child_weight* = 3
GB Regressor	*n_estimators* = 100, *learning_rate* = 0.01, *max_depth* = 10, *min_samples_split* = 2, *min_samples_leaf* = 10
Random Forest	*n_estimators* = 200, *max_features* = 'auto', *max_depth* = 10, *min_samples_split* = 10, *min_samples_leaf* = 1

Comparatively, the more sophisticated models, including DT, XGBoost, RF, and GBR, demonstrate slightly improved performance across all metrics, as detailed in Table 8. However, RF, showing the most preferable results across all metrics comes at the cost of significantly longer execution times, highlighting the trade-off between predictive accuracy and computational efficiency. Balancing both factors, XGBoost emerges as the best-performing model.

Table 8: Results - Test Set

Model	MAE	MSE	RMSE	Execution Time
Mean Baseline	0.431	0.320	0.566	< 1 min
Decision Tree	0.422	0.309	0.556	< 1 min
XGBoost Regressor	0.419	0.306	0.553	< 5 min
GB Regressor	0.420	0.306	0.553	90 min
Random Forest	0.418	0.305	0.552	111 min

Supplementary results in Appendix I further support these findings, indicating that more sophisticated ML models exhibit only marginal enhancements over the simple baseline in terms of the Maximum Error, Median Absolute Error and Explained Variance. Moreover, the similarity in performance metrics for both the test and training sets (performance on the training set is available in Appendix J) indicates that the models are not prone to overfitting. The subsequent section will explore the results of the XGBoost model, with selected outcomes for the DT, RF, and GBR models provided in Appendix K, L, and M, respectively.

Although the results of XAI will be interpreted in the next section, caution is warranted as the ML models exhibit suboptimal performance, particularly for extreme values, as detailed in the error analysis in Appendix O. The consistent prediction of values in the range of 2 to 3 highlights the need for improvement in capturing diverse patterns. Additionally, confusion matrices for relatively healthy and unhealthy baskets show a high false negative rate, challenging the model's accuracy in providing meaningful insights into nutritional quality.

4.1.1 *Feature Importance & Partial Dependence Plots*

Figure 5 displays the importance scores generated by the tuned XGBoost model. Interestingly, the scores differ from those provided previously. It is important to note that Figure 2 aimed to offer an initial understanding using the untuned DT model, while Figure 5 presents the actual scores after tuning the XGBoost model. The latter reveals that 'Year' is not as crucial as initially assumed. Instead, the ratio of private label products and the age group of panelists, in addition to the number of items per basket and the income group of panelists, appear to be significant.

Furthermore, PDPs offer a nuanced perspective alongside feature importance scores. While feature importance scores quantify each feature's overall impact on model predictions, PDPs provide a detailed insight into how a specific feature influences predictions in isolation while keeping other features constant. Using PDPs, the impact of a feature on the model's prediction is illustrated by fixing the feature values, varying the fixed feature over a range of values, and averaging the model's predictions for each value. For instance, Figure 6 depicts a respectively positive and negative linear relationship between 'Household Size' and 'Social Group' and the predicted average Nutri-Score, maintaining constant values for other features. Notably, the differences are minimal, and the y-axis is highly zoomed in for a detailed view.

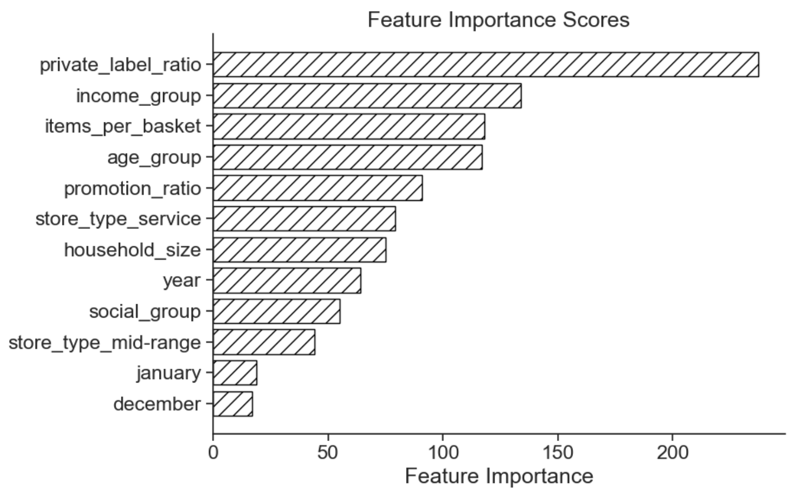


Figure 5: Feature Importance: XGBoost Regressor

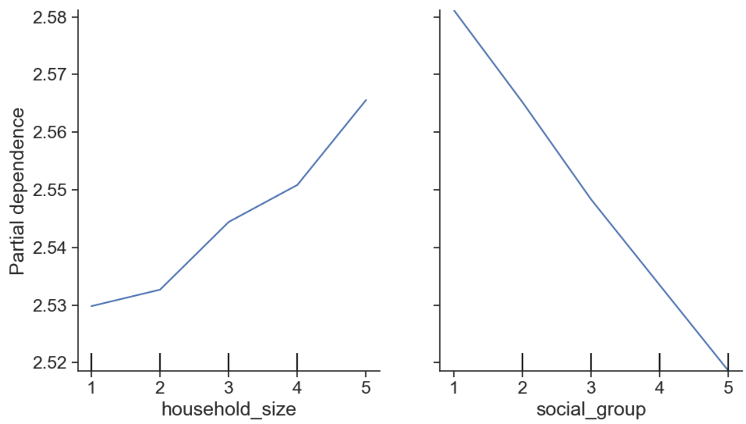


Figure 6: PDPs (XGBoost Regressor): Household Size & Social Group. Please note that the numbering of social groups corresponds to the social groups specified in Appendix B.

Furthermore, Figure 7 illustrates that as 'Private Label Ratio' increases, the predicted average Nutri-Score decreases, suggesting healthier shopping baskets, while holding all other feature values constant. A similar relationship holds for 'Promotion Ratio', although the association is less strong. The PDPs for the remaining features are provided in Appendix N.

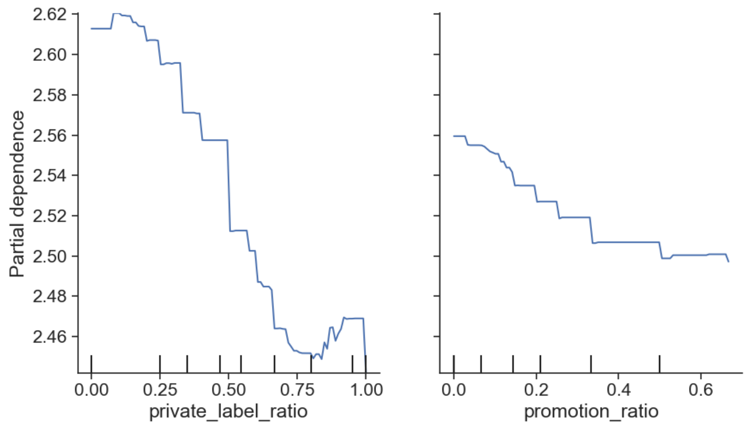


Figure 7: PDPs (XGBoost Regressor): Private Label Ratio & Promotion Ratio

4.1.2 *Shapley Additive Explanations Values*

The SHAP summary plot in Figure 8 visually demonstrates the impact of the features on the average Nutri-Score predictions, explicitly showing both the strength and direction of each feature (Chan et al., 2022). Notably, a higher ratio of private label products correlates with a lower average Nutri-Score (*i.e.*, healthier shopping baskets), while a larger shopping basket is associated with a higher average Nutri-Score (*i.e.*, less healthy shopping baskets). To delve deeper into the ML model's insights on individual features, SHAP PDPs are generated for the crucial features 'Private Label Ratio', 'Items per Basket' and 'Age Group', alongside an intriguing interaction feature. Although similar plots were explored for other features, they are not presented here, as they did not reveal distinct interactions.

A SHAP dependence plot illustrates a single feature's influence on model predictions, with dots representing dataset predictions. The x-axis denotes the feature's value, the y-axis shows the SHAP value indicating its impact, and color indicates a potential interaction effect with another feature, visible as a distinct vertical pattern (S. M. Lundberg & Lee, 2017). Figure 9 shows the relationship between income groups and the private label item ratio in shopping baskets. Lower income groups positively impact predictions, as indicated by slightly higher SHAP values. The plot highlights an interaction effect with 'Private Label Ratio', where a higher ratio contributes positively to predictions for low-income groups (depicted in red), suggesting a preference for less healthy shopping baskets. Conversely, for high-income groups, a higher ratio of private label products negatively impacts predictions, indicating healthier choices.

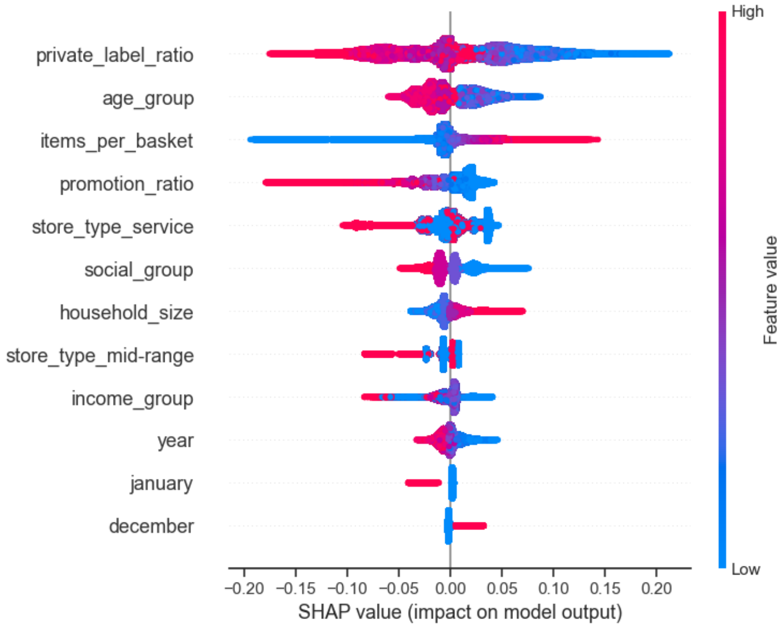


Figure 8: SHAP Summary (XGBoost Regressor)

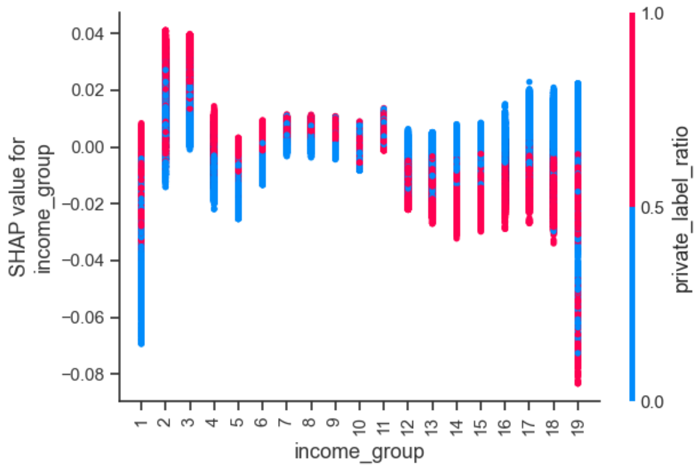


Figure 9: SHAP Dependence Plot (XGBoost Regressor): Income Group vs. Private Label Ratio. Please note that the numbering of income groups corresponds to the income groups specified in Appendix B.

Figure 10 depicts the interaction between 'Household Size' and 'Items per Basket'. Smaller households generally result in slightly lower SHAP values, indicating a negative effect on the model's predictions. The plot suggests a positive impact on the model's predictions for larger household sizes with an increased number of items per basket, reflected in higher (red) dots. Conversely, smaller household sizes negatively affect the model's predictions with larger shopping baskets, suggesting that smaller households tend to make healthier purchases when the shopping basket is larger.

Finally, Figure 11 illustrates the SHAP dependence plot for 'Age Group' and 'Promotion Ratio'. Younger age groups generally positively impact the model's predictions, as indicated by slightly higher SHAP values. The plot suggests that, for younger age groups, an increase in the ratio of products on promotion has a positive effect on predictions, while for older age groups, it has a negative impact. This implies that households with younger household heads tend to make less healthy purchases when there are more promoted items in the basket, whereas households with older household heads make healthier purchases in such scenarios. Please consider that the interaction effect for the features in Figure 11 is relatively weaker compared to Figure 9 and 10. This is evident in the less distinct separation of red and blue dots.

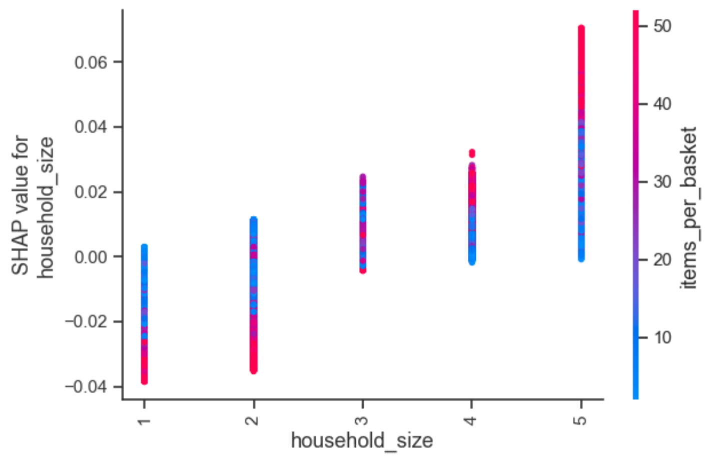


Figure 10: SHAP Dependence Plot (XGBoost Regressor): Household Size vs. Items per Basket

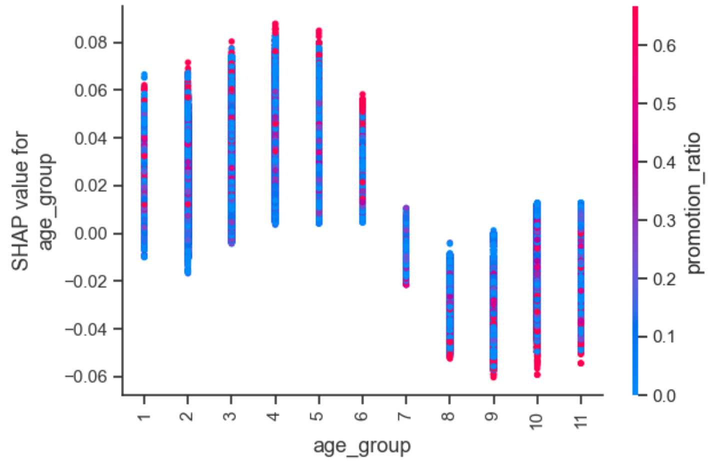


Figure 11: SHAP Dependence Plot: Age Group vs. Promotion Ratio (XGBoost Regressor). Please note that the numbering of age groups corresponds to the age groups specified in Appendix B.

4.2 *Association Rule Mining*

In this study, both for healthy and unhealthy shopping baskets, the minimum support level is set at 0.5% to find meaningful patterns in retail transactions with numerous products. Additionally, the study focuses on the 706 specific category names of products rather than individual barcodes, ensuring the identification of patterns across different supermarkets. Furthermore, the minimum confidence and lift levels are set at 0.3 and 2, respectively, to filter out less significant associations. With these settings, the analysis reveals a total of 509 distinct rules for unhealthy shopping baskets, and a total of 67 rules for healthy shopping baskets. To streamline the presentation of results, this study reports the first 20 rules when arranged in descending order of the lift measure.

Table 9 presents association rules and associated metrics for unhealthy shopping baskets. Notable patterns include the strong positive relationships between certain antecedents and the consequent 'extruded products[6]'. For instance, when 'peanuts and nuts' and 'chips' are both present, there is a support of 0.006, confidence of 0.344, and lift of 4.564, indicating a high likelihood of 'extruded products' being part of the basket. Extruded prod-

[6] In the context of food, extruded refers to a process where the food material is forced through a machine, typically in a specific shape or form. This process is commonly used in the production of certain snacks, cereals, and pasta.

ucts, distinct from extruded bread substitutes in the GfK dataset, include specific types of chips such as Cheetos®, processed into a particular form. Similar patterns emerge with other antecedents such as 'deli meats', 'cookies' and combinations such as 'milk' and 'chips'. These rules, characterized by high confidence and lift values, suggest a consistent association between specific items and the inclusion of 'extruded products' in the shopping basket, providing insights into consumer behavior and preferences for unhealthy food choices.

While the focus of the association rules is on unhealthy shopping baskets, it is intriguing to note the inclusion of seemingly healthier products such as 'milk', 'yogurt', and 'cheese' in some rules. This observation highlights the complexity of shopping behavior, where individuals may opt for a mix of both healthy and unhealthy items in their overall purchases. The association of these relatively healthier products with items such as 'chips', 'soft drinks' and 'extruded products' suggests that consumers might be making conscious choices to balance their shopping baskets with both indulgent and nutritious options.

Table 9: Association Rules and Metrics: Unhealthy Shopping Baskets

Antecedent	Consequent	Sup.	Conf.	Lift
(milk, extruded products)	(soft drinks, chips)	0.007	0.310	6.146
(Dutch cheese, soft drinks, chips)	(extruded products)	0.005	0.365	4.832
(salty biscuits-cookies, chips)	(extruded products)	0.005	0.358	4.748
(peanuts and nuts, chips)	(extruded products)	0.006	0.344	4.564
(deli meats, chips)	(extruded products)	0.007	0.344	4.553
(foreign cheese, chips)	(extruded products)	0.007	0.333	4.415
(custard, chips)	(extruded products)	0.006	0.330	4.367
(small cookies, chips)	(extruded products)	0.006	0.328	4.347
(biscuit, chips)	(extruded products)	0.006	0.323	4.278
(large cookies, chips)	(extruded products)	0.008	0.322	4.270
(soft drinks, chips)	(extruded products)	0.016	0.321	4.258
(yogurt, chips)	(extruded products)	0.007	0.319	4.233
(fruit juices-drinks, chips)	(extruded products)	0.007	0.319	4.232
(fresh deli, chips)	(extruded products)	0.008	0.314	4.156
(milk, soft drinks, custard)	(yogurt)	0.006	0.460	4.054
(bread spread, custard)	(yogurt)	0.005	0.460	4.048
(milk, chips)	(extruded products)	0.011	0.303	4.022
(Dutch cheese, chips)	(extruded products)	0.009	0.303	4.010
(gingerbread, custard)	(yogurt)	0.006	0.444	3.914
(biscuit, custard)	(yogurt)	0.007	0.442	3.893
(milk, Dutch cheese, deli meats)	(fresh deli)	0.005	0.471	3.890

Notes. 'sup.' indicates support; 'conf.' indicates confidence

The association rules presented in Table 10 offer insights into shopping patterns associated with healthier nutritional choices. Notably, the presence of certain items, such as 'bell pepper' is positively correlated with the likelihood of purchasing 'vegetables other' with a support of 0.006, confidence of 0.425, and lift of 4.941. Similarly, combinations such as 'bell pepper', 'yogurt' and 'vegetables other salad' are linked, indicating a 0.442 confidence and a lift of 3.632. Distinct vegetables identified within the specified categories include leafy greens, bell peppers, onions, carrots, cucumbers, corn, and mushrooms. Broader classifications include 'vegetables other', 'frozen vegetables', and 'canned vegetables'. The rule involving 'pears' suggests a connection with 'fruit large other' with a confidence of 0.391 and a lift of 4.118. Distinct fruit categories within the specified categories include but are not limited to apples, pears, citrus fruits (other), and oranges. The category labeled 'vegetables other salad' includes snackable vegetables, such as cherry tomatoes, (small) cucumbers, and celery. Moreover, the designation 'K' corresponds to 'koelvers', indicating products that are chilled or refrigerated.

Table 10: Association Rules and Metrics: Healthy Shopping Baskets

Antecedent	Consequent	Sup.	Conf.	Lift
(veg. o. K, salad veg. o. K)	(veg. o. K salad)	0.005	0.315	5.096
(bell pepper K)	(veg. o. K)	0.006	0.425	4.941
(veg. o., veg. o. K salad)	(veg. o. K)	0.006	0.367	4.262
(pears K)	(fruit large other K)	0.006	0.391	4.118
(veg. o. K salad, fruit large other K)	(veg. o.r K)	0.006	0.349	4.055
(apples K)	(fruit large other K)	0.014	0.374	3.947
(veg. o., fruit large other K)	(veg. o. K)	0.007	0.335	3.892
(yogurt, veg. o. K)	(fruit large other K)	0.006	0.366	3.856
(veg. o. salad, fruit large other K)	(veg. o. K)	0.006	0.326	3.784
(fruit compote)	(vegetables preserves)	0.007	0.369	3.706
(bell pepper, yogurt)	(veg. o. salad)	0.005	0.442	3.632
(veg. o., bell pepper)	(veg. o. salad)	0.010	0.425	3.491
(mixed vegetables, bell pepper)	(veg. o. salad)	0.006	0.418	3.435
(veg. o., leafy vegetables lettuce)	(veg. o. salad)	0.008	0.411	3.377
(leafy vegetables lettuce, yogurt)	(veg. o. salad)	0.005	0.397	3.257
(buttermilk)	(yogurt)	0.006	0.377	3.243
(veg. o., fruit large other)	(veg. o. salad)	0.007	0.387	3.181
(bell pepper, fresh mushrooms)	(veg. o. salad)	0.006	0.380	3.119
(foreign cheese)	(veg. o. salad)	0.009	0.377	3.097
(milk, veg. o. salad)	(yogurt)	0.005	0.356	3.062

Notes. 'veg. o.' indicates 'vegetables other'; 'sup.' indicates support; 'conf.' indicates confidence

5 DISCUSSION

5.1 *Discussion of Results*

Predicting the average Nutri-Score of a shopping basket poses challenges, with complex ML models offering only marginal improvements over a simple baseline. The XGBoost model's MAE of 0.419 and RMSE of 0.553, together with the model's suboptimal performance for extreme values, raise concerns about assessing basket healthiness. XGBoost is regarded as the top-performing model, despite RF displaying slightly more favorable metrics, when considering both computation time and accuracy. Comparisons with existing literature are complicated by the novel approach of this study. For instance, Tachie et al. (2023) predict individual product Nutri-Scores on a different scale (-15 to 40), achieving MAE and MSE of 0.97 and 2.62, respectively. Furthermore, feature variations exist, as Tachie et al. (2023) focus on micronutrients and product-level Nutri-Scores, while this study emphasizes household panel features and the average Nutri-Score of a shopping basket.

The decision to predict the average Nutri-Score of a shopping basket using demographic details and basket-related features, rather than individual purchase data, is grounded in three key considerations. Firstly, this aligns with the study's goal of addressing unhealthy dietary choices by exploring factors influencing consumers' nutritional decisions during grocery shopping. Secondly, methodologically, opting for demographic details and basket-related features is a strategic decision. Since products in the basket are directly tied to the average Nutri-Score, using individual purchase data could result in a less complex and less informative model. Lastly, choosing broader demographic and basket-level features enhances method applicability across diverse data contexts where detailed individual purchase data might be limited or unavailable (Egan et al., 2014).

Subsequently, the feature importance results will be explored in connection with existing literature. Hossain et al. (2019) emphasize the critical role of factors such as education level, household size, and household assets in predicting caloric intake using RF. While XGBoost is the primary focus in this study, the results demonstrate a consistent pattern. The feature importance scores presented in Figure 5 suggest that 'Income Group', 'Household Size', and 'Social Group' are potentially influential features in predicting the average Nutri-Score. This aligns with the emphasis on household assets, household size, and education level in the context of food security, as noted by Hossain et al. (2019). Additionally, SHAP values in Figure 8 provide insights into the direction of the features' impact on the model's predictions. Higher values of 'Private Label Ratio', 'Age

Group', 'Promotion Ratio', 'Store Type Service', 'Social Group', 'Store Type Mid-range', and 'Income Group' generally negatively impact predictions, indicating slightly healthier shopping baskets. Conversely, higher values of 'Items per Basket' and 'Household Size' generally have a positive impact, indicating slightly less healthy shopping baskets.

The observed feature trends align with prior research. Ogundijo et al. (2021) propose that older individuals lean towards healthier food choices, while higher-income individuals prefer organic and locally sourced foods. Although not directly tied to the Nutri-Score, these findings highlight the preferences of specific demographic groups for healthier food selections. In a related context, Volpe et al. (2018) evaluate consumers' food purchases using the USDA's Healthy Eating Index and Consumer Nutrition Environment Index, finding that higher household income, larger household size, and higher education correlate with healthier shopping baskets. While generally consistent, this study reveals a positive impact of larger household sizes on Nutri-Score predictions, suggesting less healthy shopping baskets. This influence may be attributed to the number of items in a basket, as shown in Figure 10, where larger baskets are associated with less healthy predictions, especially for larger household sizes. It is prudent to interpret these findings with caution, given the models' challenges in accurately predicting extremely (un)healthy average Nutri-Scores.

Beyond predicting average Nutri-Scores, this study delves into healthy and unhealthy shopping baskets through ARM. Literature typically sets support and confidence thresholds around 0.1% and 50-80%, respectively (Rana & Mondal, 2021; Supriyadi, 2020). The rules presented in Tables 9 and 10 align with these thresholds, with slightly higher support and lower confidence levels. However, the emphasis on Lift scores over support and confidence metrics aims to spotlight rules that unveil surprising associations, offering profound insights into consumers' nutritional choices. Unlike support and confidence, which measure frequency and reliability, respectively, Lift identifies the strength and significance of associations between items. This supplementary analysis becomes particularly evident in the following subsection, underscoring the significance of the findings for diverse stakeholders.

5.2 *Relevance*

The findings have potential implications for various stakeholders, including policymakers, researchers, and marketers. Policymakers addressing public health concerns related to nutrition may find valuable insights into factors influencing shopping basket healthiness, facilitating more targeted interventions. For instance, the SHAP values suggest that a higher ratio of private label products is associated with healthier shopping baskets, particularly for individuals in higher income groups, aligning with existing literature emphasizing the role of income in shaping dietary choices. Policymakers could consider using this information to design incentives promoting healthier private label options. Additionally, the study contributes to the nutritional analysis field by shedding light on the complexities of predicting average Nutri-Scores. Insights from SHAP values and DPDs, such as the impact of household size and private label ratio, present avenues for further research into socioeconomic factors influencing shopping choices and overall basket healthiness. This aligns with existing literature emphasizing comprehensive studies considering multiple demographic and contextual factors in understanding dietary patterns (Hossain et al., 2019; Ogundijo et al., 2021; Volpe et al., 2018).

Marketers in the food industry can capitalize on the ARM results, which uncover patterns associated with (un)healthy shopping baskets. For instance, the association rule linking 'peanuts and nuts' and 'chips' to the presence of 'extruded products' provides concrete insights into consumer snack preferences. Marketers can strategically position healthier snack alternatives, such as 'cherry tomatoes' or 'cucumbers' which are apparent in healthier baskets, in proximity to these popular combinations to encourage healthier choices. Moreover, the identification of seemingly healthier products such as 'milk' and 'yogurt' in unhealthy baskets highlights the complexity of consumer choices. Marketers can use this information to develop targeted marketing campaigns that emphasize the nutritional benefits of these products, potentially influencing consumer perceptions and choices. This aligns with existing literature on the importance of product placement and marketing strategies in shaping consumer behavior towards healthier food choices (Hecht et al., 2020).

5.3 *Limitations & Future Work*

The study is not without limitations, and future work could address several aspects to enhance the robustness of the findings. Notably, the Nutri-Scores are estimated at the category level, following the approach by Clark et al. (2022), rather than at the more granular product level. This distinction is crucial as it affects the precision of the nutritional assessment. Future research could explore the implications of estimating Nutri-Scores at the product level, acknowledging that nutritional content can vary within categories. Additionally, the linear mapping of Nutri-Scores to a 1-5 scale by Clark et al. (2022) raises concerns about treating it as a ratio-scaled variable. A more accurate representation might be as an ordinal-scaled variable, considering the different cut-offs for Nutri-Scores labelled on the A-E scale (refer to Appendix D, point 5).

Another limitation related to the Nutri-Score is that the Nutri-Score may not be the optimal variable for expressing the healthiness of a shopping basket. This could also clarify why the model faces challenges, even when the features themselves may serve as indicators of shopping basket healthiness. According to Van Der Bend et al. (2022), the Nutri-Score may not fully capture food healthfulness due to the exclusion of certain essential nutrients. Including components such as vitamins or minerals could enhance differentiation among food products. Additionally, the "per 100 g or 100 ml" reference unit may inaccurately represent healthfulness, especially for foods consumed in larger portions (Van Der Bend et al., 2022). Accordingly, future research might explore alternative predictor variables for a more accurate depiction of basket healthiness.

Moreover, studying more countries would enhance the global perspective on dietary patterns, addressing the current limitation of focusing solely on the Netherlands in this study. Finally, optimizing hyperparameters on the entire training set rather than on a sample, could further refine the models and enhance the generalizability of the findings. These avenues for future research would contribute to a more nuanced and comprehensive understanding of the factors influencing shopping basket healthiness.

6 CONCLUSION

This section concludes the study by addressing the established research questions.

RQ1. *To what extent can ML algorithms, specifically DT, RF, GBR and XGBoost, effectively predict the average Nutri-Score of consumer shopping baskets based on household panel features, with a focus on assessing predictive accuracy using MAE, MSE, and RMSE metrics?*

In contrast to the baseline model, the fine-tuned DT, RF, GBR, and XGBoost algorithms exhibit marginal improvement, achieving MAE, MSE, and RMSE values of 0.422, 0.309, and 0.556; 0.418, 0.305, and 0.552; 0.420, 0.306, and 0.553; and 0.419, 0.306, and 0.553, respectively. Although RF and produces the lowest error metrics, the computation time of 90 minutes is also considerably higher. Considering that XGBoost results are only about 0.001 higher in error across the different metrics, but estimated with considerably lower computation time (around 5 minutes), XGBoost emerges as the best-performing model. The results of the models, however, raise concerns about potential inaccuracies in assessing basket healthiness, and comparisons with existing literature are challenging due to the novelty of this study's approach.

RQ2. *How can XAI techniques, such as feature importance analysis, partial dependence plots, and SHAP values, be effectively applied to enhance the interpretability of ML models in explaining the factors influencing nutritional choices in shopping settings?*

XAI enhances complex model interpretability, meeting evidence-based policy requirements and providing actionable insights. In this study, XGBoost feature importance analysis, PDPs, and SHAP values bridge the gap between data-driven predictive metrics and actionable decisions. The feature importance scores suggest that factors such as 'Private Label Ratio', 'Items per Basket', 'Age Group', and 'Income Group' may play influential roles in predicting average Nutri-Scores. PDPs visually illustrate relationships, including positive associations of the average Nutri-Score with 'Household Size' and negative linear relationships with 'Social Group', 'Private Label Ratio', and 'Promotion Ratio'. SHAP values offer detailed insights into individual feature contributions, revealing that higher values of 'Private Label Ratio' and 'Age Group' negatively impact predictions, while higher values of 'Items per Basket' have a positive effect. SHAP dependence plots unveil interaction effects, indicating that a higher ratio of private label products is associated with healthier shopping baskets, particularly for individuals in higher income groups. While these insights are valuable, it's crucial to acknowledge the complexity of predicting average Nutri-Scores and interpret the results with consideration of the models' limitations.

RQ3. *What insights can be derived from the association rules extracted through the Apriori algorithm, revealing patterns of co-occurrence and relationships between specific food and beverage items or categories in both healthy and unhealthy shopping baskets?*

Insights drawn from ARM, generated through the Apriori algorithm, unveil valuable patterns of co-occurrence and relationships among specific food and beverage items or categories in both healthy and unhealthy shopping baskets. For instance, identified associations such as the pairing of 'peanuts and nuts' and 'chips' with the presence of 'extruded products', or 'bell pepper' with 'chilled vegetables other' provide an intuitive understanding of consumer preferences. Besides, the presence of seemingly healthier items such as 'milk' and 'yogurt' in unhealthy baskets adds depth to this understanding, emphasizing the multifaceted nature of shopping decisions. Overall, these insights contribute to a richer comprehension of the factors influencing nutritional choices in diverse shopping settings.

RQ. *To what extent can ML algorithms predict the average Nutri-Score of consumer shopping baskets based on household panel features, and what insights can be gained regarding the factors influencing nutritional choices, including the exploration of ARM to uncover patterns within shopping baskets?*

The fine-tuned ML models, DT, RF, GBR, and XGBoost, demonstrate marginal improvement over the baseline, with XGBoost emerging as the best-performing model, achieving a MAE of 0.419 and RMSE of 0.553. However, these outcomes reveal suboptimal performance, particularly for extreme values. Furthermore, XAI techniques, such as feature importance analysis, PDPs, and SHAP values, enhance interpretability by revealing influential factors and visualizing relationships. Finally, insights from ARM using the Apriori algorithm uncover valuable patterns in both healthy and unhealthy shopping baskets, offering a nuanced understanding of consumer preferences and the multifaceted nature of shopping decisions.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/access.2018.2870052

Agarwal, S. (2013). Data mining: Data mining concepts and techniques. *2013 International Conference on Machine Intelligence and Research Advancement*, 203–207. https://doi.org/10.1109/ICMIRA.2013.45

Aggarwal, C. C. (2015, April). *Data mining*. Springer.

Agrawal, R., & Srikant, R. (1998). Fast algorithms for mining association rules. *Very Large Data Bases*, 580–592. http://pages.cs.wisc.edu/~nil/764/DADS/39_agrafa94.pdf

Agrawal, T. (2020, November). *Hyperparameter optimization in machine learning*. Apress.

apyori. (2019, November). https://pypi.org/project/apyori/1.1.2/

Bergmann, N. C., Davies, J. M., Lingvay, I., & Knop, K. F. (2022). Semaglutide for the treatment of overweight and obesity: A review. *Diabetes, Obesity and Metabolism*, *25*(1). https://doi.org/https://doi.org/10.1111/dom.14863

Bonaccorso, G. (2017, July). *Machine learning algorithms*. Packt Publishing Ltd.

Chan, M.-C., Pai, K.-C., Su, S.-A., Wang, M.-S., Wu, C.-L., & Chao, W.-C. (2022). Explainable machine learning to predict long-term mortality in critically ill ventilated patients: a retrospective study in central Taiwan. *BMC Medical Informatics and Decision Making*, *22*(1). https://doi.org/10.1186/s12911-022-01817-6

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Clark, M., Springmann, M., Rayner, M., Scarborough, P., Hill, J., Tilman, D., Macdiarmid, J. I., Fanzo, J., Bandy, L., & Harrington, R. A. (2022). Estimating the environmental impacts of 57,000 food products. *Proceedings of the National Academy of Sciences*, *119*(33). https://doi.org/10.1073/pnas.2120584119

De Temmerman, J., Heeremans, E., Slabbinck, H., & Vermeir, I. (2021). The impact of the nutri-score nutrition label on perceived healthiness and purchase intentions. *Appetite*.

de Carvalho, M. S., & da Silva, G. L. (2021). Inside the black box: Using explainable ai to improve evidence-based policies. *2021 IEEE 23rd Conference on Business Informatics (CBI)*, *02*, 57–64. https://doi.org/10.1109/CBI52690.2021.10055

Dowle, M., & Srinivasan, A. (2023). *Data.table: Extension of 'data.frame'*. https://r-datatable.com

Dréano-Trécant, L., Egnell, M., Hercberg, S., Galan, P., Soudon, J., Fialon, M., Touvier, M., Kesse-Guyot, E., & Julia, C. (2020). Performance of the front-of-pack nutrition label nutri-score to discriminate the nutritional quality of foods products: A comparative study across 8 european countries. *Nutrients*, *12*. https://doi.org/10.3390/nu12051303

Dubois, P., Albuquerque, P., Allais, O., Bonnet, C., Bertail, P., Combris, P., Lahlou, S., Rigal, N., Ruffieux, B., & Chandon, P. (2020). Effects of front-of-pack labels on the nutritional quality of supermarket food purchases: Evidence from a large-scale randomized controlled trial. *Journal of the Academy of Marketing Science*, *49*, 119–138. https://doi.org/10.2139/ssrn.3567974

Egan, M., Pliakas, T., Grace, D., McGill, E., Jones, A., Wong, J., Aalders, S., & Lock, K. (2014). Use of retail data in the assessment of natural experiments: The case of reducing the strength, an intervention to reduce alcohol availability. *The Lancet*, *384*. https://doi.org/10.1016/S0140-6736(14)62156-5

Fialon, M., Egnell, M., Talati, Z., Galan, P., Dréano-Trécant, L., Touvier, M., Pettigrew, S., Hercberg, S., & Julia, C. (2020). Effectiveness of different front-of-pack nutrition labels among italian consumers: Results from an online randomized controlled trial. *Nutrients*, *12*. https://doi.org/10.3390/nu12082307

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Glanz, K., Bader, M., & Iyer, S. (2012). Retail grocery store marketing strategies and obesity. *American Journal of Preventive Medicine*, *42*(5), 503–512. https://doi.org/10.1016/j.amepre.2012.01.013

Haag, F., Hopf, K., Vasconcelos, P. M., & Staake, T. (2022). Augmented cross-selling through explainable AI – a case from energy retailing. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2208.11404

Han, J., Kamber, M., & Pei, J. (2011, June). *Data Mining: Concepts and techniques*. Elsevier.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hecht, A. A., Perez, C., Polascek, M., Thorndike, A., Franckle, R., & Moran, A. (2020). Influence of food and beverage companies on retailer marketing strategies and consumer behavior. *International Journal of Environmental Research and Public Health*, *17*. https://doi.org/10.3390/ijerph17207381

Hemalatha, B. (2020). A survey on the online shopping customer review data using association rule mining. *International Journal of Data Mining Techniques and Applications*. https://doi.org/10.20894/ijdmta.102.009.001.011

Hossain, M., Mullaly, C., & Asadullah, M. N. (2019). Alternatives to calorie-based indicators of food security: An application of machine learning methods. *Food Policy*, *84*. https://doi.org/https://doi.org/10.1016/j.foodpol.2019.03.001

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Json encoder and decoder. (n.d.). Python Software Foundation. https://docs.python.org/3/library/json.html

Julia, C., Arnault, N., Agaësse, C., Fialon, M., Deschasaux-Tanguy, M., Andreeva, V., Fezeu, L., Kesse-Guyot, E., Touvier, M., Galan, P., & Hercberg, S. (2021). Impact of the front-of-pack label nutri-score on the nutritional quality of food choices in a quasi-experimental trial in catering. *Nutrients*, *13*. https://doi.org/10.3390/nu13124530

Julia, C., & Hercberg, S. (2018). Nutri-score: Evidence of the effectiveness of the french front-of-pack nutrition label. https://api.semanticscholar.org/CorpusID:209544289

Kamiński, M., Kręgielska-Narożna, M., & Bogdański, P. (2021). Seasonal variation in lifestyle behavior in Poland: Google searches and market sales analysis. *BMC Public Health*, *21*(1). https://doi.org/10.1186/s12889-021-11543-9

Karpyn, A., McCallops, K., Wolgast, H., & Glanz, K. (2020). Improving consumption and purchases of healthier foods in retail environments: A systematic review. *International Journal of Environmental Research and Public Health*, *17*(20). https://doi.org/https://doi.org/10.3390/ijerph17207524

Khan, M. I. H., Sablani, S. S., Nayak, R., & Gu, Y. (2022). Machine learning-based modeling in food processing applications: State of the art. *Comprehensive Reviews in Food Science and Food Safety*, *21*(2), 1409–1438. https://doi.org/10.1111/1541-4337.12912

Kuhn & Max. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kumarr, K. S., Naidu, M., & Radha, K. (2019). Pattern discovery with web usage mining using apriori and fp-growth algorithms. *International Journal of Computer Trends Technology*. https://doi.org/10.14445/22312803/IJCTT-V67I3P101

Li, C. (2009). Towards the healthy nutritional dietary patterns. *2009 Fourth International Conference on Digital Information Management*. https://doi.org/10.1109/icdim.2009.5356791

Lundberg, S., Erion, G., Chen, H., DeGrave, A. J., Prutkin, J. M., Nair, B. G., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, *2*(1), 2522–5839.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Martin, J. H., Nenycz-Thiel, M., Dawes, J., Tanusondjaja, A., Cohen, E., McColl, B., & Bogomolova, S. (2020). Fundamental basket size patterns and their relation to retailer performance. *Journal of Retailing and Consumer Services*, *54*, 102032. https://doi.org/10.1016/j.jretconser.2020.102032

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the Python in Science Conferences*. https://doi.org/10.25080/majora-92bf1922-00a

Melovic, B., Ćirović, D., Dudic, B. a. Z., Backović, T., & Gregus, M. (2020). The analysis of marketing factors influencing consumers' preferences and acceptance of organic food products—recommendations for the optimization of the offer in a developing market. *Foods*, *9*. https://doi.org/10.3390/foods9030259

Monterrosa, E., Frongillo, E. A., Drewnowski, A., De Pee, S., & Vandevijvere, S. (2020). Sociocultural influences on food choices and implications for sustainable healthy diets. *Food and Nutrition Bulletin*, *41*. https://doi.org/10.1177/0379572120975874

Nichols, B. S., Raška, D., & Flint, D. J. (2014). Effects of consumer embarrassment on shopping basket size and value: A study of the millennial consumer. *Journal of Consumer Behaviour*, *14*(1), 41–56. https://doi.org/10.1002/cb.1500

NOS. (2024). Kieshulp voor fris en chips is aangepast, maar Nutri-Score niet vrij van kritiek. https://nos.nl/artikel/2503427-kieshulp-voor-fris-en-chips-is-aangepast%20maar-nutri-score-niet-vrij-van-kritiek

Ogundijo, D. A., Tas, A., & Onarinde, B. (2021). Factors influencing the perception and decision-making process of consumers on the choice of healthier foods in the United Kingdom: a systematic review using narrative synthesis. *International Journal of Food Science and Technology*, *57*(2), 881–897. https://doi.org/10.1111/ijfs.15478

OpenAI. (2023). ChatGPT (version 3.5). https://openai.com/chatgpt

Öztoprak, S., & Orman, Z. (2022). A new model-agnostic method and implementation for explaining the prediction on finance data. *European Journal of Science and Technology*. https://doi.org/10.31590/ejosat.1079145

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pramoditha, R. (2023). What is the Dummy Variable Trap and How to Avoid it? https://medium.com/data-science-365/what-is-the-dummy-variable-trap-and-how-to-avoid-it-aeb227c2cd92

Prasad, A., Strijnev, A., & Zhang, Q. (2008). What can grocery basket data tell us about health consciousness? *International Journal of Research in Marketing*, *25*(4), 301–309. https://doi.org/10.1016/j.ijresmar.2008.05.001

Puddephatt, J.-A., Keenan, G. S., Fielden, A., Reaves, D. L., Halford, J. C., & Hardman, C. A. (2020). 'Eating to survive': A qualitative analysis of factors influencing food choice and eating behaviour in a food-insecure population. *Appetite*, *147*, 104547. https://doi.org/10.1016/j.appet.2019.104547

Radwan, A. M. (2017). Enhancing prediction on imbalance data by thresholding technique with noise filtering. *2017 8th International Conference on Information Technology (ICIT)*, 399–404. https://doi.org/10.1109/ICITECH.2017.8080033

Rallis, I., Markoulidakis, Y., Georgoulas, I., Kopsiaftis, G., Kaselimi, M., Doulamis, N., & Doulamis, A. (2022). Interpretation of net promoter score attributes using explainable ai. *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*. https://doi.org/10.1145/3529190.3529212

Rana, S., & Mondal, M. N. I. (2021). A seasonal and multilevel association based approach for market basket analysis in retail supermarket. *European Journal of Information Technologies and Computer Science*. https://doi.org/10.24018/compute.2021.1.4.31

Raorane, A., & Kulkarni, R. (2011). Data mining techniques: A source for consumer behavior analysis. *ArXiv*, *abs/1109.1202*. https://doi.org/10.5121/IJDMS.2011.3304

Sani, A., Samuel, P, N. N., Waseso, B., Gunadi, G., & Haryanto, T. (2022). Data mining on sales transaction data using the association method with apriori algorithm. *2022 10th International Conference on Cyber and IT Service Management (CITSM)*, 1–5. https://doi.org/10.1109/CITSM56380.2022.9935890

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., & Keim, D. (2019). Towards a rigorous evaluation of xai methods on time series. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 4197–4201. https://doi.org/10.1109/ICCVW.2019.00516

Shah, M., Shah, N., Shetty, A., Shah, D., & Gotmare, P. (2016). A comparative study of pattern recognition algorithms on sales data. *International Journal of Computer Applications*, *141*, 38–41. https://doi.org/10.5120/IJCA2016909463

Sharma, S. (2017). Concept of association rule of data mining assists mitigating the increasing obesity. *International journal of information retrieval research*, *7*(2), 1–18. https://doi.org/10.4018/ijirr.2017040101

Supriyadi, D. (2020). Penerapan association rule mining berbasis algoritma frequent pattern growth untuk rekomendasi penjualan. *7*, 135–148. https://doi.org/10.35957/jatisi.v7i2.339

Tachie, C., Tawiah, N. A., & Aryee, A. N. (2023). Using machine learning models to predict the quality of plant-based foods. *Current Research in Food Science*, *7*. https://doi.org/https://doi.org/10.1016/j.crfs.2023.100544

Tallón-Ballesteros, A., & Chen, C.-h. (2020, December). *Machine learning and artificial intelligence*. IOS Press.

Tarrahi, F. (2023). Obesitas. *TvPO*, *18*(2). https://doi.org/https://doi.org/10.1007/s12503-023-1587-5

Ter Borg, S., Steenbergen, E., Milder, I. E. J., & Temme, E. H. M. (2021). Evaluation of Nutri-Score in Relation to Dietary Guidelines and Food Reformulation in The Netherlands. *Nutrients*, *13*(12), 4536. https://doi.org/10.3390/nu13124536

Time access and conversions. (n.d.). Python Software Foundation. https://docs.python.org/3/library/time.html

Tiwari, R. (2023). Explainable ai (xai) and its applications in building trust and understanding in ai decision making. *INTERANTIONAL*

JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. https://doi.org/10.55041/ijsrem17592

Van Benthem, D. (2022). 'Met de Nutri-Score houd je de consument dom'. https://nos.nl/nieuwsuur/artikel/2437005-met-de-nutri-score-houd-je-de-consument-dom

Van Der Bend, D. L. M., Van Eijsden, M., Van Roost, M. H. I., De Graaf, C., & Roodenburg, A. J. C. (2022). The Nutri-Score algorithm: Evaluation of its validation process. *Frontiers in Nutrition*, *9*. https://doi.org/10.3389/fnut.2022.974003

Verma, Y. (2021, October). A complete guide to categorical data encoding -. https://analyticsindiamag.com/a-complete-guide-to-categorical-data-encoding/

Volpe, R., Jaenicke, E. C., & Chenarides, L. (2018). Store formats, market structure, and consumers' food shopping decisions. *Applied Economic Perspectives and Policy*, *40*(4), 672–694. https://doi.org/10.1093/aepp/ppx033

Walters, R. G., & Jamil, M. (2003). Exploring the relationships between shopping trip type, purchases of products on promotion, and shopping basket profit. *Journal of Business Research*, *56*(1), 17–29. https://doi.org/10.1016/s0148-2963(01)00201-6

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of open source software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wicaksono, D., Jambak, M. I., & Saputra, D. M. (2020). The Comparison of Apriori Algorithm with Preprocessing and FP-Growth Algorithm for Finding Frequent Data Pattern in Association Rule. *Advances in Intelligent Systems Research*, *172*. https://doi.org/10.2991/aisr.k.200424.047

Wickham, H. (2011). The Split-Apply-Combine Strategy for data analysis. *Journal of Statistical Software*, *40*(1). https://doi.org/10.18637/jss.v040.i01

Wickham, H. (2023). *Stringr: Simple, consistent wrappers for common string operations*. https://stringr.tidyverse.org

Wickham, H., & Bryan, J. (2023). *Readxl: Read excel files*. https://readxl.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://dplyr.tidyverse.org

Wickham, H., Hester, J., & Bryan, J. (2023). *Readr: Read rectangular text data*. https://readr.tidyverse.org

APPENDIX A: PACKAGES AND VERSIONS

Data cleaning is performed in R (version 4.3.2), while EDA and modeling are performed in Python (version 3.9.13).

Table 11: List of R and Python Packages with Versions and Sources

Package	Version	Source
R Packages		
`plyr`	1.8.9	(Wickham, 2011)
`dplyr`	1.1.3	(Wickham, François, et al., 2023)
`data.table`	1.14.8	(Dowle & Srinivasan, 2023)
`readr`	2.1.4	(Wickham, Hester, & Bryan, 2023)
`readxl`	1.4.3	(Wickham & Bryan, 2023)
`stringr`	1.5.1	(Wickham, 2023)
`caret`	6.0-94	(Kuhn & Max, 2008)
Python Packages		
`numpy`	1.21.5	(Harris et al., 2020)
`pandas`	1.4.4	(McKinney, 2010)
`scikit-learn`	1.0.2	(Pedregosa et al., 2011)
`xgboost`	2.0.2	(Chen & Guestrin, 2016)
`shap`	0.43.0	(S. Lundberg et al., 2020)
`shap TreeExplainer`	0.43.0	(S. M. Lundberg et al., 2020)
`apyori`	1.1.2	("apyori", 2019)
`matplotlib`	3.5.2	(Hunter, 2007)
`seaborn`	0.11.2	(Waskom, 2021)
`json`	2.0.9	(*JSON encoder and decoder*, n.d.)
`time`		(*Time access and conversions*, n.d.)

APPENDIX B: CODEBOOKS RAW DATA FILES

Table 12: Column Descriptions - Purchase file

Column Name	Column Description
Panelist	The internal ID of the household
Date_of_Purchase	The date that the purchase took place
Banner_Name	The name of the retailer in which the purchase occasion took place
Barcode	The barcode of the product that was purchased
Total_Unit_Sales	The amount of pieces of the product that was purchased
Total_Value_Sales	The amount of euro-cents that was spent for that purchase
Total_Volume_Sales	The amount of volume (gr, ml, pieces, etc.) that was purchased for that product
Quarter	The quarter the purchase took place in
Promo	Indicator if the product was on promotion (yes or no)
Purchase_method	Channel where the purchase was made (offline, online-delivery, online-pickup)

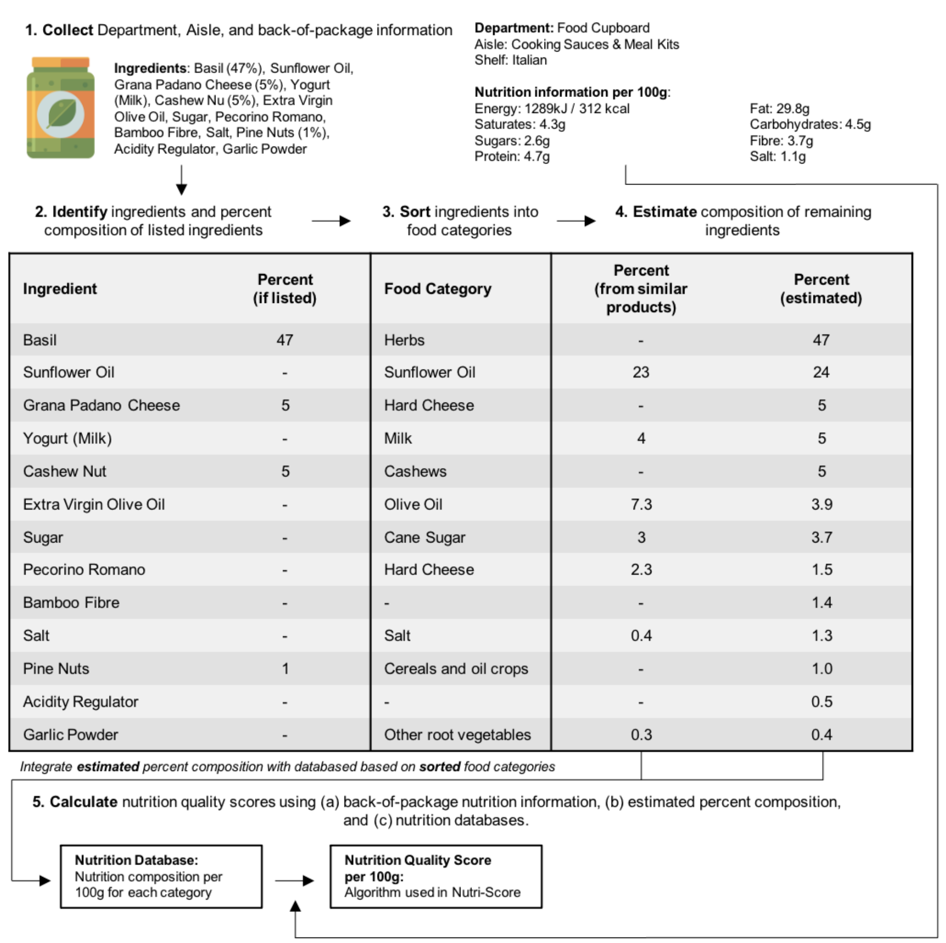
Table 13: Column Descriptions - Barcode file

Column Name	Column Description
Barcode	The barcode of the product
Barcode_Description	The description of the product
PL	Whether or not this product is a Private Label
Brand	The name of the brand that this product belongs to
Sub_Brand	The name of the sub-brand that this product belongs to
Category_Name	The name of the category that this product belongs to
Measurement_unit	The unit of mass that this product is measured with ("SU-Waarde", "grams", "ml", "stuks", "n/a")
Volume_per_unit	The size of the product
BG_category_name	BG20 category name the product belongs to, as identified by the country
BG_category_number	BG20 category number the product belongs to, as identified by the country
Country specific additional information	Additional variables delivered by the country all starting with "pa", e.g. paEKO, paDieet, paGezondheid, etc.

Table 14: Column Descriptions - Panelist file

Column Name	Column Description
Panelist	The internal ID of the household
Quarter	The quarter in which the panelist was present
Postal_code	The zip code within which the household resides
Age	The age of the head of the household. Age is delivered in 11 different age bands (12-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-64, 65-74, 75 or older)
Household_size	The number of household members
Social_class	The social band/Education of the head of the household ('A', 'B-plus', 'B-minus', 'C', 'D', Unknown)
Income_class	Net income of the household per month, based on all income sources for all household members ('Below 700 euro', '700-900 euro', '900-1100 euro', '1100-1300 euro', '1300-1500 euro', '1500-1700 euro', '1700-1900 euro', '1900-2100 euro', '2100-2300 euro', '2300-2500 euro', '2500-2700 euro', '2700-2900 euro', '2900-3100 euro', '3100-3300 euro', '3300-3500 euro', '3500-3700 euro', '3700-3900 euro', '3900-4100 euro', '4100 euro or more)
Region	The region within which the household resides
Province	The province within which the household resides

APPENDIX C: MACHINE LEARNING APPROACH NUTRI-SCORES



Ingredient	Percent (if listed)	Food Category	Percent (from similar products)	Percent (estimated)
Basil	47	Herbs	-	47
Sunflower Oil	-	Sunflower Oil	23	24
Grana Padano Cheese	5	Hard Cheese	-	5
Yogurt (Milk)	-	Milk	4	5
Cashew Nut	5	Cashews	-	5
Extra Virgin Olive Oil	-	Olive Oil	7.3	3.9
Sugar	-	Cane Sugar	3	3.7
Pecorino Romano	-	Hard Cheese	2.3	1.5
Bamboo Fibre	-	-	-	1.4
Salt	-	Salt	0.4	1.3
Pine Nuts	1	Cereals and oil crops	-	1.0
Acidity Regulator	-	-	-	0.5
Garlic Powder	-	Other root vegetables	0.3	0.4

Figure 12: Machine Learning Approach to Estimate Nutri-Scores of Products (Adapted from Clark et al. (2022), p. 3)

APPENDIX D: EXAMPLE CALCULATION NUTRI-SCORE

Example Calculation of Nutri-Score Estimation (Adapted from Supplementary Text; Estimating Nutrition Quality of Food Products; Example Calculations of a Product's Nutrition Impact Score (Clark et al., 2022)) Consider a hypothetical product containing 100g of the following: 400kJ of energy, 3g of sugar, 0.5g of saturated fats, 700mg of sodium, 5g of fiber, 10g of protein, and with 30 percent of its composition being fruit, vegetables, nuts, or healthy oils (**FVNO**).

1. Each aspect receives a score within its potential range: 0-10 for energy, sugar, saturated fat, and sodium, and 0-5 for fiber, protein, and FNVO. Higher scores indicate a higher content of the respective component in the product. For this example, the scores are as follows: energy: 1, sugar: 0, saturated fats: 0, sodium: 7, fiber: 5, protein: 5, FVNO: 0.

2. The scores for negative components (energy, sugar, saturated fat, and sodium) are summed to assess aspects associated with poorer health. In this example, the sum is 8 (1 + 0 + 0 + 7).

3. The scores for positive components (fiber, protein, and FVNO) are summed to evaluate aspects linked to health benefits. For this example, the sum is 10 (5 + 5 + 0).

4. The overall Nutri-Score is determined by subtracting the sum of positive component scores from the sum of negative component scores. In this example, this results in a value of -2 (8 - 10).

5. Finally, the numeric value is mapped to the A-E scale used in Nutri-Score, which involves specific thresholds. Solid foods are classified as 'A' if their score falls between -15 and -1, 'B' for scores from 0 to 2, 'C' for scores from 3 to 10, 'D' for scores from 11 to 18, and 'E' for scores from 19 to 40. Beverage thresholds are slightly different. Water is 'A,' beverages with a score of 1 or less are 'B,' scores from 2 to 5 are 'C,' scores from 6 to 9 are 'D,' and scores from 10 to 40 are 'E.'

APPENDIX E: CATEGORY LIST

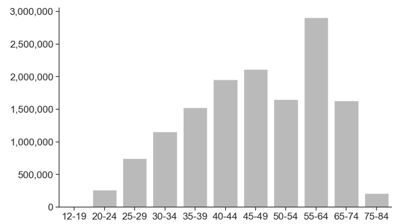
Table 15: Clark et al. (2022) Categories

50% Beef and 50% Pork
Bakery Free From
Beef and Lamb
Biscuits & Cereal Bars
Bread & Rolls
Breakfast Cereals
Cakes, Cake Bars, Slices & Pies
Cheese
Chilled Desserts
Chocolate
Coffee
Cooking Ingredients
Cooking Sauces & Meal Kits
Crackers & Crispbreads
Crisps, Snacks & Popcorn
Croissants, Brioche & Pastries
Crumpets, Muffins & Pancakes
Dairy Alternatives
Deli Meat and Cheese
Desserts
Doughnuts, Cookies & Muffins
Dried Fruit, Nuts, Nutrient Powders & Seeds
Dried Pasta, Rice, Noodles & Cous Cous
Easy Entertaining
Fish & Seafood
Fizzy Drinks & Cola
Fresh Fruit and Nuts
Fresh Salad & Dips
Fresh Soup, Sandwiches & Salad Pots
From our Bakery
Frozen Breakfast, Fruit & Pastry
Frozen Chips, Onion Rings, Potatoes & Rice
Frozen Desserts, Ice Cream & Ice Lollies
Frozen Party Food & Sausage Rolls
Frozen World Foods & Halal
Frozen Yorkshire Puddings & Stuffing

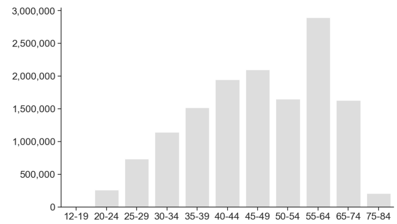
Table 16: Clark et al. (2022) Categories - Continued

Gluten Free Range
Home Baking
Hot Chocolate & Malted Drinks
Hot Drinks
Jams, Sweet & Savoury Spreads
Juices & Smoothies
Kids & Lunchbox Drinks
Meat
Meat Alternatives
Milk, Butter & Eggs
Milkshake
Olives, Antipasti, Pickles & Chutneys
Pies, Quiches & Party Food
Pizza & Garlic Bread
Premium Drinks & Mixers
Ready Meals
Sports & Energy Drinks
Squash & Cordial
Sugar & Sweeteners
Sweets, Mints & Chewing Gum
Table Sauces, Marinades & Dressings
Tea
Teacakes, Fruit Loaves & Scones
Tins, Cans & Packets
Vegetables
World Foods
Wraps, Pittas, Naan & Thins
Yoghurts

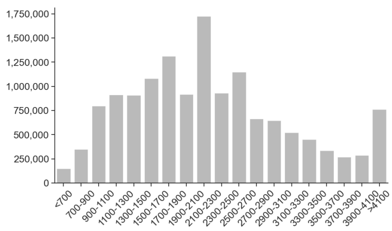
APPENDIX F: DISTRIBUTIONS BEFORE AND AFTER REMOVAL MISSING VALUES



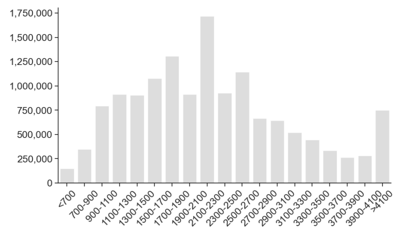
(a) Age Group (Before)



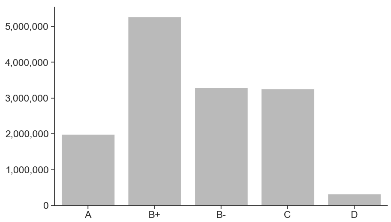
(b) Age Group (After)



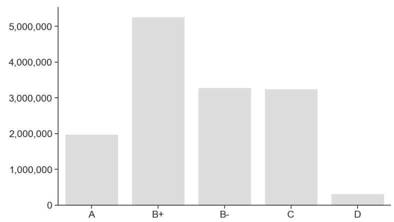
(c) Income Group (Before)



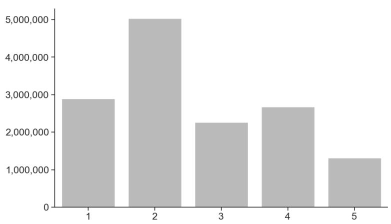
(d) Income Group (After)



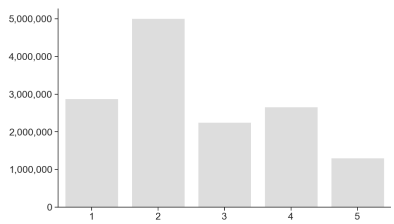
(e) Social Group (Before)



(f) Social Group (After)

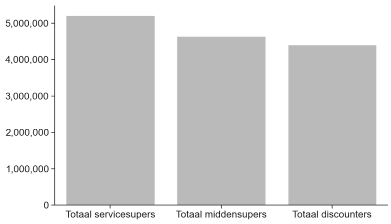


(g) Household Size (Before)

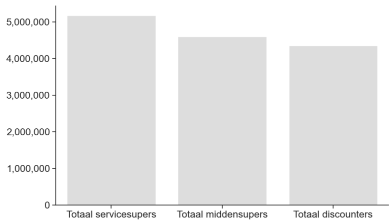


(h) Household Size (After)

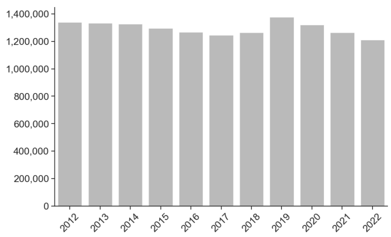
Figure 13: Distributions of Socioeconomic Variables Before (Dark-Grey) and After (Light-Grey) Removal of Missing Values



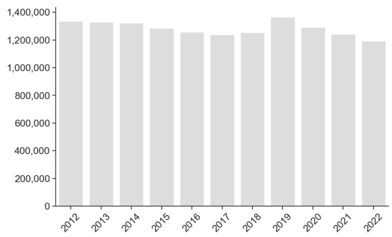
(a) Store Type (Before)



(b) Store Type (After)

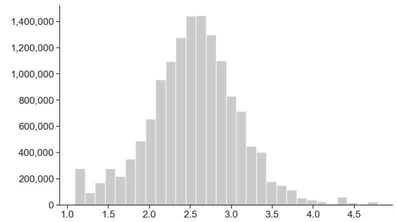


(c) Year (Before)

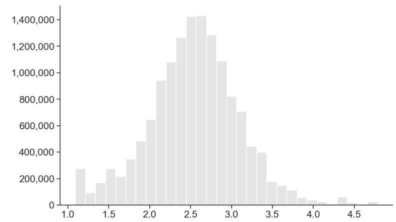


(d) Year (After)

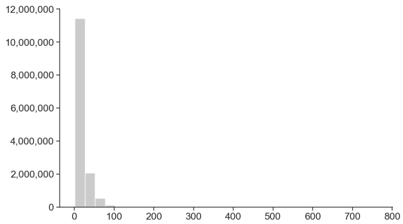
Figure 14: Distributions of Other Categorical Variables Before (Dark-Grey) and After (Light-Grey) Removal of Missing Values



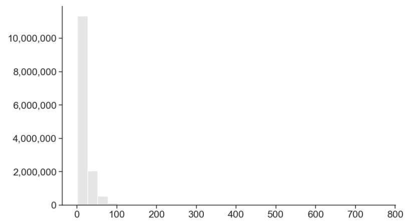
(a) Nutri-Score (Before)



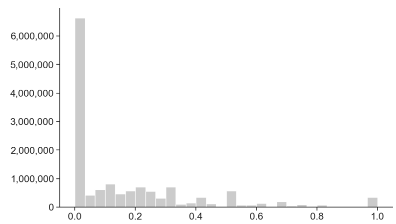
(b) Nutri-Score (After)



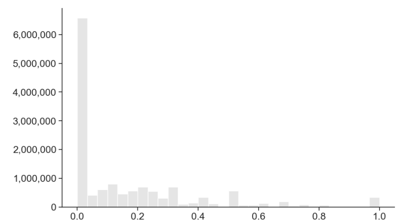
(c) Items per Basket (Before)



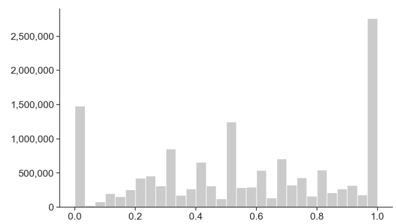
(d) Items per Basket (After)



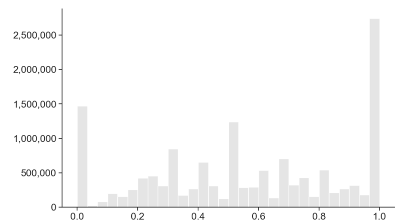
(e) Promotion Ratio (Before)



(f) Promotion Ratio (After)



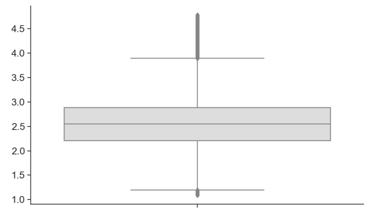
(g) Private Label Ratio (Before)



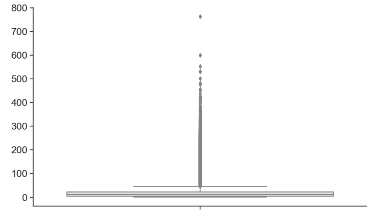
(h) Private Label Ratio (After)

Figure 15: Distributions of Numerical Variables Before (Dark-Grey) and After (Light-Grey) Removal of Missing Values

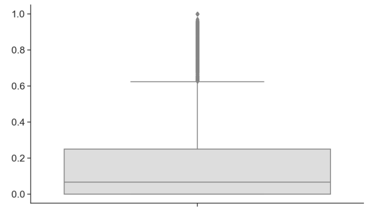
APPENDIX G: OUTLIER ANALYSIS



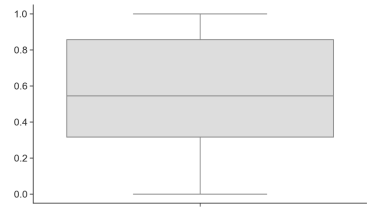
(a) Nutri-Score



(b) Items Per Basket



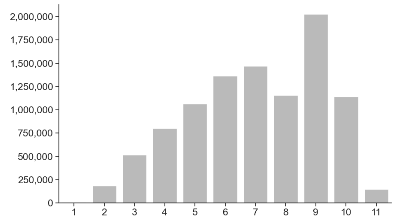
(c) Promotion Ratio



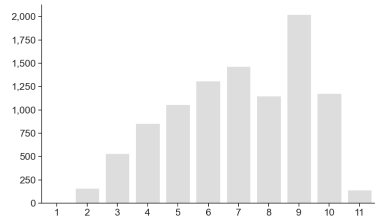
(d) Private Label Ratio

Figure 16: Boxplots of Numerical Variables

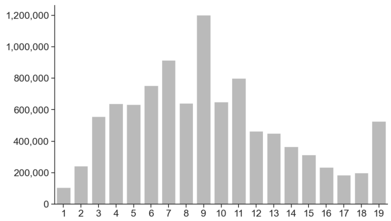
APPENDIX H: DISTRIBUTIONS TRAINING SET



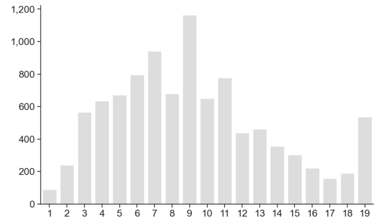
(a) Age Group (Complete)



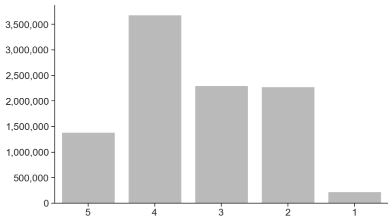
(b) Age Group (Sample)



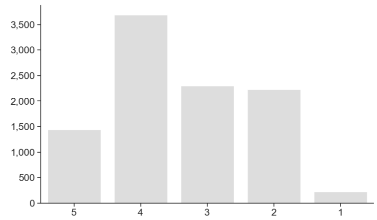
(c) Income Group (Complete)



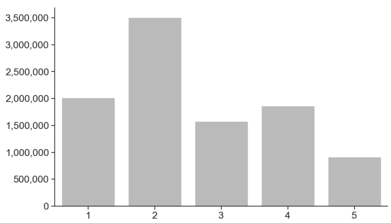
(d) Income Group (Sample)



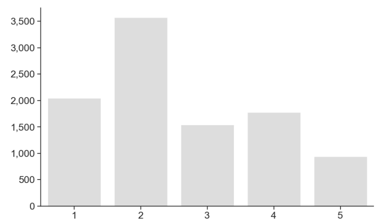
(e) Social Group (Complete)



(f) Social Group (Sample)

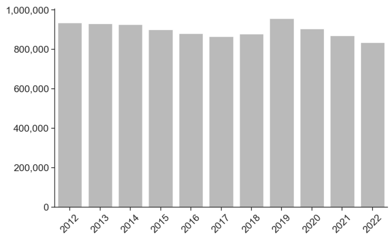


(g) Household Size (Complete)

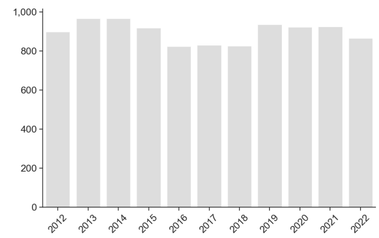


(h) Household Size (Sample)

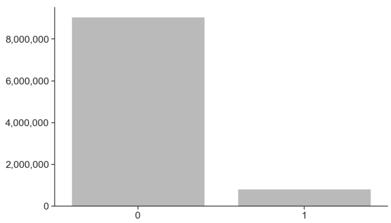
Figure 17: Distributions of Socioeconomic Variables in Training Set: Complete (Dark-Grey) and Sample (Light-Grey)



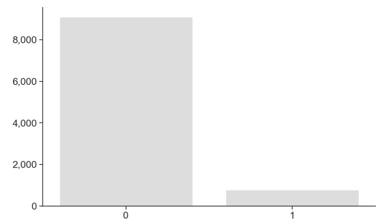
(a) Year (Complete)



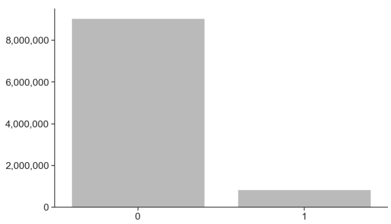
(b) Year (Sample)



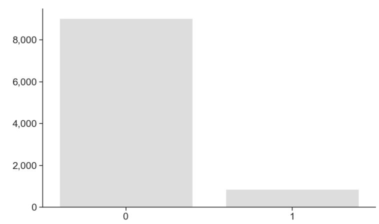
(c) December (Complete)



(d) December (Sample)

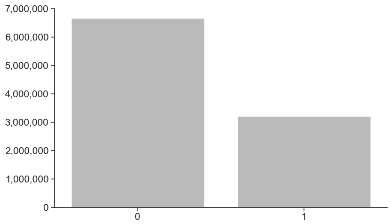


(e) January (Complete)

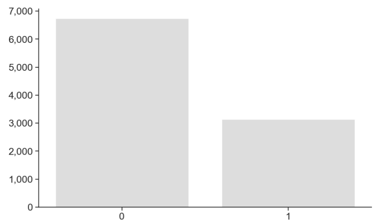


(f) January (Sample)

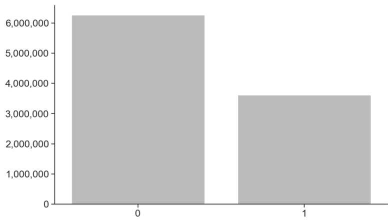
Figure 18: Distributions of Other Categorical Variables in Training Set: Complete (Dark-Grey) and Sample (Light-Grey). Note that 'December' and 'January' dummy variables are added to address seasonality.



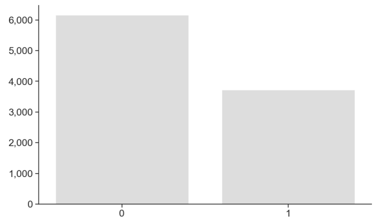
(a) Store Type: Mid-Range (Complete)



(b) Store Type: Mid-Range (Sample)

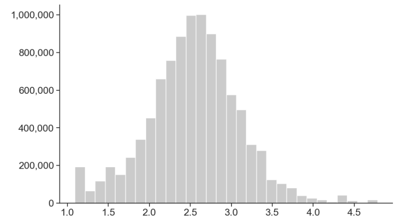


(c) Store Type: Service (Complete)

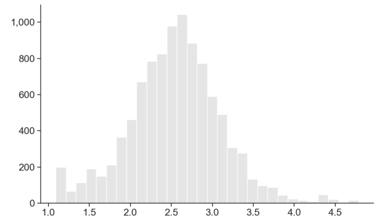


(d) Store Type: Service (Sample)

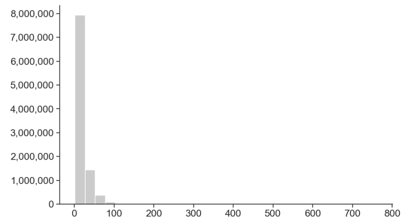
Figure 19: Distributions of Other Categorical Variables in Training Set - Continued: Complete (Dark-Grey) and Sample (Light-Grey). 'Store Type' has undergone one-hot encoding, such that direct comparisons of sample and complete distributions to those in Appendix F are less straightforward.



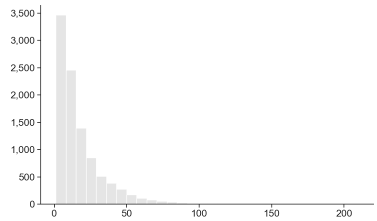
(a) Nutri-Score (Complete)



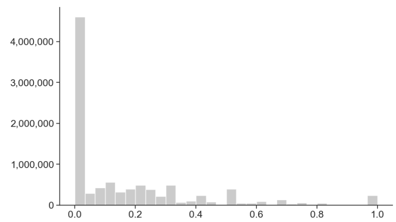
(b) Nutri-Score (Sample)



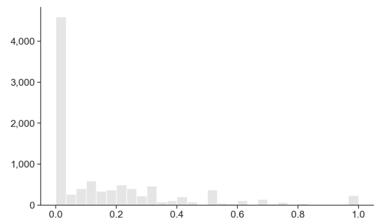
(c) Items per Basket (Complete)



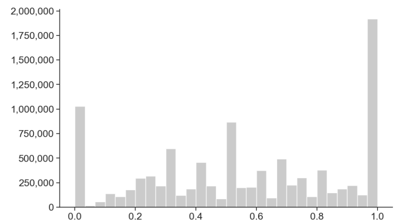
(d) Items per Basket (Sample)



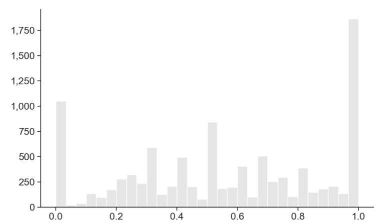
(e) Promotion Ratio (Complete)



(f) Promotion Ratio (Sample)



(g) Private Label Ratio (Complete)



(h) Private Label Ratio (Sample)

Figure 20: Distributions of Numerical Variables in Training Set: Complete (Dark-Grey) and Sample (Light-Grey)

APPENDIX I: ADDITIONAL RESULTS TEST SET

Table 17: Additional Results - Test Set

Model	Max Error	Median AE	Explained Variance
Mean Baseline	2.240	0.337	0.000
Decision Tree	2.737	0.327	0.034
XGBoost Regressor	2.867	0.325	0.043
GB Regressor	2.669	0.325	0.043
Random Forest	2.914	0.323	0.048

Examining the additional results in Table 17 provides insights into the models' performance beyond the primary evaluation metrics.

- **Max Error:** This metric indicates the maximum absolute error observed in the predictions. Lower values are desirable, reflecting reduced discrepancies between predicted and actual values. The Mean Baseline exhibits the lowest max error, indicating that it makes less extreme errors compared to the more complex ML models.

- **Median Absolute Error:** This metric represents the median value of the absolute errors, providing insight into the central tendency of prediction accuracy. The median absolute error is slightly smaller for the more complex ML models compared to the Mean Baseline, indicating that the more complex ML models tend to have smaller errors around the median.

- **Explained Variance:** This metric quantifies the proportion of variance in the target variable that the model explains. A value of 1 indicates perfect prediction, while 0 suggests that the model does not capture any variance. The explained variance is close to 0 for all models, suggesting that only a small proportion of the variance in the target variable is captured by each model.

APPENDIX J: (ADDITIONAL) RESULTS TRAINING SET

Table 18: Results - Training Set

Model	MAE	MSE	RMSE
Decision Tree	0.422	0.309	0.556
XGBoost Regressor	0.420	0.306	0.553
GB Regressor	0.420	0.306	0.553
Random Forest	0.418	0.304	0.552

Table 19: Additional Results - Training Set

Model	Max Error	Median AE	Explained Variance
Decision Tree	2.737	0.326	0.034
XGBoost Regressor	2.857	0.324	0.043
GB Regressor	2.666	0.325	0.044
Random Forest	2.952	0.323	0.049

APPENDIX K: XAI RESULTS DECISION TREE

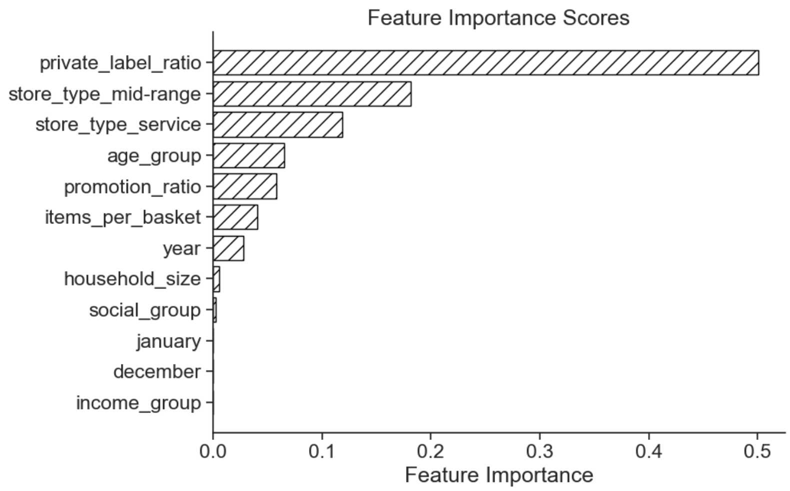


Figure 21: Feature Importance: Decision Tree

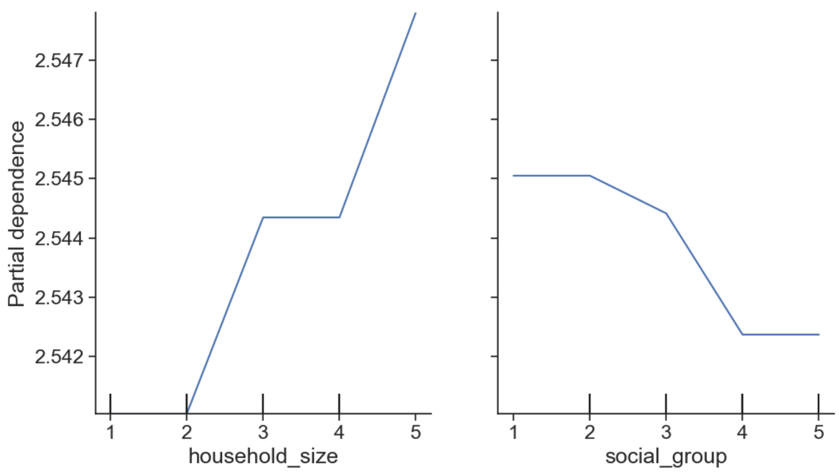


Figure 22: PDPs (Decision Tree): Household Size & Social Group

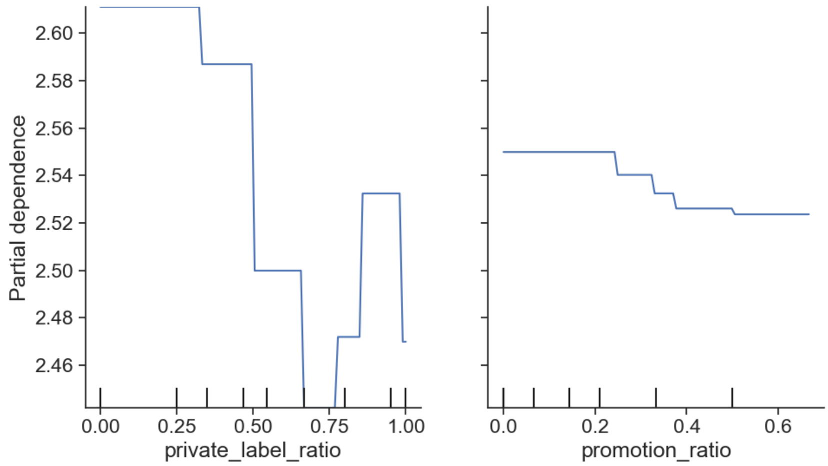


Figure 23: PDPs (Decision Tree): Private Label Ratio & Promotion Ratio

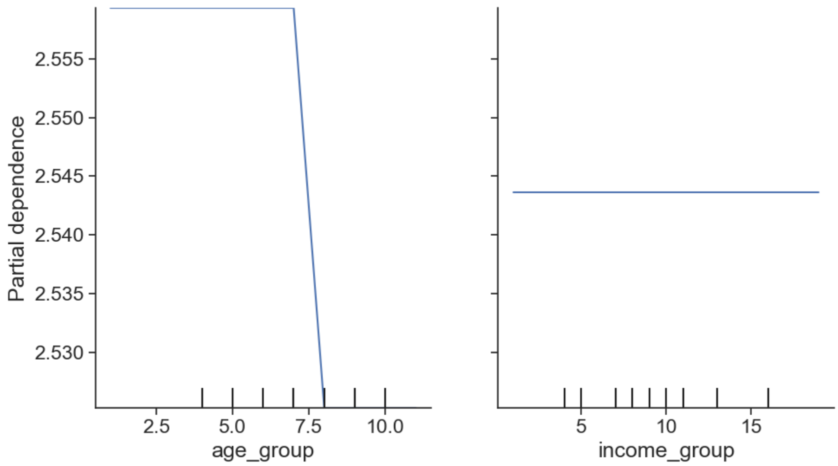


Figure 24: PDPs (Decision Tree): Age Group & Income Group

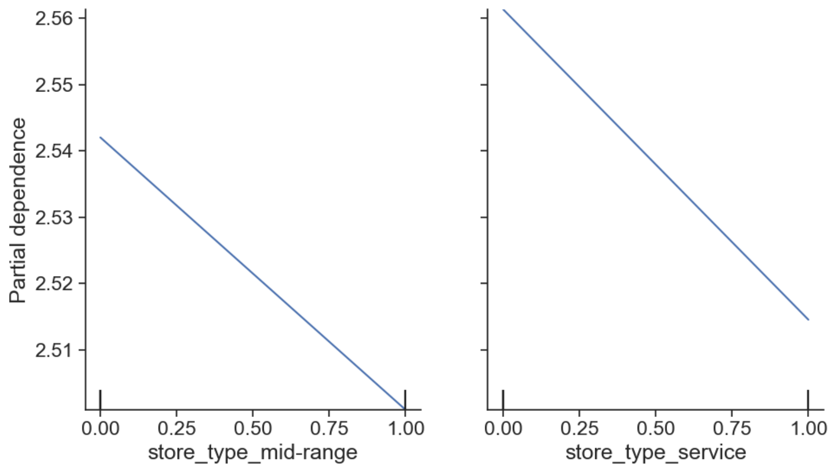


Figure 25: PDPs (Decision Tree): Store Type Mid-Range & Store Type Service

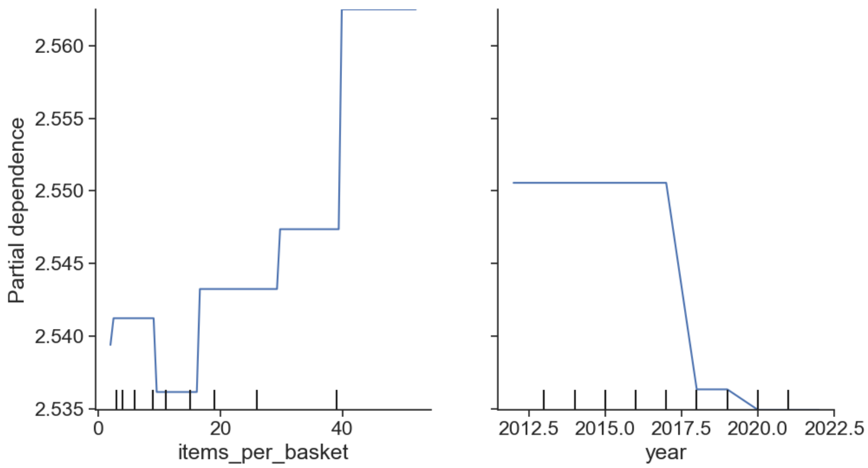


Figure 26: PDPs (Decision Tree): Items per Basket & Year

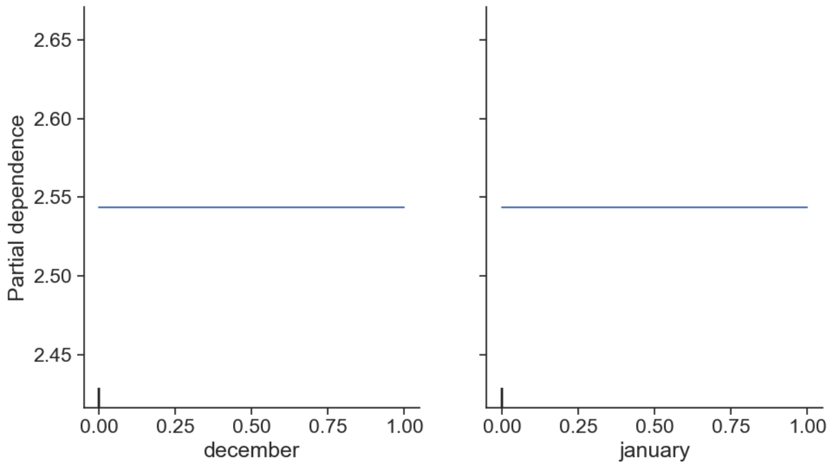


Figure 27: PDPs (Decision Tree): December & January

APPENDIX L: XAI RESULTS RANDOM FOREST

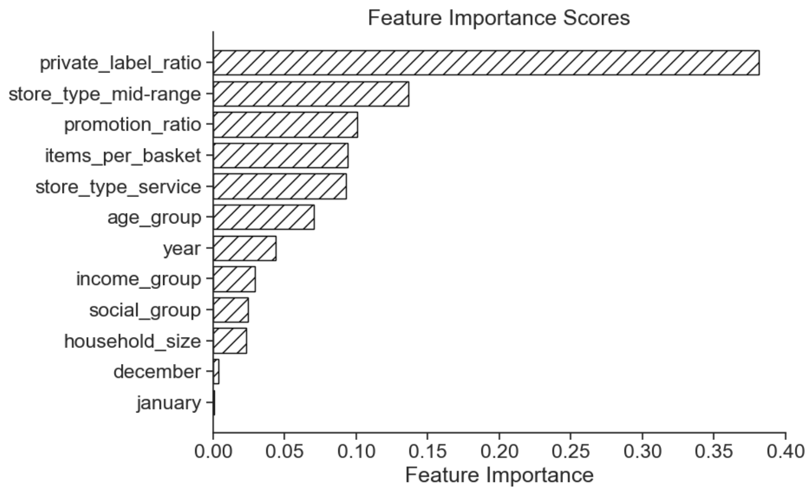


Figure 28: Feature Importance: Random Forest

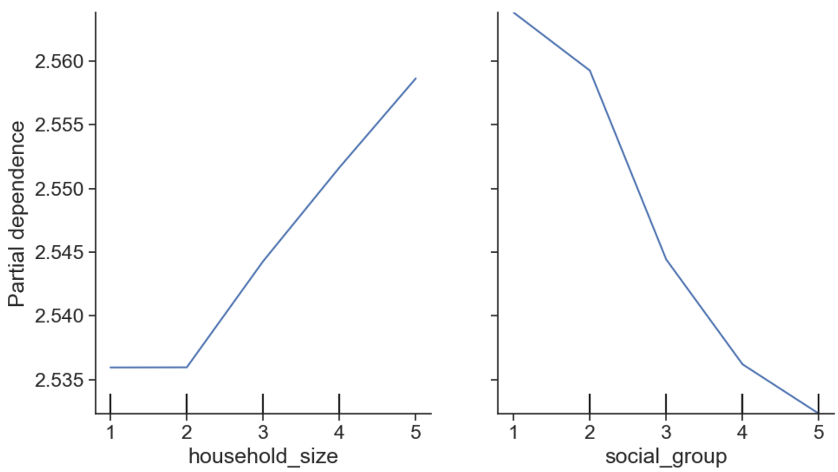


Figure 29: PDPs (Random Forest): Household Size & Social Group

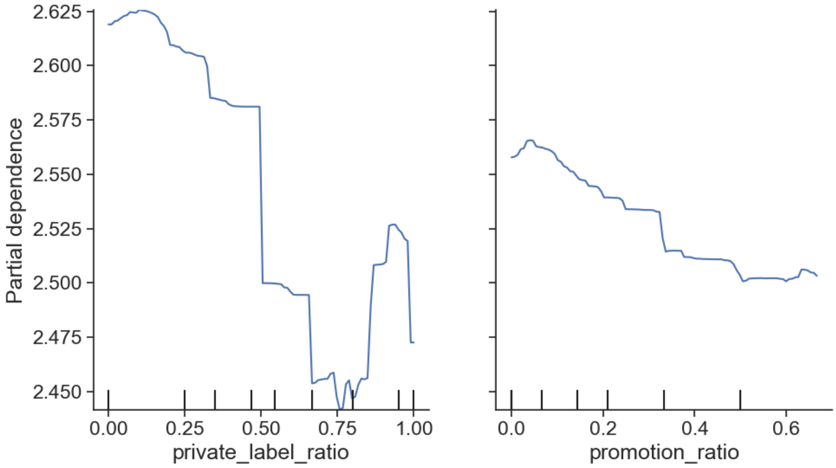


Figure 30: PDPs (Random Forest): Private Label Ratio & Promotion Ratio

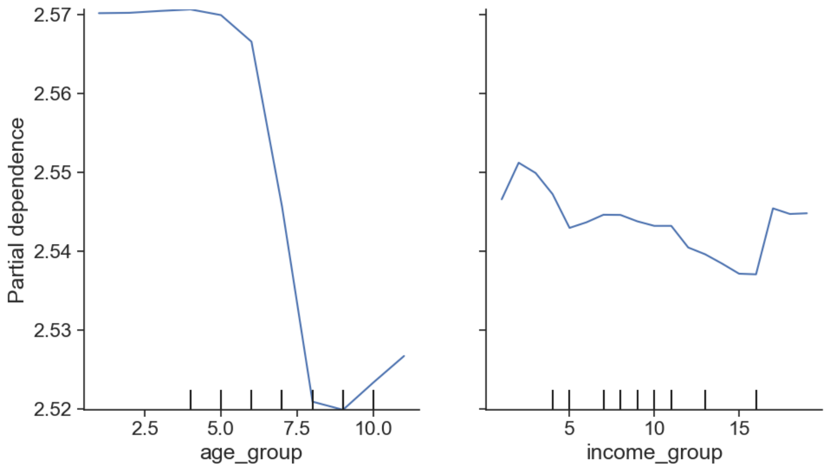


Figure 31: PDPs (Random Forest): Age Group & Income Group

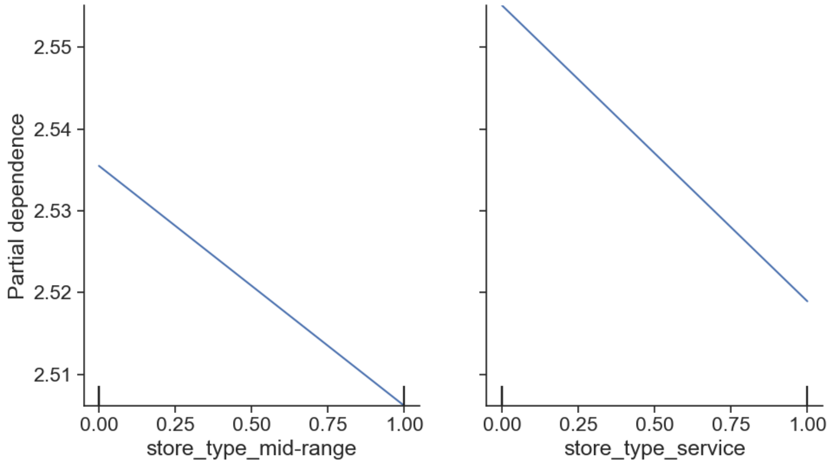


Figure 32: PDPs (Random Forest): Store Type Mid-Range & Store Type Service

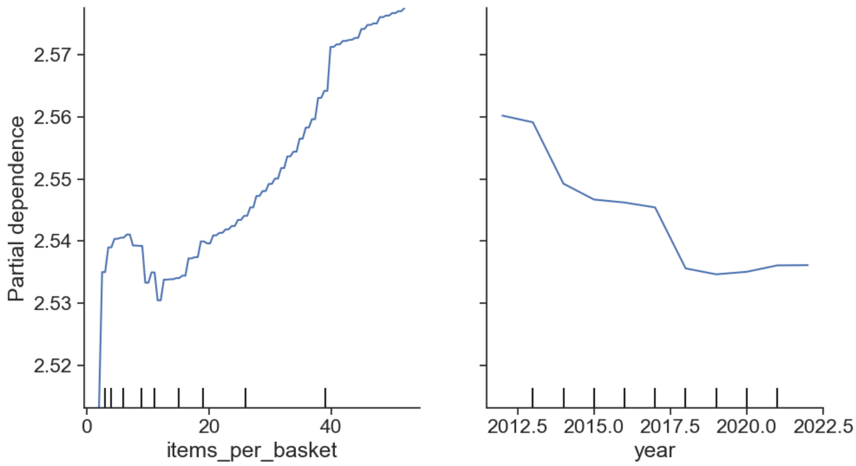


Figure 33: PDPs (Random Forest): Items per Basket & Year

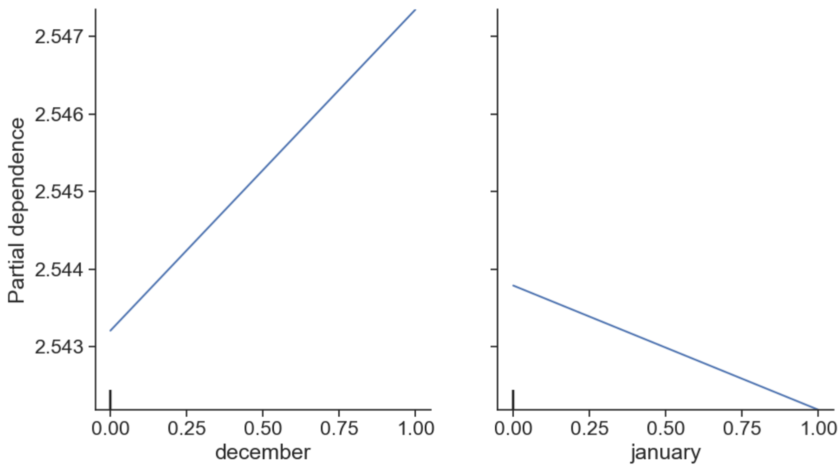


Figure 34: PDPs (Random Forest): December & January

APPENDIX M: XAI RESULTS GRADIENT BOOSTING REGRESSOR

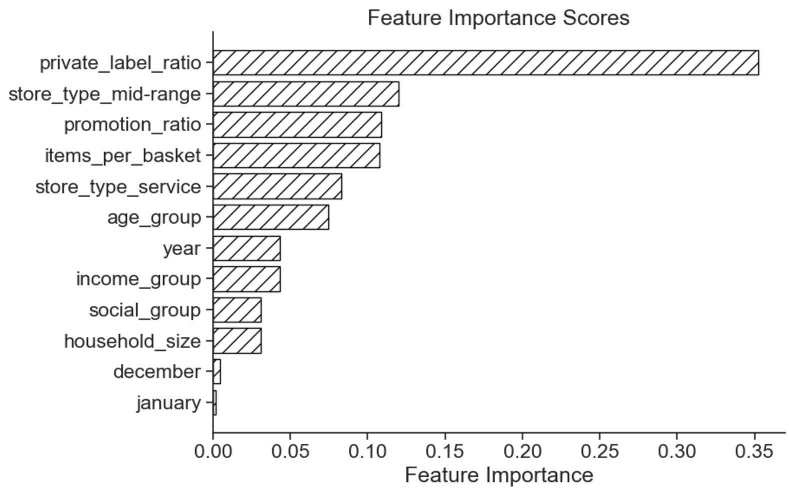


Figure 35: Feature Importance: Gradient Boosting Regressor

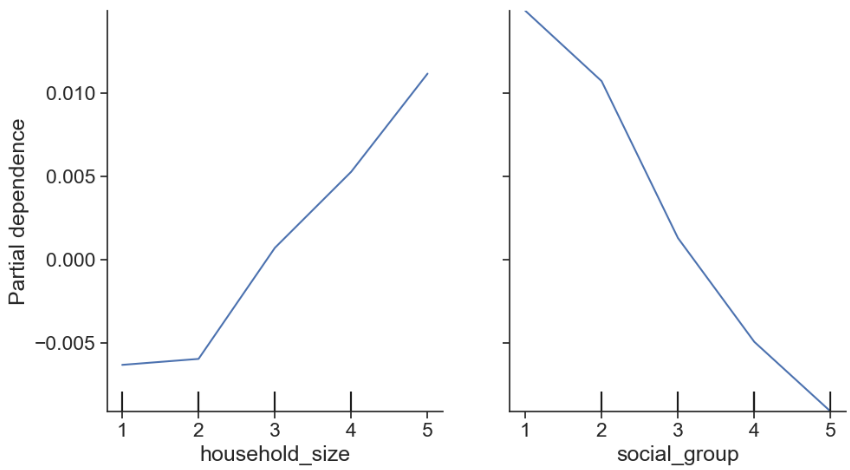


Figure 36: PDPs (Gradient Boosting Regressor): Household Size & Social Group

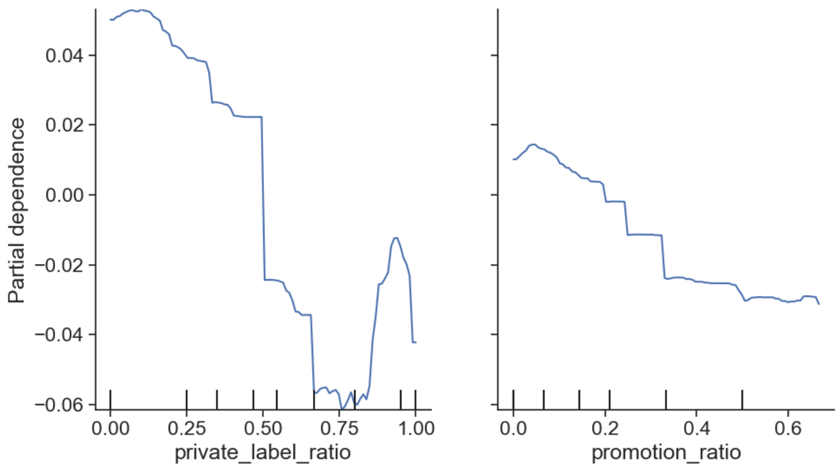


Figure 37: PDPs (Gradient Boosting Regressor): Private Label Ratio & Promotion Ratio

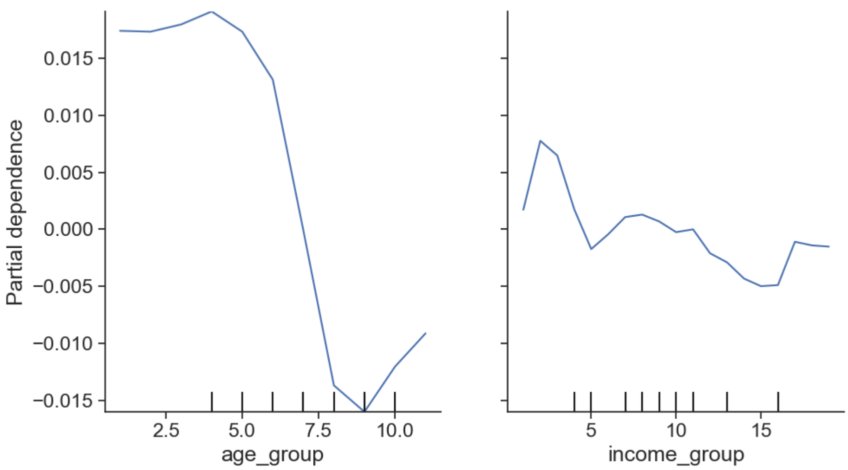


Figure 38: PDPs (Gradient Boosting Regressor): Age Group & Income Group

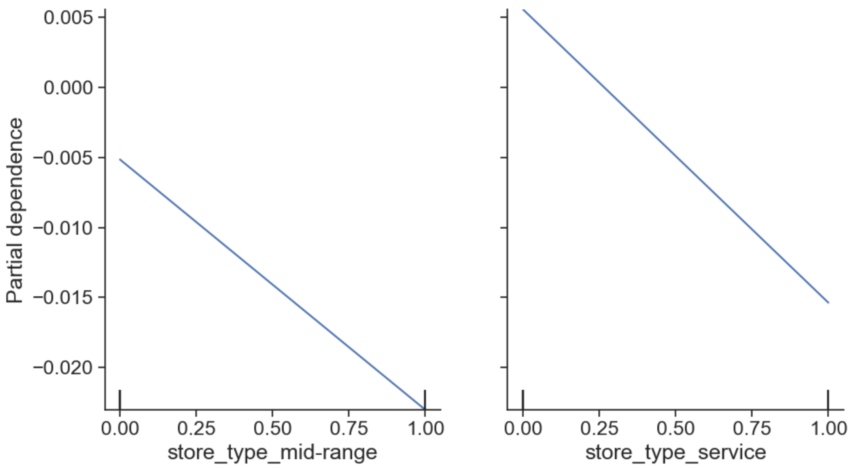


Figure 39: PDPs (Gradient Boosting Regressor): Store Type Mid-Range & Store Type Service

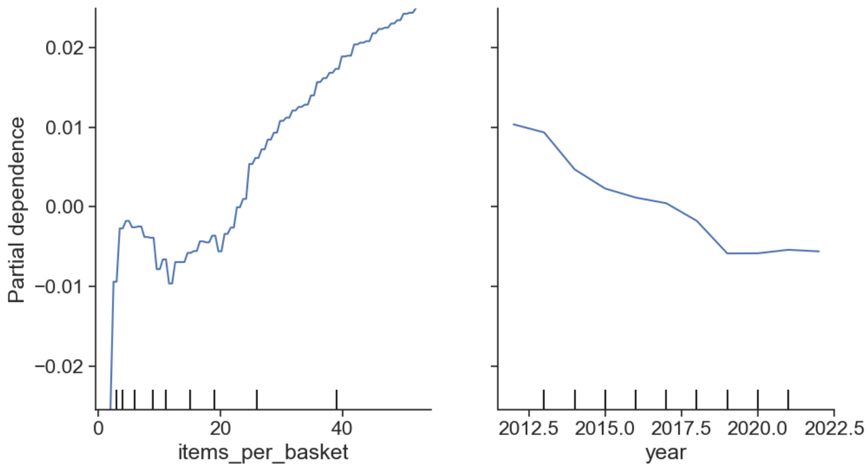


Figure 40: PDPs (Gradient Boosting Regressor): Items per Basket & Year

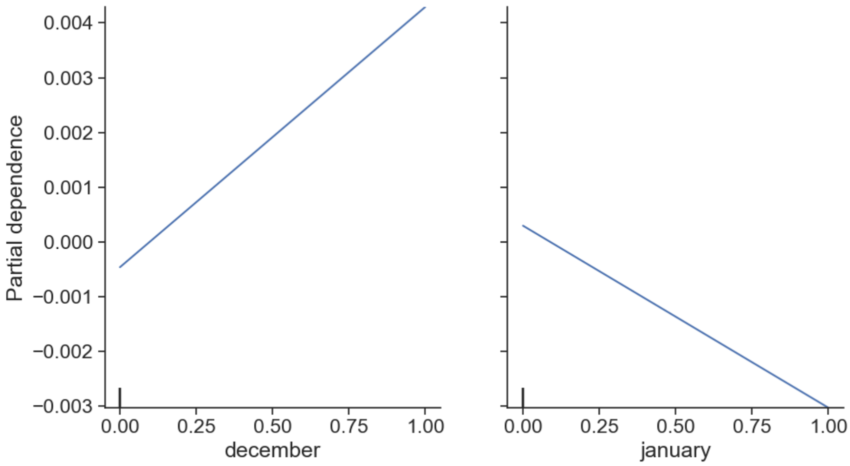


Figure 41: PDPs (Gradient Boosting Regressor): December & January

APPENDIX N: ADDITIONAL XAI RESULTS (XGBOOST REGRESSOR)

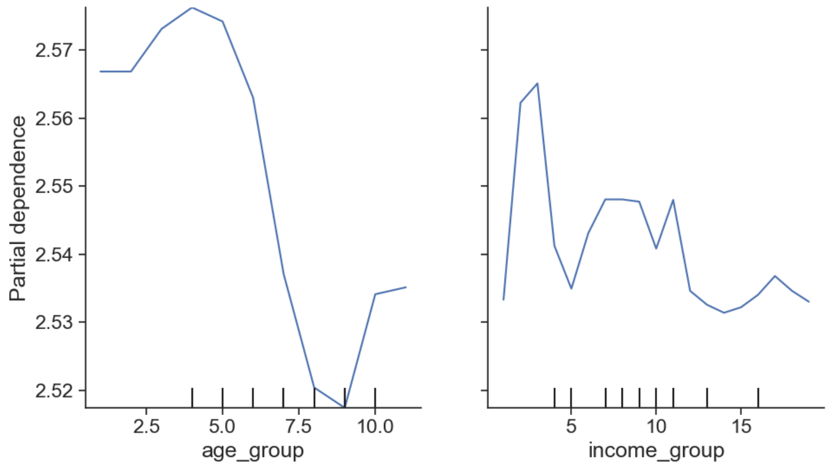


Figure 42: PDPs (XGBoost Regressor): Age Group & Income Group

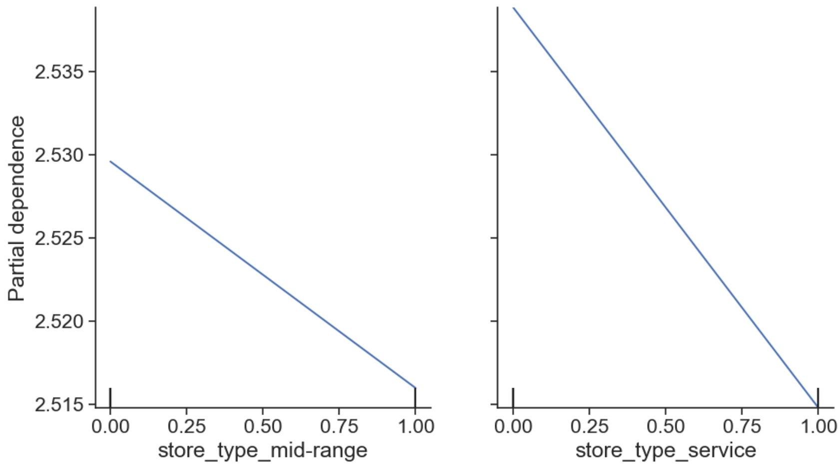


Figure 43: PDPs (XGBoost Regressor): Store Type Mid-Range & Store Type Service

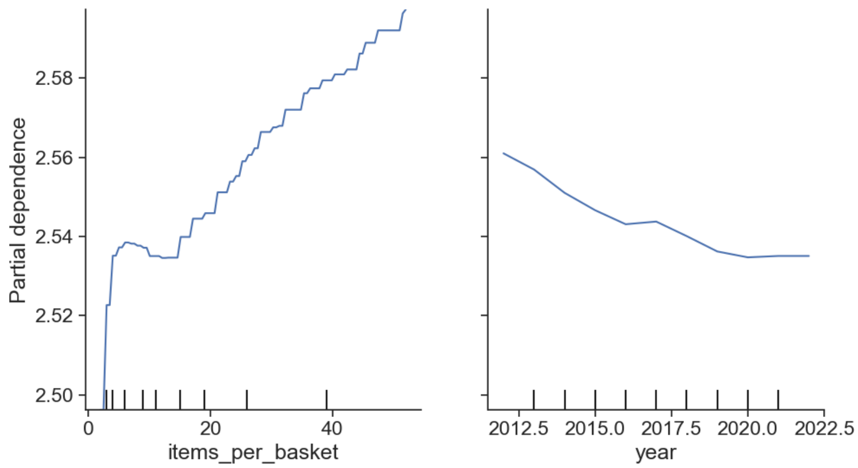


Figure 44: PDPs (XGBoost Regressor): Items per Basket & Year

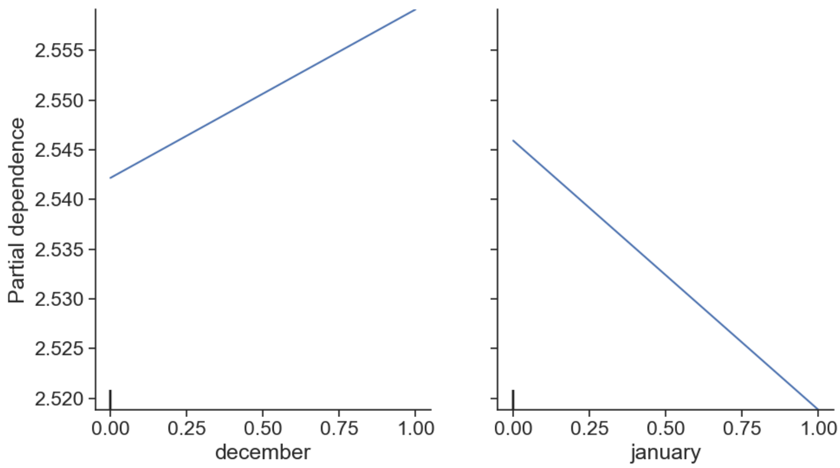


Figure 45: PDPs (XGBoost Regressor): December & January

APPENDIX O: ERROR ANALYSIS

In Figure 46, the average prediction errors of the XGBoost model are visually represented across various ranges of actual values. The horizontal dashed line at $y = 0$ acts as a baseline, indicating no prediction error. The figure clearly illustrates the model's suboptimal performance, particularly evident for more extreme values where the prediction error is large, such as in the range of 4.5 to 5.

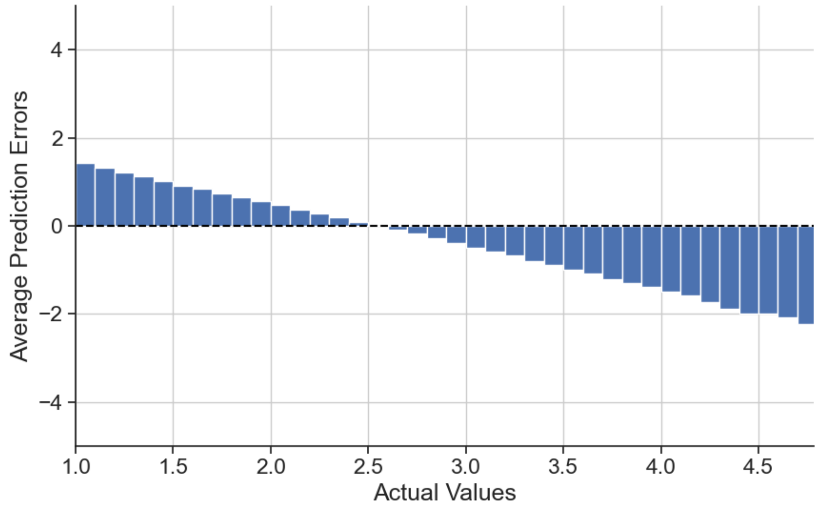


Figure 46: Average Prediction Errors (XGBoost Regressor)

Additionally, Figure 47 displays the actual and predicted values for the initial 1,000 data points. As observed in the plot, the model consistently predicts values in the range of approximately 2 to 3 for each instance. While there is some variation in predicted values, deviating from a constant average prediction (explaining the slight improvement over the blind baseline model used in this study), the observed variation is not sufficient. The plot is limited to the first 1,000 data points to avoid a cluttered illustration, but this pattern in predicted and actual values persists across the entire dataset.

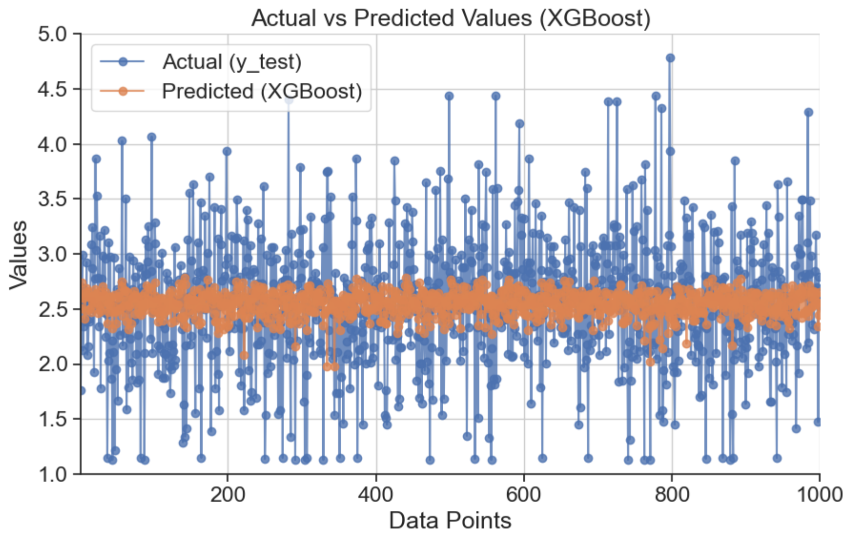


Figure 47: Actual vs. Predicted Values (XGBoost Regressor)

The mentioned observations raise concerns regarding the model's potential inaccuracies in predicting or classifying (un)healthy shopping baskets. To further explore this issue, two confusion matrices are generated: one for baskets categorized as relatively healthy and another for those considered relatively unhealthy. It is important to note that these matrices correspond to two distinct classification tasks, each determined by a different threshold. The previously established thresholds of ≤ 2 and ≥ 3 for respectively healthy and unhealthy shopping baskets are used. This implies that all shopping baskets with an average Nutri-Score of 2 or lower should be classified as 'healthy', while those with an average Nutri-Score of 3 or higher should be classified as 'unhealthy'.

Regarding the classification of healthy baskets, if a shopping basket is not categorized as healthy, it falls into either the neutral category (*i.e.*, an average Nutri-Score between 2 and 3) or the unhealthy category (*i.e.*, an average Nutri-Score of 3 or higher). For the Healthy class with a threshold of ≤ 2, the model exhibited a False Negatives rate of approximately 100%, with 628,722 instances actually healthy misclassified as not healthy (*i.e.*, either neutral or unhealthy). True Positives were limited to 2,026, reflecting instances correctly predicted as healthy. The vast majority of instances are baskets that are correctly classified as being not healthy (*i.e.*, either neutral or unhealthy).

Confusion Matrix Healthy (Threshold ≤ 2)

		Model Predictions		
		Positive	Negative	Total
Actual Values	Positive	$2,026$	$628,722$	$630,748$
	Negative	$1,056$	$3,604,906$	$3,605,962$
	Total	$3,082$	$4,233,628$	$4,236,710$

Regarding the classification of unhealthy baskets, if a shopping basket is not categorized as unhealthy, it falls into either the neutral category (*i.e.*, an average Nutri-Score between 2 and 3) or the healthy category (*i.e.*, an average Nutri-Score of 2 or lower). When considering the Unhealthy class with a threshold of ≥ 3, the model displayed a 100% False Negative rate, predicting all instances actually unhealthy as not unhealthy (*i.e.*, either neutral or healthy). This highlights a significant underestimation of unhealthy baskets, potentially impacting the model's ability to provide accurate and meaningful insights into the nutritional quality of the baskets.

Confusion Matrix Unhealthy (Threshold ≥ 3)

		Model Predictions		
		Positive	Negative	Total
Actual Values	Positive	0	$786,638$	$786,638$
	Negative	0	$3,450,072$	$3,450,072$
	Total	0	$4,236,710$	$4,236,710$