Predictive Modeling of Blood Pressure Categories: Integrating Demographic and Dietary Factors for Personalized Management

A Preprint

Li Yuan
University of Michigan
Ann Arbor, MI
leeyuan@umich.edu

December 8, 2023

Abstract

This study delves into predictive modeling of blood pressure categories, focusing on the United States, addressing the global health concern of hypertension. Mainly utilizing demographic and dietary data from the CDC National Health and Nutrition Examination Survey (NHANES) 2017-2018, aims to craft personalized management strategies. Drawing on research emphasizing the multifaceted determinants of hypertension, we leverage the multinomial regression model with lasso regularization as a baseline. Furthermore, the study advances to the extreme gradient boosting (XGB) algorithm, achieving a slightly better performance than multinomial regression. Evaluation metrics include accuracy and Area Under the Curve (AUC) in a 10-fold cross-validation framework. The study provides possible personal blood pressure management solution.

Keywords blood pressure · multinomial regression · lasso · tree · gradient boosting · xgboost · R

1 Introduction

Hypertension, a prevalent and frequently asymptomatic health condition, persists as a global health concern, significantly contributing to the burden of cardiovascular diseases (Forouzanfar et al. 2016). Against the backdrop of recent advancements in data science and machine learning, this paper initiates a meticulous investigation into the predictive modeling of blood pressure categories, specifically concentrating on the United States. The study underscores the importance of integrating demographic and dietary information to formulate personalized management strategies tailored to the unique health landscape of the U.S.

Contemporary research on hypertension highlights the multifaceted nature of its determinants, necessitating a comprehensive approach to prediction and management. A seminal study by Iqbal et al. (2021) emphasizes the significance of demographic factors in predicting hypertension prevalence, underscoring the need for nuanced models that account for individual characteristics. Additionally, the work of Johnson et al. (2009) advocates for a personalized approach, accentuating the substantial influence of dietary habits on blood pressure regulation.

The dataset utilized in this study originates from the Centers for Disease Control and Prevention (CDC) National Health and Nutrition Examination Survey (NHANES), focusing specifically on the years 2017-2018. Our dataset encompasses a comprehensive array of variables, including blood pressure measurements, demographic profiles, nutrient intakes, diabetes indicators, and health insurance details. By integrating these diverse factors, the study aims to delve into the intricate relationships between various determinants and blood pressure outcomes, ultimately seeking to develop a predictive model tailored to the specific context of blood pressure management in the United States.

As a baseline, we employ multinomial regression with lasso regularization and 10-fold cross-validation. This choice is motivated by the desire for a robust baseline that reflects the complexities of medical datasets, particularly in disease prediction scenarios. Multinomial regression, incorporating lasso regularization, brings advantages such as interpretability and the ability to handle diverse data characteristics. In further pursuit, inspired by Islam et al. (2023) in their research on hypertension modeling in Ethiopia, we leverage the extreme gradient boosting (XGB) algorithm to surpass the multinomial regression with lasso regularization baseline. The adoption of these methodologies aligns with the evolving landscape of predictive modeling in hypertension research. With a focus on predicting blood pressure categories (normal, elevated, and high), we evaluate these models based on their accuracy and Area Under the Curve (AUC), calculated using 10-fold cross-validation.

As we embark on this study, we draw from a rich tapestry of existing research to contribute novel insights into the predictive modeling of blood pressure categories. Our aim extends beyond advancing the technical aspects of machine learning applications; we seek to provide practical and personalized strategies for hypertension management, aligning with the evolving landscape of precision medicine in cardiovascular health.

2 Data

2.1 Overview

According to the Centers for Disease Control and Prevention (2023), NHANES field operations were suspended in March 2020 because of COVID-19. Consequently, data collection for the NHANES 2019-2020 cycle was incomplete, rendering it non-nationally representative. In response to the disruption caused by the COVID-19 pandemic, we only use those data collected in the 2017-2018 cycle to ensure the study's relevance and generalizability to the U.S. civilian non-institutionalized population.

NHANES employs a complex, multistage probability design for sampling the civilian, noninstitutionalized population in the U.S. In 2017-2018, 16,211 persons were selected from 30 survey locations, with 9,254 completing interviews and 8,704 undergoing examinations. Each participant has a unique identification number `SEQN`. To ensure representation, materials were translated into various languages, and cultural competency training was provided to staff (Centers for Disease Control and Prevention 2020a).

In the context of this study, the most important data we selected is the examination data of blood pressure (`BPX_J`), which "provides data for three consecutive blood pressure (BP) measurements and other methodological measurements to obtain an accurate BP. Heart rate or pulse, depending on age, are also reported (Centers for Disease Control and Prevention 2020b)." This data contains 4 readings of systolic blood pressure and 4 readings of diastolic blood pressure for each participant. In order to create a response variable about blood pressure level (`BPXLEVEL`), we first average the 4 readings of systolic blood pressure and diastolic blood pressure of each participant respectively. Then we follow the definition of normal, elevated, and hypertension provided by Centers for Disease Control and Prevention (2021) to divide our average systolic blood pressure and diastolic blood pressure into three blood pressure levels shown in table 1.

Table 1: Blood Pressure Levels Divided by Systolic and Diastolic Blood Pressure

Blood Pressure Levels	Systolic Blood Pressure		Diastolic Blood Pressure
Normal (`BPXLEVEL` = 0)	$<$ 120 mmHg	and	$<$ 80 mmHg
Elevated (`BPXLEVEL` = 1)	120-129 mmHg	and	$<$ 80 mmHg
Hypertension (`BPXLEVEL` = 2)	$\geq$ 130 mmHg	or	$\geq$ 80 mmHg

After getting the blood pressure levels (`BPXLEVEL`), we merged four other data from the NHANES, including Demographic Variables and Sample Weights (`DEMO_J`), Dietary Interview - Total Nutrient Intakes, First Day (`DR1TOT_J`), Diabetes (`DIQ_J`), and Health Insurance (`HIQ_J`), based on those participants' unique identification number `SEQN`.

Here are some description of these data:

- The Demographic Variables and Sample Weights (`DEMO_J`) data "provides individual, family, and household-level information (Centers for Disease Control and Prevention 2020c)."
- The Dietary Interview - Total Nutrient Intakes, First Day (`DR1TOT_J`) data contains "detailed dietary intake information from NHANES participants. The dietary intake data are used to estimate the

types and amounts of foods and beverages (including all types of water) consumed during the 24-hour period prior to the interview (midnight to midnight), and to estimate intakes of energy, nutrients, and other food components from those foods and beverages (Centers for Disease Control and Prevention 2020f)."

- The Diabetes (`DIQ_J`) data "provides personal interview data on diabetes, prediabetes, use of insulin or oral hypoglycemic medications, and diabetic retinopathy (Centers for Disease Control and Prevention 2020d)."
- The Health Insurance (`HIQ_J`) data "provides respondent-level interview data on insurance coverage, type of insurance coverage, coverage of prescription drugs, and uninsured status during the past 12 months (Centers for Disease Control and Prevention 2020e)."

By merging data, selecting relevant predictors, and removing some of blank data entries, we got a curated data frame with 6,125 observations and 80 variables. Detail about these 80 variables are shown in table 2.

Table 2: Variable Names and Labels in the Curated Data Frame

Name	Label	Source	Name	Label	Source
BPXLEVEL	Blood pressure levels	Derived	BPACSZ	Coded cuff size	BPX_J
BPXPLS	60 sec. pulse	BPX_J	BPXPTY	Pulse type	BPX_J
RIAGENDR	Gender	DEMO_J	RIDAGEYR	Age in years at screening	DEMO_J
RIDRETH3	Race / Hispanic origin w/ NH Asian	DEMO_J	DMDHHSIZ	Total number of people in the Household	DEMO_J
DMDHHSZA	Number of children 5 years or younger in HH	DEMO_J	DMDHHSZB	Number of children 6-17 years old in HH	DEMO_J
DMDHHSZE	Number of adults 60 years or older in HH	DEMO_J	DR1TNUMF	Number of foods / beverages reported	DR1TOT_J
DR1TKCAL	Energy (kcal)	DR1TOT_J	DR1TPROT	Protein (gm)	DR1TOT_J
DR1TCARB	Carbohydrate (gm)	DR1TOT_J	DR1TSUGR	Total sugars (gm)	DR1TOT_J
DR1TFIBE	Dietary fiber (gm)	DR1TOT_J	DR1TTFAT	Total fat (gm)	DR1TOT_J
DR1TSFAT	Total saturated fatty acids (gm)	DR1TOT_J	DR1TMFAT	Total monounsaturated fatty acids (gm)	DR1TOT_J
DR1TPFAT	Total polyunsaturated fatty acids (gm)	DR1TOT_J	DR1TCHOL	Cholesterol (mg)	DR1TOT_J
DR1TATOC	Vitamin E as alpha-tocopherol (mg)	DR1TOT_J	DR1TATOA	Added alpha-tocopherol (Vitamin E) (mg)	DR1TOT_J
DR1TRET	Retinol (mcg)	DR1TOT_J	DR1TVARA	Vitamin A, RAE (mcg)	DR1TOT_J
DR1TACAR	Alpha-carotene (mcg)	DR1TOT_J	DR1TBCAR	Beta-carotene (mcg)	DR1TOT_J
DR1TCRYP	Beta-cryptoxanthin (mcg)	DR1TOT_J	DR1TLYCO	Lycopene (mcg)	DR1TOT_J
DR1TLZ	Lutein + zeaxanthin (mcg)	DR1TOT_J	DR1TVB1	Thiamin (Vitamin B1) (mg)	DR1TOT_J
DR1TVB2	Riboflavin (Vitamin B2) (mg)	DR1TOT_J	DR1TNIAC	Niacin (mg)	DR1TOT_J
DR1TVB6	Vitamin B6 (mg)	DR1TOT_J	DR1TFOLA	Total folate (mcg)	DR1TOT_J
DR1TFA	Folic acid (mcg)	DR1TOT_J	DR1TFF	Food folate (mcg)	DR1TOT_J
DR1TFDFE	Folate, DFE (mcg)	DR1TOT_J	DR1TCHL	Total choline (mg)	DR1TOT_J
DR1TVB12	Vitamin B12 (mcg)	DR1TOT_J	DR1TB12A	Added vitamin B12 (mcg)	DR1TOT_J
DR1TVC	Vitamin C (mg)	DR1TOT_J	DR1TVD	Vitamin D (D2 + D3) (mcg)	DR1TOT_J
DR1TVK	Vitamin K (mcg)	DR1TOT_J	DR1TCALC	Calcium (mg)	DR1TOT_J
DR1TPHOS	Phosphorus (mg)	DR1TOT_J	DR1TMAGN	Magnesium (mg)	DR1TOT_J
DR1TIRON	Iron (mg)	DR1TOT_J	DR1TZINC	Zinc (mg)	DR1TOT_J
DR1TCOPP	Copper (mg)	DR1TOT_J	DR1TSODI	Sodium (mg)	DR1TOT_J
DR1TPOTA	Potassium (mg)	DR1TOT_J	DR1TSELE	Selenium (mcg)	DR1TOT_J
DR1TCAFF	Caffeine (mg)	DR1TOT_J	DR1TTHEO	Theobromine (mg)	DR1TOT_J
DR1TALCO	Alcohol (gm)	DR1TOT_J	DR1TMOIS	Moisture (gm)	DR1TOT_J
DR1TS040	SFA 4:0 (Butanoic) (gm)	DR1TOT_J	DR1TS060	SFA 6:0 (Hexanoic) (gm)	DR1TOT_J
DR1TS080	SFA 8:0 (Octanoic) (gm)	DR1TOT_J	DR1TS100	SFA 10:0 (Decanoic) (gm)	DR1TOT_J

Table 2 (Continue): Variable Names and Labels in the Curated Data Frame

Name	Label	Source	Name	Label	Source
DR1TS120	SFA 12:0 (Dodecanoic) (gm)	DR1TOT_J	DR1TS140	SFA 14:0 (Tetradecanoic) (gm)	DR1TOT_J
DR1TS160	SFA 16:0 (Hexadecanoic) (gm)	DR1TOT_J	DR1TS180	SFA 18:0 (Octadecanoic) (gm)	DR1TOT_J
DR1TM161	MFA 16:1 (Hexadecenoic) (gm)	DR1TOT_J	DR1TM181	MFA 18:1 (Octadecenoic) (gm)	DR1TOT_J
DR1TM201	MFA 20:1 (Eicosenoic) (gm)	DR1TOT_J	DR1TM221	MFA 22:1 (Docosenoic) (gm)	DR1TOT_J
DR1TP182	PFA 18:2 (Octadecadienoic) (gm)	DR1TOT_J	DR1TP183	PFA 18:3 (Octadecatrienoic) (gm)	DR1TOT_J
DR1TP184	PFA 18:4 (Octadecatetraenoic) (gm)	DR1TOT_J	DR1TP204	PFA 20:4 (Eicosatetraenoic) (gm)	DR1TOT_J
DR1TP205	PFA 20:5 (Eicosapentaenoic) (gm)	DR1TOT_J	DR1TP225	PFA 22:5 (Docosapentaenoic) (gm)	DR1TOT_J
DR1TP226	PFA 22:6 (Docosahexaenoic) (gm)	DR1TOT_J	DIQ010	Doctor told you have diabetes	DIQ_J
DIQ050	Taking insulin now	DIQ_J	HIQ011	Covered by health insurance	HIQ_J

2.2 Visualization

In this section, we present two scatterplot matrices that provide a comprehensive visual exploration of the dataset. The first matrix focuses on demographic information, offering insights into the relationships and distributions among key demographic variables. The second matrix encompasses macronutrient intakes. These visualizations aim to reveal potential patterns, correlations, and trends within the dataset.

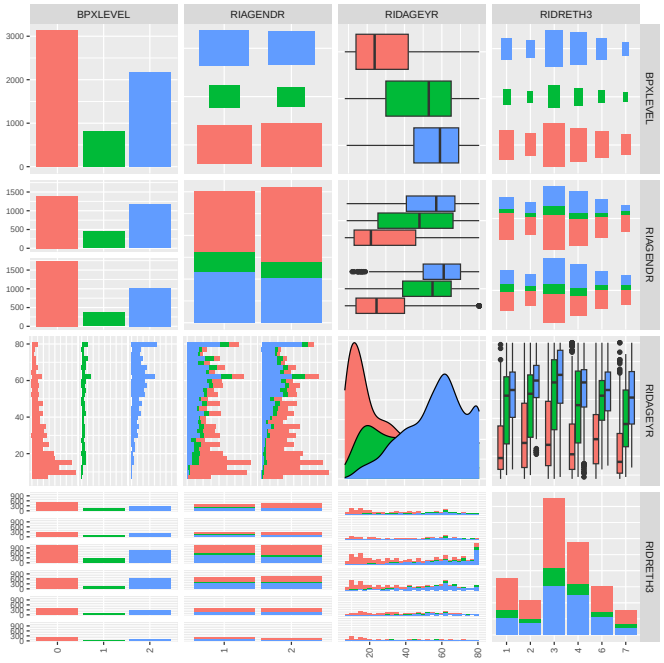

Figure 1: Scatterplot Matrix of BPXLEVEL Against Some Demographic Information

In Figure 1, we utilized a color-coded scheme to represent different blood pressure levels: red for normal, green for elevated, and blue for hypertension. By examining the relationship between blood pressure levels

(BPXLEVEL) and gender (RIAGENDR), noteworthy patterns emerge. The plot reveals a higher prevalence of elevated blood pressure and hypertension among male participants (coded as 1) compared to their female counterparts (coded as 2).

Further exploration of blood pressure levels against age (RIAGEYR) exposes intriguing insights. The distributions indicate a skewed pattern, with individuals younger than 20 predominantly exhibiting normal blood pressure levels. However, a concerning trend is observed among those around 60 years old, who are more likely to have hypertension. Age emerges as a potential influential factor for predicting blood pressure levels in future models.

Analyzing blood pressure levels against race (RIDRETH3) unveils distinct prevalence rates. Non-Hispanic White individuals (coded as 3) demonstrate the highest incidence of hypertension, followed by Non-Hispanic Black (coded as 4), Mexican American (coded as 1), and Non-Hispanic Asian (coded as 6) individuals. Categories 2 and 7, representing other Hispanic and other races (including multi-racial), exhibit the lowest hypertension cases.

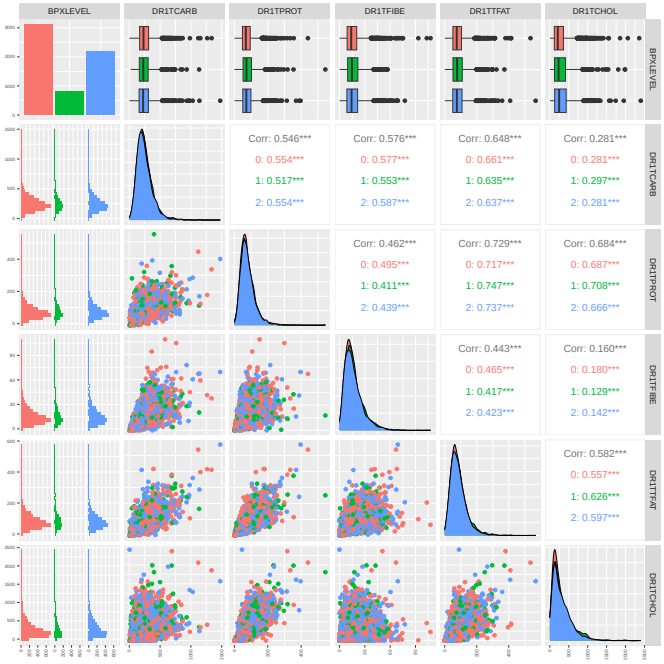

Figure 2: Scatterplot Matrix of BPXLEVEL Against Marco Nutrient Intakes

Figure 2 presents a scatterplot matrix investigating the potential impact of macronutrient intake, including carbohydrates, proteins, fats, and cholesterol (United States Department of Agriculture 2022), on blood pressure levels. Histograms of macronutrient distributions across all three blood pressure levels reveal right-skewed patterns, suggesting no single macronutrient significantly influences blood pressure.

Notably, the analysis highlights substantial correlations among macronutrient variables. The highest correlation is observed between protein intake (DR1TPROT) and dietary fat intake (DR1TTFAT), reaching 0.729. Additional pairs, such as protein intake (DR1TPROT) and cholesterol intake (DR1TCHOL) with a correlation of 0.684, indicate potential multicollinearity among predictor variables. This observation prompts caution when employing certain parametric modeling methods, such as multinomial regression, which may be sensitive to multicollinearity issues.

These findings lay the groundwork for a nuanced understanding of the dataset and underscore the importance of considering demographic and nutritional factors in predicting blood pressure levels. Subsequent sections will delve deeper into statistical analyses and modeling techniques to derive actionable insights from the presented visualizations.

3 Methods

3.1 Data Preparation (One-hot encoding and train test sets spliting)

Categorical predictors often require transformation into numerical format for compatibility with many machine learning algorithms. One-hot encoding is employed to convert categorical variables, `BPACSZ` (4 levels), `BPXPTY` (2 levels), `RIAGENDR` (2 levels), `RIDRETH3` (6 levels), `DIQ010` (4 levels), `DIQ050` (3 levels), and `HIQ011` (4 levels). This technique ensures that the categorical nature of the variables is preserved in the analysis. By applying one-hot encoding to these categorical predictors, our data frame have 91 columns of predictors in total.

In addition to one-hot encoding, we randomly selected 80% of our data (4,900 observations) as the training set without replacement and the rest 20% of the data (1,225 observations) as the testing set. In this way, we can evaluate our models objectively.

3.2 Multinomial Regression with Lasso Regularization and 10-Fold Cross Validation as a Baseline

To rigorously evaluate our model, we implement a 10-fold cross-validation strategy. This involves dividing the dataset into 10 subsets, training and testing the model 10 times, with each subset serving as the test set exactly once. This approach provides a robust estimate of the model's performance.

Multinomial regression with lasso regularization is chosen as the baseline algorithm for its interpretability and efficacy in handling diverse data characteristics. The model's prediction for a data point x_i is mathematically expressed as:

$$\hat{y}_i = \text{argmax}_j \left(\frac{e^{\beta_{0j}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\ldots+\beta_{mj}x_{mi}}}{\sum_{k=0}^{N-1} e^{\beta_{0k}+\beta_{1k}x_{1i}+\beta_{2k}x_{2i}+\ldots+\beta_{mk}x_{mi}}} \right)$$

Here, Y represents the class variable, X is the feature matrix, β denotes the coefficients, and N is the number of classes. The addition of lasso regularization to the objective function introduces a penalty term:

$$\text{argmax}_\beta \sum_{i=0}^{N-1} \log P(Y = y_i|X) - \lambda \sum_{j=0}^{N-1} |\beta_j|$$

where λ is the regularization parameter. With the lasso penalty, we will be able to identify which predictors are the most influential in predicting blood pressure types.

The model's performance is systematically assessed through accuracy and Area Under the Curve (AUC) metrics across the 10-fold cross-validation. This thorough evaluation ensures the reliability of our multinomial regression baseline in predicting blood pressure categories.

3.3 XGBoost Model with 10-Fold Cross Validation

XGBoost, an abbreviation for Extreme Gradient Boosting, stands out as a powerful ensemble learning method, widely recognized for its superior predictive capabilities. The algorithm systematically constructs a collection of weak learners, often in the form of decision trees, and amalgamates their predictions to enhance accuracy and generalize well to unseen data.

In this analysis, XGBoost is strategically employed to surpass the baseline set by Multinomial regression with lasso regularization. The algorithm iteratively builds an ensemble by sequentially introducing weak learners, each correcting errors made by its predecessors. To control model complexity and improve robustness, XGBoost incorporates regularization terms.

The mathematical formulation of the XGBoost algorithm is as follows:

Given a training dataset $\{(x_i, y_i)\}_{i=1}^n$, where x_i represents the predictors of the ith observation and y_i is the corresponding label, XGBoost aims to learn an additive model $F(x)$ of the form:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

Here:

- M is the number of weak learners (trees) in the ensemble.
- $h_m(x)$ is the mth weak learner (tree).
- γ_m is the weight (shrinkage) applied to the output of the mth weak learner.

The XGBoost algorithm minimizes the following objective function, which comprises a differentiable convex loss function $L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$ and a regularization term $\Omega(F_m)$:

$$\mathcal{L}(\{(x_i, y_i)\}_{i=1}^n) = \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma_m h_m(x_i)) + \Omega(F_m)$$

Here:

- L is the loss function measuring the difference between predicted values and true labels.
- F_{m-1} is the additive model up to the (m-1)th iteration.
- γ_m is the optimal weight for the mth weak learner.
- $h_m(x_i)$ is the prediction of the mth weak learner for the ith observation.
- $\Omega(F_m)$ is a regularization term controlling the model's complexity, typically penalizing the number of leaves in the trees and the magnitude of the weights.

The objective function is optimized in a stage-wise manner. At each stage, a new weak learner is added to the ensemble by solving for γ_m and $h_m(x_i)$, updating the model accordingly:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

The optimization involves finding values for γ_m and the weak learner's parameters that minimize the objective function. This is commonly achieved using techniques like gradient boosting, iteratively fitting a weak learner to the negative gradient of the objective function.

The performance of the XGBoost model will be assessed based on accuracy and the area under the curve (AUC) metrics. Employing 10-fold cross-validation ensures robust estimation of these metrics across different data subsets, enhancing the model's reliability and generalization capabilities.

3.4 Feature Selection with XGBoost Feature Importance

One distinctive feature of XGBoost is its ability to provide valuable insights into feature importance. This is achieved through the computation of importance scores assigned to each predictor, utilizing Gain as the metric, which represents the improvement in accuracy attributed to a specific feature across the model's trees (XGBoost Developer 2022).

This analysis leverages the XGBoost-derived importance scores to discern the most influential predictors. The incorporation of these scores aims to enhance the predictive performance of the model, resulting in improved accuracy and AUC scores. The higher the importance score assigned to a feature, the more impactful it is considered in the overall predictive capacity of the model.

4 Results

4.1 Multinomial Regression Model with Lasso Regularization

The Multinomial regression model was trained with various lasso regularization strengths, spanning a range from low to high values, using a 10-fold cross-validation strategy. The model's multinomial deviance was documented for each regularization strength.

Figure 3 helps identify the optimal λ for the lasso penalty term of the multinomial regression. We aim to get the λ which minimizes the multinomial deviance. By looking at the figure, we got the smallest multinomial deviance when $\lambda = 0.0074067$ and only 43 predictors were selected by the lasso penalty, which is indicated by the vertical dash line on the left.

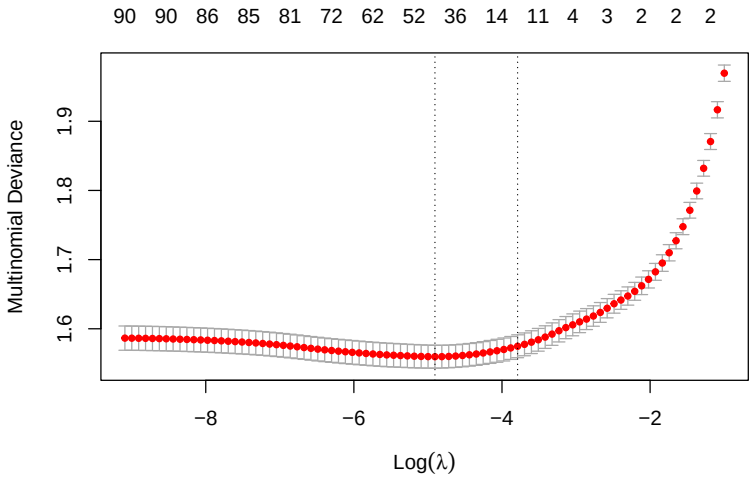

Figure 3: 10-Fold Cross Validation Findinig the Best Lasso Penalty Term

Table 3: Coefficients of Selected Predictors

Predictor	Coefficient (level 0)	Coefficient (level 1)	Coefficient (level 2)	Predictor	Coefficient (level 0)	Coefficient (level 1)	Coefficient (level 2)
Intercept	2.2521848	-0.6889857	-1.5631991	BPACSZ2	0.7865830	-0.4102865	-0.3762966
BPACSZ3	0.1767172	-0.1796646	0.0029475	BPACSZ5	-0.4005157	0.1731479	0.2273678
BPXPLS	-0.0032227	-0.0000035	0.0032262	BPXPTY2	0.1264090	0.0395048	-0.1659138
RIAGENDR2	0.2307534	-0.1249100	-0.1058434	RIDAGEYR	-0.0368332	0.0039225	0.0329107
RIDRETH32	0.0241270	-0.0054927	-0.0186343	RIDRETH33	0.1693800	-0.0519517	-0.1174283
RIDRETH34	-0.1006645	-0.0656122	0.1662767	RIDRETH36	-0.0342307	-0.0324576	0.0666883
DMDHHSZA	0.0403702	-0.0185643	-0.0218059	DMDHHSZB	0.0348970	-0.0227166	-0.0121805
DMDHHSZE	0.0189516	0.0594003	-0.0783519	DR1TNUMF	0.0064755	0.0015578	-0.0080333
DR1TSUGR	-0.0000212	-0.0001419	0.0001631	DR1TCHOL	-0.0000276	0.0000052	0.0000225
DR1TATOC	0.0000225	0.0000252	-0.0000478	DR1TACAR	-0.0000038	0.0000000	0.0000038
DR1TLYCO	0.0000002	-0.0000001	-0.0000001	DR1TNIAC	-0.0000822	-0.0000403	0.0001225
DR1TVB6	-0.0030378	-0.0047378	0.0077756	DR1TVB12	-0.0018620	0.0012349	0.0006271
DR1TVC	-0.0000052	0.0000135	-0.0000082	DR1TVD	0.0074843	-0.0026119	-0.0048724
DR1TZINC	-0.0004469	0.0008037	-0.0003569	DR1TPOTA	0.0000234	-0.0000043	-0.0000191
DR1TALCO	-0.0014905	0.0015632	-0.0000727	DR1TMOIS	-0.0000438	0.0000117	0.0000322
DR1TS060	0.1333655	-0.0404757	-0.0928898	DR1TS080	0.0251928	-0.1485159	0.1233232
DR1TS120	-0.0010639	-0.0023929	0.0034568	DR1TS180	-0.0027631	-0.0002637	0.0030268
DR1TM161	-0.0173735	0.0056004	0.0117731	DR1TM201	-0.0351797	0.0157894	0.0193903
DR1TM221	-0.0781895	0.0435119	0.0346776	DR1TP184	-0.3719599	-0.1414094	0.5133693
DR1TP204	-0.3937469	0.2857126	0.1080342	DIQ0102	0.0059360	-0.0490002	0.0430642
DIQ0502	-0.0978138	0.0281761	0.0696377	HIQ0112	-0.0250448	-0.0130697	0.0381145
HIQ0117	0.5921226	-0.3291087	-0.2630140	HIQ0119	-0.0376808	0.0126395	0.0250413

Table 3 shows all coefficients of our 43 predictors and the intercept of the model. Our multinomial model should look like:

$$\Pr(Y=0|X) = \frac{\exp(2.2521848 + 0.1767172x_{\texttt{BPACSZ3}} - 0.0032227x_{\texttt{BPXPLS}} + \cdots - 0.0376808x_{\texttt{HIQ0119}})}{\sum_{i=0}^{2}\exp(\beta_{0i} + \beta_{1i}x_{\texttt{BPACSZ3}} + \beta_{2i}x_{\texttt{BPXPLS}} + \cdots + \beta_{43i}x_{\texttt{HIQ0119}})},$$
$$\Pr(Y=1|X) = \frac{\exp(-0.6889857 - 0.1796646x_{\texttt{BPACSZ3}} - 0.0000035x_{\texttt{BPXPLS}} + \cdots + 0.0126395x_{\texttt{HIQ0119}})}{\sum_{i=0}^{2}\exp(\beta_{0i} + \beta_{1i}x_{\texttt{BPACSZ3}} + \beta_{2i}x_{\texttt{BPXPLS}} + \cdots + \beta_{43i}x_{\texttt{HIQ0119}})},$$
$$\Pr(Y=2|X) = \frac{\exp(-1.5631991 + 0.0029475x_{\texttt{BPACSZ3}} + 0.0032262x_{\texttt{BPXPLS}} + \cdots + 0.0250413x_{\texttt{HIQ0119}})}{\sum_{i=0}^{2}\exp(\beta_{0i} + \beta_{1i}x_{\texttt{BPACSZ3}} + \beta_{2i}x_{\texttt{BPXPLS}} + \cdots + \beta_{43i}x_{\texttt{HIQ0119}})},$$

where

$$\begin{aligned}
&\sum_{i=0}^{2}\exp(\beta_{0i} + \beta_{1i}x_{\texttt{BPACSZ3}} + \beta_{2i}x_{\texttt{BPXPLS}} + \cdots + \beta_{43i}x_{\texttt{HIQ0119}}) \\
=&\exp(2.2521848 + 0.1767172x_{\texttt{BPACSZ3}} - 0.0032227x_{\texttt{BPXPLS}} + \cdots - 0.0376808x_{\texttt{HIQ0119}}) + \\
&\exp(-0.6889857 - 0.1796646x_{\texttt{BPACSZ3}} - 0.0000035x_{\texttt{BPXPLS}} + \cdots + 0.0126395x_{\texttt{HIQ0119}}) + \\
&\exp(-1.5631991 + 0.0029475x_{\texttt{BPACSZ3}} + 0.0032262x_{\texttt{BPXPLS}} + \cdots + 0.0250413x_{\texttt{HIQ0119}}).
\end{aligned}$$

In our multinomial regression model, we achieved a test accuracy of 68% and the Area Under the Curve (AUC) was calculated as 0.6791, demonstrating the model's robust discriminative ability across multiple classes. These results establish a baseline for our future modeling efforts, showcasing the effectiveness of the multinomial regression approach in capturing and understanding underlying patterns within the dataset.

4.2 XGBoost Model

The XGBoost model was trained using Extreme Gradient Boosting with exact tree method, a powerful ensemble learning method. The following hyperparameters were modified and utilized in the model:

- Learning Rate (eta): 0.005
- Subsample: 0.75
- Column Subsample: 0.8
- Maximum Depth: 10
- Number of Trees (Rounds): 35

With these hyperparameters, we used 10-fold cross-validation to get a test accuracy of 68.08163% and an AUC of 0.6845. The test accuracy is 0.08163% higher than that of multinomial regression model with lasso regularization, and the test AUC is 0.0054 higher than that of multinomial model. These indicate that the XGBoost model is slightly better than multimonial regression model with lasso regularization when classifying the blood pressure levels.

4.3 XGBoost Model with Selected Predictors

Figure 4 shows the Gain scores of the predictors used in the XGBoost model. A higher bar (higher Gain score) represents more important the predictor is. Notably, key predictors such as `RIDAGEYR` (age), `DMDHHSZB` (household size), and `BPXPLS` (pulse rate) emerged as significant contributors to the predictive power of the model.

In our pursuit of refining the model and unveiling the most impactful predictors, we executed a meticulous feature selection process. We initiated this process by systematically eliminating predictors, starting with the least important (the one with the lowest Gain score), and subsequently assessed the impact on both test accuracy and AUC. This methodical stepwise elimination allowed us to pinpoint a subset of predictors that consistently upheld optimal predictive performance. During this process, we keep using the same hyperparameters we used in the original XGBoost model with 10-fold cross validation at each step.

The results of this feature selection journey revealed a compelling trade-off between the number of predictors and predictive accuracy. Significantly, in figure 5, the model showcased a remarkable test accuracy of 68.08163% and an AUC of 0.6857238 even with just the top 17 most important predictors. This underscores

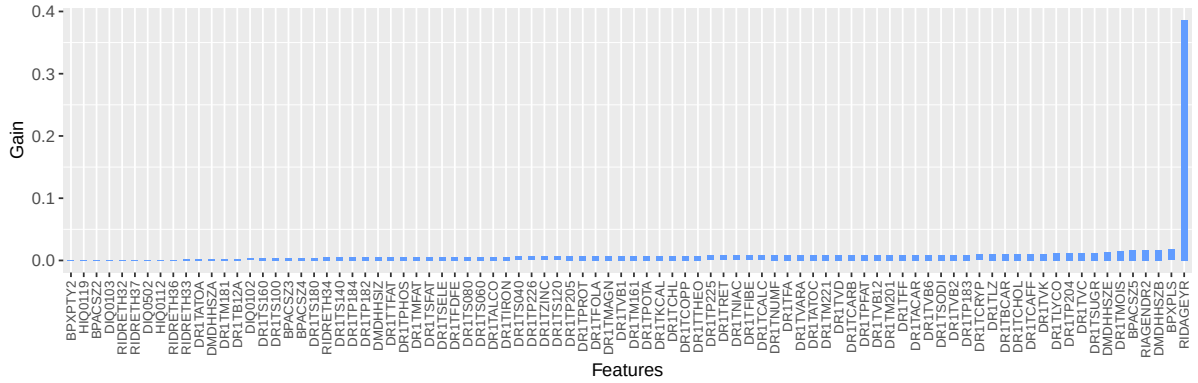

Figure 4: Bar Plots of Gain Score of Each Feature in the XGBoost Model

the efficiency of the selected predictors in encapsulating crucial information for the accurate prediction of health outcomes. As indicated by the red dash line in figure 5, the model achieved the highest test accuracy of 68.65306% and the highest AUC of 0.6887414 with the top 35 predictors.

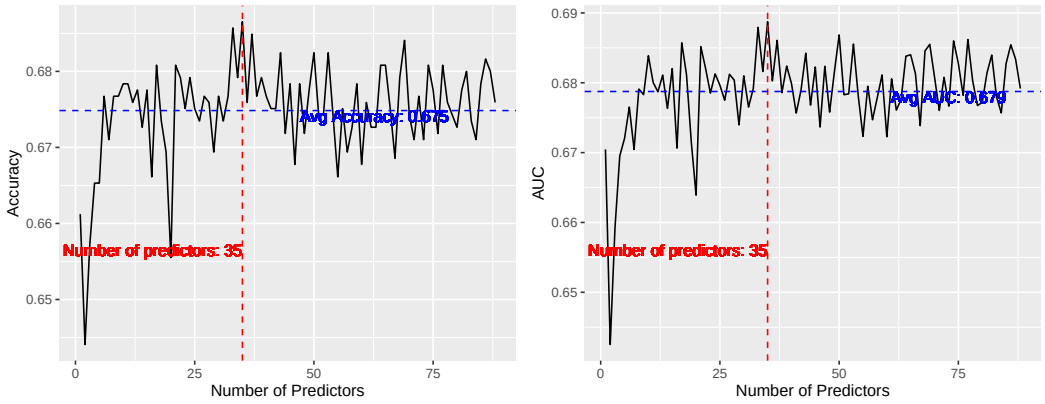

Figure 5: XGBoost Model Test Accuracy and AUC from 1 Predictor to 88 Predictors

Table 4 provides a comprehensive overview of the top predictors identified by the XGBoost model with test accuracy higher than 68.08163%, presenting their corresponding threshold Gain scores, accuracy, and AUC values. The table is thoughtfully organized, with entries sorted based on descending test accuracy, prioritizing higher accuracy models. In cases of ties, the sorting is further refined by considering descending AUC values and, if necessary, the top number of predictors in descending order.

Table 4: XGBoost Model Top Predictors and Performance Metrics

Number of Top Predictors	Threshold Gain Score	Accuracy	AUC
35	0.007949711	68.65306%	0.6887414
33	0.008132943	68.57143%	0.6879340
37	0.007893822	68.48980%	0.6860826
69	0.004570309	68.40816%	0.6855072
50	0.006381227	68.24490%	0.6868555
53	0.006278792	68.24490%	0.6855381
43	0.007424771	68.24490%	0.6842324
86	0.000273781	68.16327%	0.6854440

Table 5 provides a summary of the performance of multinomial regression model with lasso penalty, XGBoost model with all predictors, and XGBoost model with reduced predictors. As our focus lies on the accuracy and AUC metrics, and, based on these, the model with the top 35 most important predictors stands out as the preferred choice. This model exhibits higher increases in accuracy, with 0.65306% and 0.57143% improvements compared to the multinomial regression model with lasso regularization and the XGBoost model using all predictors, respectively. Moreover, it demonstrates a 0.0096414 and 0.0042414 increase in AUC compared to the multinomial regression model with lasso regularization and the XGBoost model using all predictors, respectively.

Table 5: Summary of Three Models' Test Accuracies and AUCs

Model	Test Accuracy	Test AUC
Multinomial Regression with Lasso	68.00000%	0.67910
XGBoost Using 91 Predictors	68.08163%	0.68450
XGBoost Using 35 Predictors	68.65306%	0.68874

Among these 35 selected predictors, there are 20 predictors selected by both the multinomial model and the XGBoost model. These predictors are `RIDAGEYR`, `BPXPLS`, `DMDHHSZB`, `RIAGENDR2`, `BPACSZ5`, `DR1TMOIS`, `DMDHHSZE`, `DR1TSUGR`, `DR1TVC`, `DR1TP204`, `DR1TLYCO`, `DR1TCHOL`, `DR1TVB6`, `DR1TACAR`, `DR1TM201`, `DR1TVB12`, `DR1TVD`, `DR1TM221`, `DR1TATOC`, and `DR1TNUMF`. Since both models selected these predictors, indicating the importance of these predictors on predicting blood pressure levels. The other 15 predictors selected by the XGBoost model but not the multinomial model are `DR1TVK`, `DR1TCAFF`, `DR1TBCAR`, `DR1TLZ`, `DR1TCRYP`, `DR1TP183`, `DR1TVB2`, `DR1TSODI`, `DR1TFF`, `DR1TPFAT`, `DR1TCARB`, `DR1TVARA`, `DR1TFA`, `DR1TCALC`, and `DR1TFIBE`.

Our systematic approach to feature selection not only fine-tuned the model but also provided insightful perspectives on the pivotal factors influencing its predictive power. This enhanced interpretability contributes to a more robust and effective health outcome prediction system.

5 Conclusions

Despite the advancements made in developing predictive models, it's crucial to acknowledge certain limitations. One prominent drawback is the challenge of achieving high accuracy, particularly in the context of health-related predictions. Accurate blood pressure classification is paramount for providing meaningful health insights, and any inaccuracies in predictions could have significant implications. Notably, discrepancies in predicting health outcomes can impact the reliability of personalized recommendations and interventions, potentially leading to suboptimal health management.

Numerous studies emphasize the importance of accuracy in health-related predictive models. For instance, a study by Sofogianni et al. (2022) highlighted the critical role of accurate predictions in cardiovascular risk assessment models, underscoring the potential consequences of misclassification on patient care. Additionally, research conducted by Grover and Joshi (2014) emphasized the need for robust predictive models in chronic disease management, as inaccuracies can compromise the effectiveness of preventive measures and early interventions. These findings underscore the broader concern within the scientific community about the implications of suboptimal accuracy in health-related predictions.

Addressing the aforementioned drawbacks requires a multi-faceted approach. Feature engineering, the process of refining and creating new predictors, could enhance the models' ability to capture intricate patterns in the data, potentially boosting predictive performance. Additionally, acquiring more high-quality data, especially with a focus on diverse demographic groups and health conditions, could contribute to a more comprehensive and representative model. Exploring advanced machine learning techniques, such as deep learning methods like neural networks, holds promise in uncovering complex relationships within the data, potentially elevating predictive accuracy.

An exciting application of our predictive models lies in the integration with health apps, such as Apple Health, Samsung Health, and so on. Implementing our models in these platforms could empower individuals to receive personalized daily blood pressure suggestions based on their recorded dietary intakes, known health conditions, and demographic information. To demonstrate, we deployed the multionomial regression model into a website interface at `https://bpmodel.ly.gd.edu.kg/`, where users can input their demographic, health, and dietary information to get a blood pressure prediction. This practical application could serve as a proactive tool for users to manage their health more effectively, offering real-time insights and guidance.

In conclusion, while our predictive models showcase promising results, there is ongoing work to be done in refining their accuracy and applicability. By addressing the identified drawbacks through feature engineering, data enrichment, and the exploration of advanced machine learning techniques, we can move closer to developing highly reliable and impactful predictive models for blood pressure classification. The envisioned integration with health apps presents an exciting avenue for translating our research into actionable insights, fostering proactive health management among individuals.

6 Computational Details

The analysis was conducted using R version 4.3.2 for Windows, with the utilization of various R libraries from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/` to facilitate data manipulation, statistical modeling, and visualization. The following R libraries were employed in this study:

- `caret`: for classification and regression training.
- `dplyr`: for data manipulation and summarization.
- `GGally`: for extension to ggplot2 for correlation plots.
- `ggpubr`: for creating publication-ready plots with ggplot2.
- `glmnet`: for fitting generalized linear models with regularization.
- `grid`: for arranging and combining multiple plots.
- `gridExtra`: for arranging and combining multiple plots.
- `haven`: for reading and writing SPSS, Stata, and SAS files.
- `knitr`: for dynamic report generation in R Markdown.
- `patchwork`: for arranging and combining multiple plots.
- `pROC`: for analyzing ROC curves and assessing model performance.
- `tidyr`: for data tidying and reshaping.
- `xgboost`: for extreme gradient boosting.

The analyses were conducted in the RStudio integrated development environment (IDE) version "Mountain Hydrangea" Release (583b465e, 2023-06-05) for Windows. RStudio can be downloaded at `https://posit.co/`.

An Intel-compatible 64-bit platform is preferred. At least 2048 MB of RAM is recommended to run the whole script. An operating system of Windows 7 or higher or Mac OS X 10.6 or higher is preferred.

7 Reproducibility

Ensuring the reproducibility of this study is of utmost importance. The entire analysis, including data preprocessing, model development, and result generation, is encapsulated in an RMarkdown document. The RMarkdown file, along with the necessary BibTeX and style files, has been made available on GitHub for easy access and replication: `https://github.com/lygitdata/bpmodel/`.

The RMarkdown file and its relevant files can be downloaded at the following link:

`https://bpmodel.ly.gd.edu.kg/manuscript/download.zip`

To reproduce the findings and generate the same results presented in this paper, follow these steps:

1. Download the Necessary Files:

- Navigate to the provided link in your browser.
- Unzip the downloaded file to a directory of your choice.

2. Open RMarkdown in RStudio:

- Ensure you have R and RStudio installed on your machine.
- Open RStudio and navigate to the directory where you unzipped the files.
- Open the RMarkdown file (`manuscript.Rmd`) in RStudio.

3. Install Required Packages:

- If not already installed, install the required R packages from the CRAN.

4. Knit the Document:

- Knit the RMarkdown file to reproduce the analysis. This will execute the code chunks, perform the analysis, and generate the final document.

By following these steps, you can recreate the entire analysis and verify the results presented in this paper. This approach ensures transparency and allows others to validate and build upon the findings of this study.

Reference

Centers for Disease Control and Prevention. 2020a. "NHANES 2017-2018 Overview." wwwn.cdc.gov. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2017.

———. 2020b. "2017-2018 Data Documentation, Codebook, and Frequencies Blood Pressure (BPX_j)." wwwn.cdc.gov. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BPX_J.htm.

———. 2020c. "2017-2018 Data Documentation, Codebook, and Frequencies Demographic Variables and Sample Weights (DEMO_j)." wwwn.cdc.gov. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.htm.

———. 2020d. "2017-2018 Data Documentation, Codebook, and Frequencies Diabetes (DIQ_j)." wwwn.cdc.gov. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DIQ_J.htm.

———. 2020e. "2017-2018 Data Documentation, Codebook, and Frequencies Health Insurance (HIQ_j)." wwwn.cdc.gov. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/HIQ_J.htm.

———. 2020f. "2017-2018 Data Documentation, Codebook, and Frequencies Dietary Interview - Total Nutrient Intakes, First Day (DR1TOT_j)." wwwn.cdc.gov. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DR1TOT_J.htm.

———. 2021. "Facts about Hypertension." Centers for Disease Control; Prevention. https://www.cdc.gov/bloodpressure/facts.htm.

———. 2023. "2019-2020 Examination Data - Continuous NHANES." wwwn.cdc.gov. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2019.

Forouzanfar, Mohammad H, Ashkan Afshin, Lily T Alexander, H Ross Anderson, Zulfiqar A Bhutta, Stan Biryukov, Michael Brauer, et al. 2016. "Global, Regional, and National Comparative Risk Assessment of 79 Behavioural, Environmental and Occupational, and Metabolic Risks or Clusters of Risks, 1990–2015: A Systematic Analysis for the Global Burden of Disease Study 2015." *The Lancet* 388 (October): 1659–1724. https://doi.org/10.1016/s0140-6736(16)31679-8.

Grover, Ashoo, and Ashish Joshi. 2014. "An Overview of Chronic Disease Models: A Systematic Literature Review." *Global Journal of Health Science* 7 (October): 210–27. https://doi.org/10.5539/gjhs.v7n2p210.

Iqbal, Afrin, Karar Zunaid Ahsan, Kanta Jamil, M. Moinuddin Haider, Shusmita Hossain Khan, Nitai Chakraborty, and Peter Kim Streatfield. 2021. "Demographic, Socioeconomic, and Biological Correlates of Hypertension in an Adult Population: Evidence from the Bangladesh Demographic and Health Survey 2017–18." *BMC Public Health* 21 (June). https://doi.org/10.1186/s12889-021-11234-5.

Islam, M N, Md. Jahangir Alam, Md. Maniruzzaman, N. A. M. Faisal Ahmed, Mohammad Ali, Md. Jahanur Rahman, and Dulal Chandra Roy. 2023. "Predicting the Risk of Hypertension Using Machine Learning Algorithms: A Cross Sectional Study in Ethiopia." *PLOS ONE* 18 (August): e0289613–13. https://doi.org/10.1371/journal.pone.0289613.

Johnson, Rachel K., Lawrence J. Appel, Michael Brands, Barbara V. Howard, Michael Lefevre, Robert H. Lustig, Frank Sacks, Lyn M. Steffen, and Judith Wylie-Rosett. 2009. "Dietary Sugars Intake and Cardiovascular Health." *Circulation* 120 (September): 1011–20. https://doi.org/10.1161/circulationaha.109.192627.

Sofogianni, Areti, Nikolaos Stalikas, Christina Antza, and Konstantinos Tziomalos. 2022. "Cardiovascular Risk Prediction Models and Scores in the Era of Personalized Medicine." *Journal of Personalized Medicine* 12 (July): 1180. https://doi.org/10.3390/jpm12071180.

United States Department of Agriculture. 2022. "Macronutrients | National Agricultural Library." Usda.gov. https://www.nal.usda.gov/human-nutrition-and-food-safety/food-composition/macronutrients.

XGBoost Developer. 2022. "Understand Your Dataset with XGBoost — Xgboost 2.0.2 Documentation." xgboost.readthedocs.io. https://xgboost.readthedocs.io/en/stable/R-package/discoverYourData.html#build-the-feature-importance-data-table.