Module 6 Final Project | Nutrition Data

Trang Tran

CPS Analytics, Northeastern University

ALY6020 | Predictive Analytics

Professor Prashant Mittal

Dec 16, 2023

Introduction

With a cleaned version of the original Nutrition data, which is sourced from the Food and Drug Administration (FDA), the analysis focuses on nutrient information of around nine thousand food items. The objectives include four main parts: Data Cleaning, Imputation, Preprocessing, and Predictive Modeling. Selected predictors undergo imputation, preprocessing, and outlier detection, paving the way for various classification techniques like KNN, Logistic Regression, SVM, Random Forest, Gradient Boosting, and XGBoost. The report concludes by prescribing the best-performing model for predicting a binary calorie response variable, offering insights for informed decision-making in food nutrition labeling.

Data Cleaning and Imputation

The cleaned Nutrition dataset consists of 8789 rows and 45 columns. We chose to keep 15 predictors and one target variable, ‘Calories_binary’, for the analysis. Figure 1 below shows the number of missing values (NAs) per column in the dataset.

```
all_nutrition.isna().sum()
```

```
Sugar_Tot_g        1832
Vit_B6_mg           905
Vit_C_mg            818
Magnesium_mg        739
Zinc_mg             706
Niacin_mg           637
Thiamin_mg          634
Riboflavin_mg       616
Fiber_TD_g          594
Phosphorus_mg       579
Potassium_mg        426
Cholestrl_mg        410
Calcium_mg          348
Iron_mg             144
Sodium_mg            83
Calories_binary    3083
dtype: int64
```

Figure 1: Missing Values Checking

Subsequently, we randomly chose three predictors, 'Vit_B6_mg', 'Sugar_Tot_g', 'Vit_C_mg' for imputation purposes. After applying three multidimensional imputation methods, including

MICE (Multiple Imputation by Chained Equations), KNN-based Imputer, and Random Forest Imputer, we have three versions of imputation values for three chosen variables as follows (Figure 2).

MICE

Original data with missing values

	Vit_B6_mg	Sugar_Tot_g	Vit_C_mg
7	0.074	NaN	0.0
9	0.074	NaN	0.0
20	0.271	NaN	0.0
38	0.124	NaN	0.0
40	0.065	NaN	0.0
...	...	...	...
8540	0.084	2.83	NaN
8541	0.085	0.75	NaN
8542	0.077	4.20	NaN
8543	0.382	0.00	NaN
8544	0.613	NaN	NaN

[2938 rows x 3 columns]

Imputed Data

	Vit_B6_mg	Sugar_Tot_g	Vit_C_mg
7	0.074	8.026649	0.000000
9	0.074	8.026649	0.000000
20	0.271	8.365787	0.000000
38	0.124	8.112724	0.000000
40	0.065	8.011155	0.000000
...	...	...	...
8540	0.084	2.830000	0.829390
8541	0.085	0.750000	0.317606
8542	0.077	4.200000	0.957257
8543	0.382	0.000000	9.920145
8544	0.613	9.184172	19.946978

KNN

Original data with missing values

	Vit_B6_mg	Sugar_Tot_g	Vit_C_mg
7	0.074	NaN	0.0
9	0.074	NaN	0.0
20	0.271	NaN	0.0
38	0.124	NaN	0.0
40	0.065	NaN	0.0
...	...	...	...
8540	0.084	2.83	NaN
8541	0.085	0.75	NaN
8542	0.077	4.20	NaN
8543	0.382	0.00	NaN
8544	0.613	NaN	NaN

[2938 rows x 3 columns]

Imputed Data

	Vit_B6_mg	Sugar_Tot_g	Vit_C_mg
7	0.074	3.612	0.00
9	0.074	3.612	0.00
20	0.271	2.396	0.00
38	0.124	1.386	0.00
40	0.065	18.832	0.00
...	...	...	...
8540	0.084	2.830	15.08
8541	0.085	0.750	3.74
8542	0.077	4.200	0.28
8543	0.382	0.000	0.08
8544	0.613	0.280	2.70

Random Forest

Original data with missing values

	Vit_B6_mg	Sugar_Tot_g	Vit_C_mg
7	0.074	NaN	0.0
9	0.074	NaN	0.0
20	0.271	NaN	0.0
38	0.124	NaN	0.0
40	0.065	NaN	0.0
...	...	...	...
8540	0.084	2.83	NaN
8541	0.085	0.75	NaN
8542	0.077	4.20	NaN
8543	0.382	0.00	NaN
8544	0.613	NaN	NaN

[2938 rows x 3 columns]

Imputed Data

	Vit_B6_mg	Sugar_Tot_g	Vit_C_mg
7	0.074	8.600540	0.000
9	0.074	8.600540	0.000
20	0.271	1.254025	0.000
38	0.124	2.857915	0.000
40	0.065	14.129800	0.000
...	...	...	...
8540	0.084	2.830000	4.644
8541	0.085	0.750000	1.376
8542	0.077	4.200000	1.338
8543	0.382	0.000000	0.063
8544	0.613	12.335000	37.472

Figure 2: Imputation Results

There is a big difference between the imputation results of the three techniques. We decided to keep the KNN method's result because we believe that the dataset exhibits localized patterns and relationships that take advantage of KNN's principle of proximity. KNN is also well-suited for datasets with multiple predictors and suitable for imputing missing values while avoiding overfitting to noise. In this case, the result of KNN imputation is not unusually large fluctuation amplitude like the other two methods.

Data Preprocessing

With the imputed dataset in place, we proceed to remove the remaining missing values from the data. The final dataset comprises 4413 observations and 16 columns. After defining predictors and response variables, we conducted the standardization process for all numerical predictors. The next step involves applying three outlier detection methods, including Local Outlier Factor,

Isolation Forest, and One-class SVM. Figures 3 and 4 present a summary of the outlier detection process, indicating that the total number of outliers detected by all three methods is less than 1% (~0.63%) of the data.

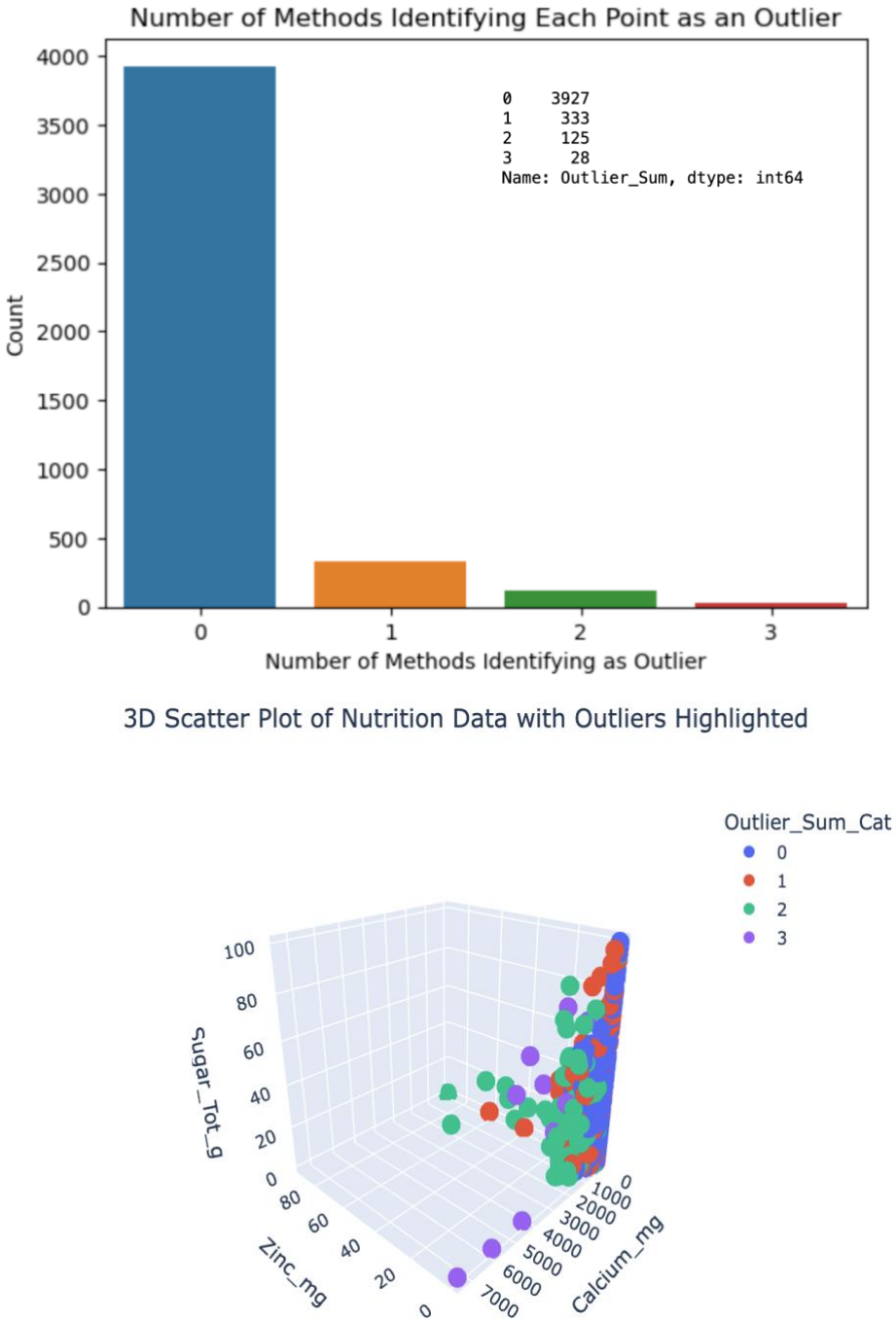

Figure 4: An example of 3D scatter plot with Outliers Highlighted

Predictive Modeling (Classification Models)

This section illustrates various classification techniques applied to the dataset for predicting the "Calories_binary" (High – Low) target variable. We divided the data into training and test sets with an 80-20 ratio and performed Randomized Search cross-validation for hyperparameter tuning to identify optimal configurations for each model. Here are the results for six models:

	precision	recall	f1-score	support
High	0.98	0.96	0.97	522
Low	0.95	0.97	0.96	361
accuracy			0.97	883
macro avg	0.96	0.97	0.96	883
weighted avg	0.97	0.97	0.97	883

KNN:Overall Accuracy on Test Set with best k: 0.9660249150622876

	precision	recall	f1-score	support
High	0.93	0.89	0.91	522
Low	0.85	0.91	0.88	361
accuracy			0.90	883
macro avg	0.89	0.90	0.89	883
weighted avg	0.90	0.90	0.90	883

Logistic Regression: Overall Accuracy on Test Set: 0.8969422423556059

	precision	recall	f1-score	support
High	0.94	0.95	0.95	522
Low	0.93	0.91	0.92	361
accuracy			0.94	883
macro avg	0.93	0.93	0.93	883
weighted avg	0.94	0.94	0.94	883

SVM: Overall Accuracy on Test Set: 0.9354473386183465

	precision	recall	f1-score	support
High	0.97	0.97	0.97	522
Low	0.96	0.96	0.96	361
accuracy			0.97	883
macro avg	0.97	0.97	0.97	883
weighted avg	0.97	0.97	0.97	883

Random Forest: Overall Accuracy on Test Set: 0.9671574178935447

	precision	recall	f1-score	support
High	0.97	0.97	0.97	522
Low	0.96	0.96	0.96	361
accuracy			0.97	883
macro avg	0.97	0.97	0.97	883
weighted avg	0.97	0.97	0.97	883

Gradient Boosting: Overall Accuracy on Test Set: 0.9682899207248018

	precision	recall	f1-score	support
0	0.97	0.97	0.97	522
1	0.96	0.96	0.96	361
accuracy			0.97	883
macro avg	0.97	0.97	0.97	883
weighted avg	0.97	0.97	0.97	883

XGBoost: Overall Accuracy on Test Set: 0.9671574178935447

Figure 5: Classification Model Results

Among the six classification models, KNN, Random Forest Classifier, Gradient Boosting Classifier, and XGBoost stand out with an impressive overall accuracy score of approximately 97% and well-balanced highest scores in other metrics such as precision, recall, and F1. In contrast, the Logistic Regression model exhibits a lower overall accuracy score of around 90%. It also presents other metrics, including precision, recall, and F1 score, lower than those of other models. The SVM model achieves a median accuracy score of 94%. While it may not surpass the

top models, it still demonstrates competitive performance. If the highest accuracy is the primary consideration, models like KNN, Random Forest Classifier, Gradient Boosting Classifier, or XGBoost may be preferred.

Lastly, we delved deeper into the variable importance information by evaluating the feature importance of the Random Forest model. Figure 6 indicates that the nutrient 'Phosphorus' is the most important feature in predicting the binary calorie response (with a weight of >0.12), followed by 'Niacin', 'Sugar', 'Iron', 'Zinc', etc.

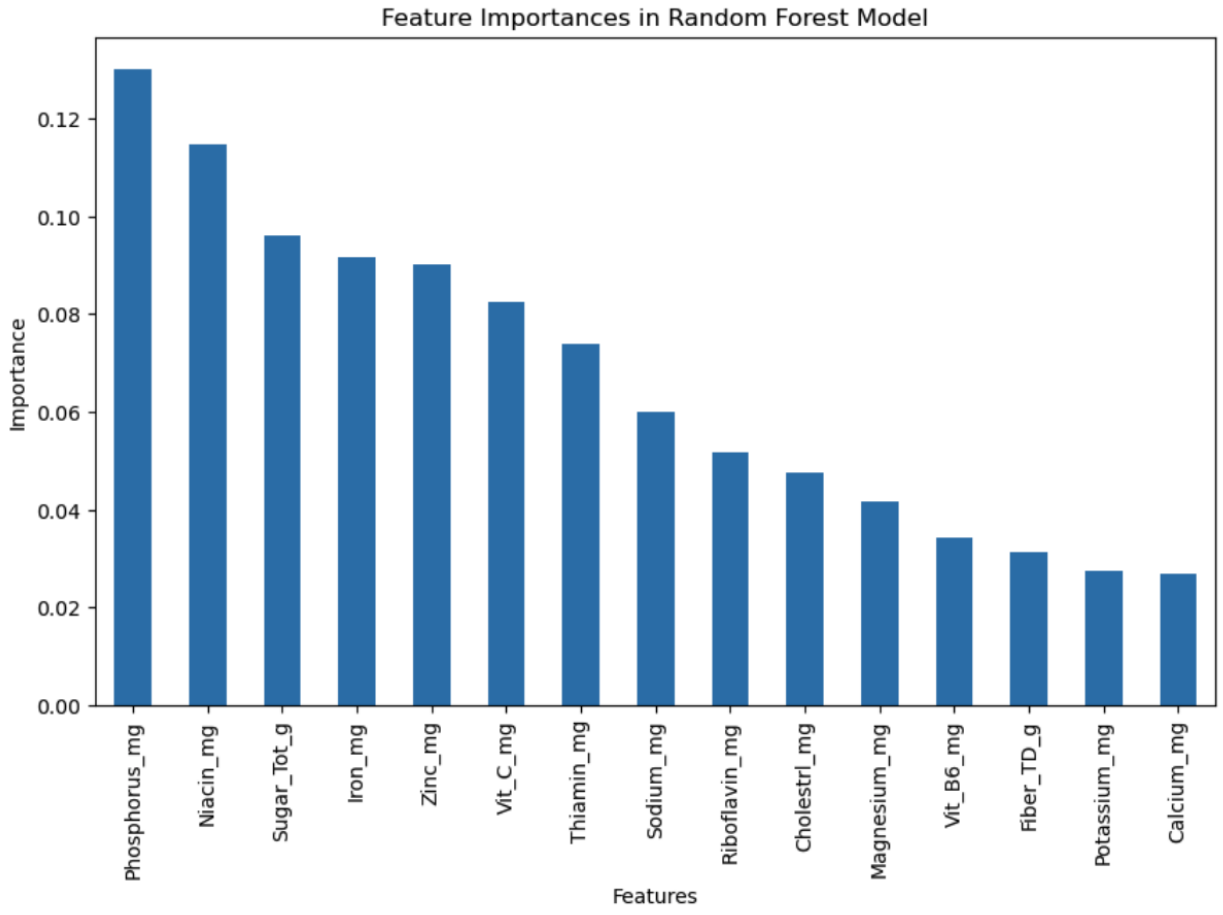

Figure 6: Variable Importance Information