Machine Learning Classification and Regression

5 algorithms for Classification and 5 algorithms for regression

Abstract

This comprehensive study explored the application of machine learning techniques across two distinct domains: powerlifting performance classification and nutritional content prediction in foods.

In the first phase, five classification algorithms (Adaptive Boosting, Gradient Boosting, Naive Bayes, Random Forest, and Logistic Regression) were evaluated for their effectiveness in categorizing powerlifters. The analysis incorporated performance metrics including lift ratios (squat/bench/deadlift), absolute strength values, and categorical variables. A notable methodological adaptation involved the implementation of quantile transformation for numerical features in the Naive Bayes algorithm, enhancing its performance with non-normally distributed data.

The models showed varying degrees of success in discriminating between gender categories, weight classes, and competition divisions. The second phase employed five regression algorithms (Elastic Net, Huber Regressor, Linear Regression, Lasso, and Ridge) to predict B12 and protein concentrations in food items. The regression models demonstrated consistent patterns in their predictive accuracy, with an MSE of 6.06 for protein prediction. Water and carbohydrate content emerged as the strongest negative predictors, while various lipid components showed moderate influence. The models exhibited robust performance for moderate concentrations but demonstrated increased variance at extreme values.

Both analyses highlighted the critical importance of appropriate feature engineering and domain-specific data preprocessing. The study revealed the potential of machine learning in sports science and nutritional analysis while emphasizing the necessity of careful algorithm selection based on data characteristics. These findings contribute to the growing body of knowledge regarding the practical application of machine learning techniques in sports performance analysis and nutritional science.

Introduction

In the realm of sports science and nutrition, the quest for understanding and prediction has led researchers down fascinating paths of inquiry. This study embarks on a journey through the landscape of machine learning, exploring its potential to unravel the scheme behind two very different data, performance in powerlifting and food composition.

Picture, if you will, a bustling powerlifting competition, where athletes strain against unyielding iron. Here, we deploy five sophisticated algorithms - Adaptive Boosting, Gradient Boosting, Naive Bayes, Random Forest, and Logistic Regression - each like a discerning judge, evaluating the lifters based on their feats of strength and personal attributes. These digital arbiters sift through a tapestry of data, from the raw power displayed in squats, bench presses, and deadlifts to the subtle interplay of weight classes and divisions.

Let us venture into the realm of the kitchen, a laboratory of sustenance where the hidden alchemy of nutrition resides in every fragment of food. Here, a specialized ensemble of five algorithms-Elastic Net, Huber Regressor, Linear Regression, Lasso, and Ridge-assume the mantle of analytical gastronomes, tasked with decoding the intricate profiles of B12 and protein content in our daily sustenance. Like enlightened chefs experimenting with complex recipes, these computational tools meticulously calibrate their models, adjusting for the subtle interplay of nutritional variables to distill a refined synthesis of predictive accuracy and scientific insight.

As we embark further on this analytical journey, we will reveal the unique capabilities and peculiarities of each approach, crafting a compelling narrative of how cutting-edge analytics can shed light on the domains of athletic achievement and nutritional understanding. Accompany us as we navigate the confluence of human endeavor and computational expertise, where raw data intertwines with physical strength, and micronutrients emerge as critical variables in a sophisticated framework of quantitative analysis.

Data

PowerLifting

The powerlifting database is a comprehensive collection of data from powerlifting competitions, featuring information on thousands of athletes across various weight classes, age groups, and divisions. This dataset, maintained by organizations like OpenPowerlifting, provides valuable insights for researchers, coaches, and enthusiasts interested in analyzing trends and performance in the sport of powerlifting.

The dataset is substantial in size, containing over 2.3 million rows and 41 columns. This vast amount of data covers a wide range of information about powerlifting competitions and athletes. The columns include both numerical and categorical data types, providing a rich set of features for analysis.

Some of the key columns in the dataset include:

Name: A string column containing the full name of each athlete.

Sex: A categorical column with two levels - 'M' for males and 'F' for females.

Age: A numerical column representing the athlete's age.

BodyweightKg: A numerical column showing the athlete's bodyweight in kilograms.

WeightClassKg: A numerical column indicating the athlete's weight class.

Best3SquatKg, Best3BenchKg, Best3DeadliftKg: Numerical columns representing the best lifts for each discipline.

TotalKg: A numerical column showing the sum of the best lifts across all three disciplines.

Equipment: A categorical column with multiple levels, such as 'Raw', 'Single-ply', etc.

Federation: A categorical column with numerous levels, representing different powerlifting federations.

Date: A date column indicating when the competition took place.

The categorical columns in the dataset can have varying numbers of levels. For example, the 'Equipment' column might have 3-5 levels, while the 'Federation' column could have dozens of levels representing different organizing bodies worldwide.

Powerlifting, as a strength sport, consists of three main lifts: the squat, the bench press, and the deadlift. Each of these movements tests different aspects of an athlete's strength and technique.

The squat is a full-body exercise that primarily targets the legs and core. In this lift, the athlete places a barbell across their upper back, squats down until their thighs are parallel to the ground or lower, and then stands back up. This movement requires significant leg strength, core stability, and overall body control.

The bench press focuses on upper body strength, particularly the chest, shoulders, and triceps. For this lift, the athlete lies on a bench and lowers a barbell to their chest before pressing it back up to full arm extension. Proper technique in the bench press involves a careful balance of power and control.

Finally, the deadlift is often considered the ultimate test of overall body strength. In this lift, the athlete bends down and lifts a barbell from the ground to a standing position with the bar at hip level.

The deadlift engages nearly every major muscle group in the body, making it a true measure of an athlete's raw strength.

This dataset offers significant opportunities for machine learning applications aimed at uncovering patterns in lifting performance or predicting specific outcomes based on input variables. One particularly interesting approach involves creating new variables that represent proportions—namely how much each lift (squat, bench press, or deadlift) contributes to an athlete's total weight lifted (TotalKg). These proportions can provide deeper insights into lifting strategies and styles while serving as valuable features for machine learning models.

To create these variables:

$$BestMovmentProprotion = \frac{BestMovmentKg}{TotalKg}$$

It indicates how significant deadlifts are to their overall total.

Using these newly created proportions as features allows us to explore how men and women differ in their lifting styles. For example, men may rely more heavily on deadlifts compared to women who might distribute their strength differently across lifts due to physiological differences such as muscle mass distribution or biomechanics.

A machine learning model can be trained using these proportions (SquatProportion, BenchProportion, DeadliftProportion) as input features to classify gender (Sex). By training models such as logistic regression or random forests on this data, we can predict whether an athlete is male or female based on their relative strengths in these three lifts. This approach not only helps identify patterns but also demonstrates how feature engineering can enhance machine learning applications in sports analytics.

Beyond gender classification, machine learning can also be applied to predict other outcomes such as weight class (WeightClassKg) or competitive division (Division) using raw lift values like Best3SquatKg, Best3BenchKg, and Best3DeadliftKg. These tasks involve building classification models that use lifting performance metrics to categorize athletes into appropriate groups based on their competition data.

In summary, this dataset provides an excellent opportunity to apply machine learning techniques by leveraging both raw lift values and engineered features like proportions to uncover patterns in athletic performance. By focusing on tasks such as gender classification or weight class prediction, we can gain valuable insights into how different factors influence success in powerlifting while showcasing how data-driven approaches can enhance our understanding of this strength sport.

Food

The dataset from Kaggle titled "Food Vitamins, Minerals, Macronutrient" provides a comprehensive collection of nutritional information for various food items. This dataset is particularly well-suited for applying regression machine learning models to predict vitamin B12 and protein content in foods.

The dataset contains numerous numerical variables, including macronutrients (protein, fat, carbohydrates), vitamins, and minerals. These variables offer a rich set of features that can be utilized to build predictive models for B12 and protein content.

For predicting vitamin B12 content, relevant numerical features might include other B-complex vitamins (B1, B2, B3, B6), as well as minerals like iron and zinc, which are often found in B12-rich foods. Additionally, macronutrient percentages could provide valuable information, as B12 is primarily found in animal-based products, which tend to be higher in protein and fat.

When focusing on protein content prediction, key numerical variables to consider would be the other macronutrients (fat and carbohydrates), as well as various minerals like iron, zinc, and phosphorus, which are often associated with protein-rich foods. Energy content (calories) could also be a useful predictor, as higher-protein foods often have higher caloric density.

The dataset's diverse range of food items allows for the development of robust regression models that can generalize well across different food categories. This variety is crucial for creating models that can accurately predict B12 and protein content in a wide array of foods, from animal products to plant-based alternatives.

To build effective regression models, techniques such as multiple linear regression, random forests, or gradient boosting machines could be employed. These models can leverage the relationships between various nutritional components to make accurate predictions of B12 and protein content.

By utilizing this dataset for machine learning applications, researchers and nutritionists can gain valuable insights into the relationships between different nutrients and develop tools for estimating important nutritional values in foods, potentially aiding in dietary planning and nutritional research.

Experiments

Classification

Adaptive Boosting

BinaryClass Men Women

Based on the provided visualizations, the Adaptive Boosting model shows interesting performance in predicting gender based on lifting proportions:

The SHAP feature importance plot reveals that both BenchProportion and SquatProportion contribute significantly to the model's predictions, with both features showing both high and low impact interactions. This suggests that the way athletes distribute their strength between lifts varies meaningfully between genders.

The confusion matrix demonstrates that the model performs better at identifying male lifters (93% accuracy) than female lifters (48% accuracy). Specifically:

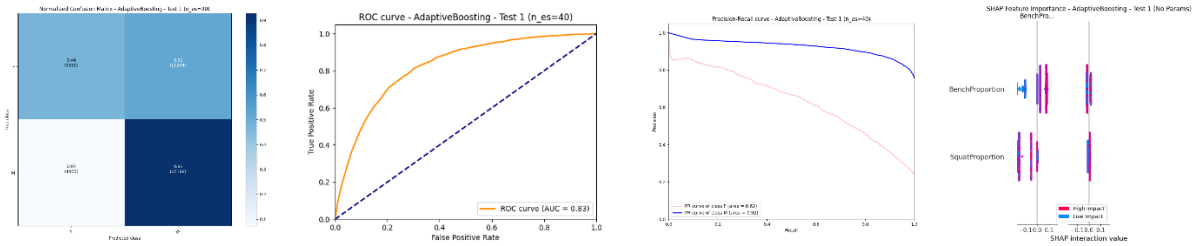
57,195 males were correctly identified (93% of male cases)

9,346 females were correctly identified (48% of female cases)

There is a notable imbalance in the dataset, with many more male than female lifters

The ROC curve shows good overall discriminative ability with an AUC of 0.83, well above the random chance baseline of 0.5. This indicates that the model has learned meaningful patterns in how lifting proportions differ between men and women, despite the class imbalance.

These results suggest that while gender can be predicted from lifting proportions with reasonable accuracy, the model's lower performance on female lifters indicates that either women have more variable lifting patterns or the class imbalance in the training data affected the model's ability to learn female lifting patterns effectively.



Multiclass

Division

We can see with the graph after that we can analyze to summurize the AdaptiveBoosting model's performance for multiclass classification of powerlifting divisions:

ROC Curve Analysis:

The ROC curves show varying performance across different divisions. The "Boys" division has the best performance with an AUC of 0.79, followed by "Open Men" with an AUC of 0.73. Other divisions like "Amateur Open", "Open", and "R-O" have moderate performance with AUCs around 0.60-0.63. The model struggles more with divisions like "Junior", "Junior 19-23", and "O", which have AUCs closer to 0.55-0.60.

Precision-Recall Curve:

The PR curves reveal significant class imbalance issues. The "Boys" class performs best with an area of 0.60, while "Open" has an area of 0.42. Most other classes have very low areas (0.03-0.06), indicating poor precision-recall trade-offs for those divisions.

Confusion Matrix:

The normalized confusion matrix shows that the model has varying success in correctly classifying different divisions:

"Boys" has the highest correct classification rate at 87%

"Open Men" is correctly classified 67% of the time

Most other divisions are often misclassified as either "Boys" or "Open"

There's significant confusion between similar categories (e.g., Junior divisions)

Feature Importance:

The SHAP value plots show different feature importances for various divisions:

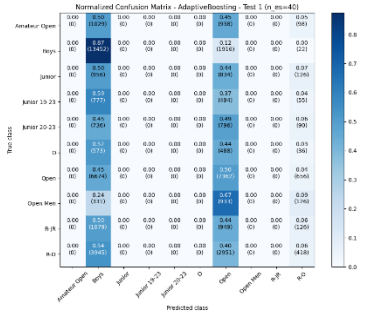
For "Boys", "BestDeadliftKg" has the highest impact, followed by "BestBenchKg"

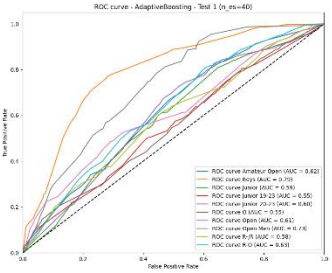
For "Open", "BestDeadliftKg" and "BestSquatKg" are most important

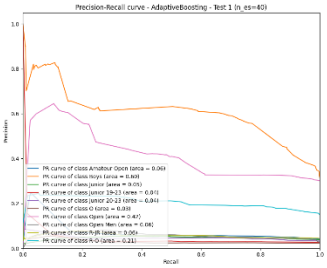
"Amateur Open" is most influenced by "BestBenchKg" and "BestDeadliftKg"

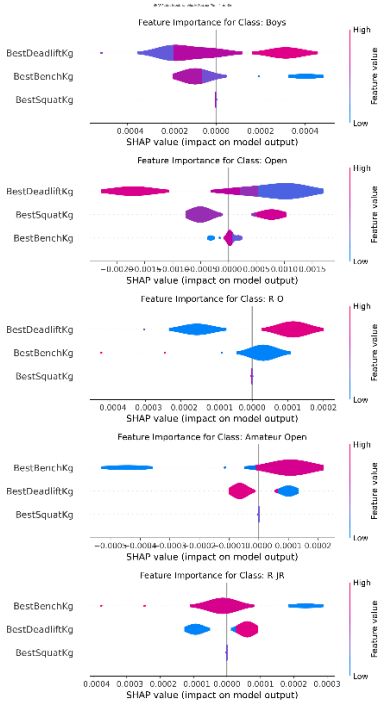
"R-JR" classification depends mainly on "BestBenchKg" and "BestDeadliftKg"

In conclusion, the AdaptiveBoosting model shows varying performance across different powerlifting divisions, with better results for some divisions (like Boys and Open Men) and poorer performance for others. The model struggles with class imbalance and there's significant confusion between similar categories. Feature importance varies by division, but generally, deadlift and bench press performances are the most influential factors in classification.









WeightClass

The analysis of AdaptiveBoosting performance for weightclass classification reveals several interesting aspects:Overall Performance:

- The 60kg class shows the best performance with an AUC of 0.86 and a precision-recall (PR) of 0.20
- The 67.5kg class follows with an AUC of 0.79 and a PR of 0.23
- Heavier classes (90kg, 82.5kg) show weaker performance with respective AUCs of 0.57 and 0.58

Confusion Matrix:

- The 60kg class shows the best precision with 84% correct classifications
- The 67.5kg class achieves 69% correct classifications

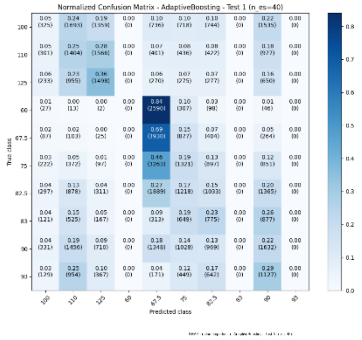
- Significant confusion is observed between adjacent classes, particularly in middle weight categories
- Heavier classes (90kg, 93kg) show more confusion with other categories

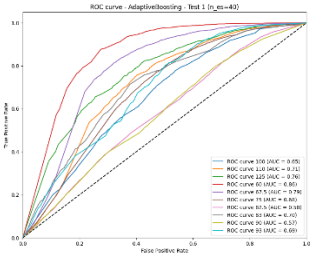
Feature Analysis (SHAP values):

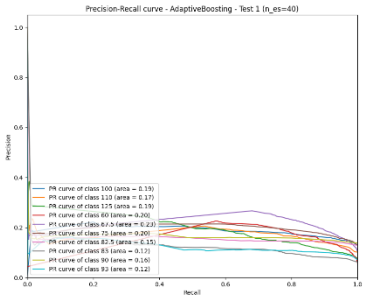
- For 90kg class: BenchKg is the most important feature, followed by SquatKg
- For 75kg class: BenchKg and SquatKg have similar impacts
- For 100kg class: SquatKg and BenchKg are the most determinant
- For 82.5kg class: BenchKg and DeadliftKg are the main predictors
- For 67.5kg class: BenchKg shows the greatest influence

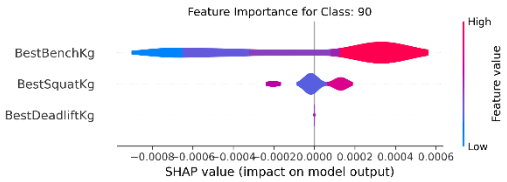
This analysis reveals that:

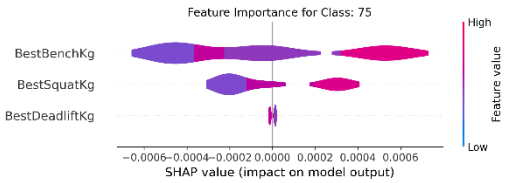
- Extreme weight categories (very light or very heavy) are more easily identifiable
- Performance varies significantly across weight categories
- The relative importance of different lifts varies by weight category
- The model struggles to discriminate between adjacent weight categories, particularly in middle weights

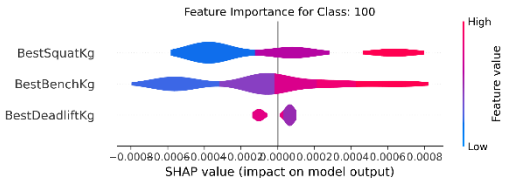


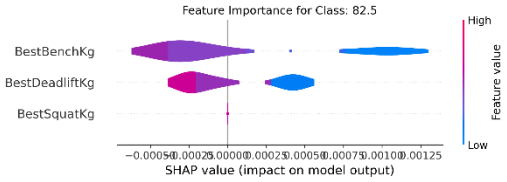


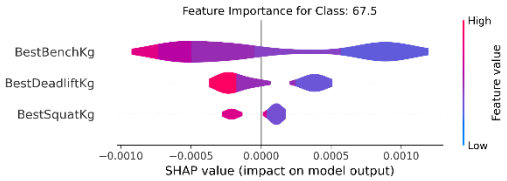












GradientBoosting

BinaryClass

WomenMen

Before analyzing the performance of the GradientBoosting model for discriminating between men and women based on Squat, Bench, and Deadlift proportions, it's important to understand how this algorithm works.

GradientBoosting is a machine learning technique that combines multiple weak models (typically decision trees) to create a powerful predictive model. The algorithm works iteratively, with each new model being trained to correct the errors of previous models.

Here are the main steps of the algorithm:

An initial simple model is created to make baseline predictions

The errors (residuals) of this model are calculated

A new model is trained to predict these errors

The predictions of this new model are added to previous predictions, with a learning rate to control the new model's influence

This process is repeated a predetermined number of times, each iteration aiming to further reduce residual errors

GradientBoosting is particularly effective at capturing complex relationships in data and can handle both classification and regression problems.

Now, analyzing the model's performance based on the provided graphs:

Overall Performance:

The ROC curve shows strong model performance with an AUC of 0.83, indicating good discriminative ability between men and women.

Confusion Matrix:

The model performs better at identifying males (M) with 93% accuracy (57,526 correct predictions)

For females (F), accuracy is lower at 48% (9,423 correct predictions)

There is a significant class imbalance in the dataset, with many more male than female entries

Precision-Recall Curves:

The curve for class M (male) has an area of 0.93, showing excellent performance

The curve for class F (female) has an area of 0.65, indicating moderate performance

Feature Importance:

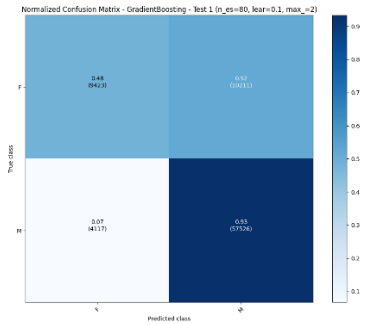
According to the SHAP values graph:

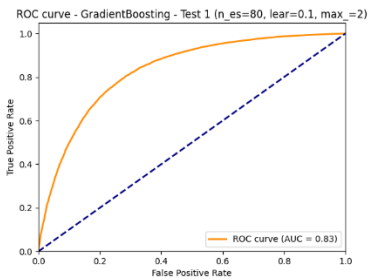
BenchProportion has the highest impact on model predictions

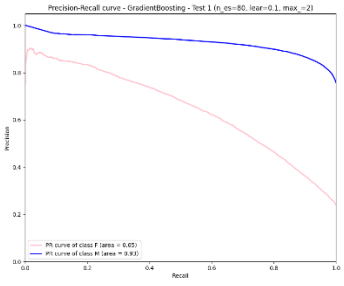
DeadliftProportion shows moderate importance

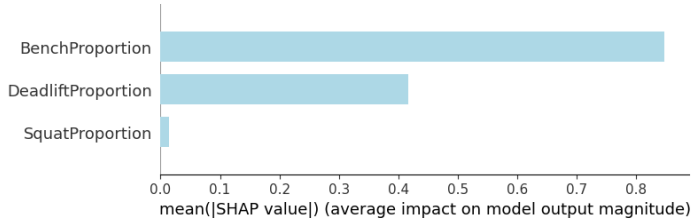
SquatProportion has the least influence on predictions

In conclusion, the GradientBoosting model shows good ability to discriminate between men and women, though with notably better performance for male identification. The class imbalance in the data likely contributes to this difference. Bench press proportion appears to be the most determinant factor for this classification.









Multiclass

Division

Based on the provided images, I can analyze the GradientBoosting model's performance for classifying powerlifting divisions using BestSquat, BestBench, and BestDeadlift as features:

ROC Curve Analysis:

The ROC curves show varying performance across different divisions. The "Boys" division has the best performance with an AUC of 0.92, followed by "Open Men" and "R-O" with AUCs of 0.73 and 0.74 respectively. Other divisions like "Open" and "Junior 19-23" have moderate performance with AUCs around 0.71 and 0.66. The model struggles more with divisions like "Junior" and "O", which have AUCs around 0.62.

Precision-Recall Curve:

The PR curves reveal significant class imbalance issues. The "Boys" class performs exceptionally well with an area of 0.85, while "Open" has an area of 0.47. Most other classes have very low areas (0.05-0.15), indicating poor precision-recall trade-offs for those divisions.

Confusion Matrix:

The normalized confusion matrix shows that:

"Boys" has the highest correct classification rate at 89%

"Open Men" is correctly classified 82% of the time

Most other divisions are often misclassified as "Open"

There's significant confusion between similar categories (e.g., Junior divisions)

Feature Importance:

The SHAP value plots show different feature importances for various divisions:

For "Open", BestDeadliftKg has the highest impact, followed closely by BestBenchKg

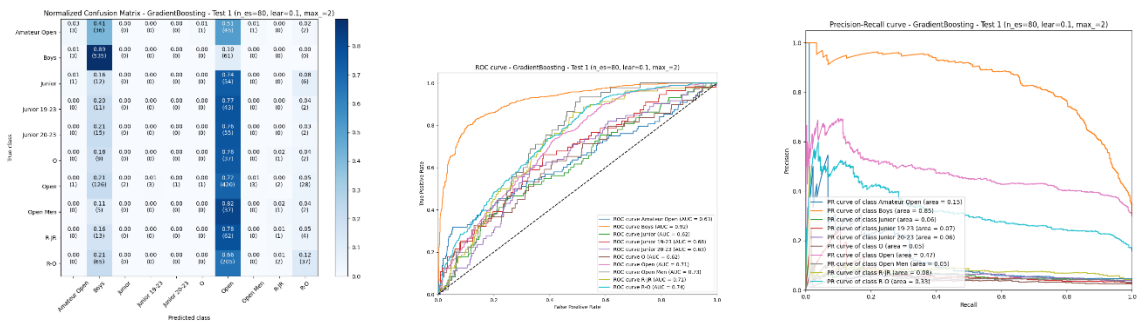
For "R-O", BestDeadliftKg is most important, followed by BestSquatKg

For "Boys", BestDeadliftKg and BestBenchKg are equally important

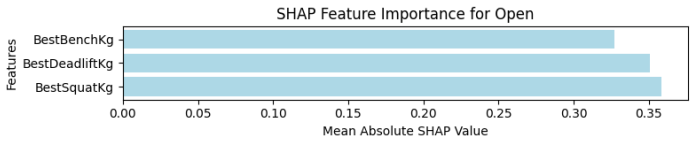
For "O", BestBenchKg has the highest impact

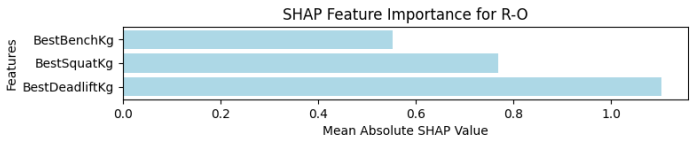
For "Open Men", all three lifts have similar importance, with BestDeadliftKg slightly higher

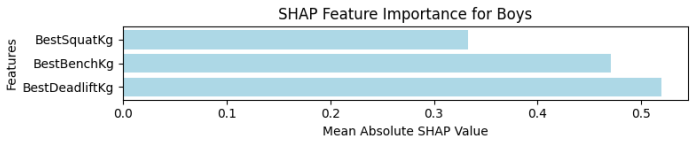
In conclusion, the GradientBoosting model shows varying performance across different powerlifting divisions, with better results for some divisions (like Boys and Open Men) and poorer performance for others. The model struggles with class imbalance and there's significant confusion between similar categories. Feature importance varies by division, but generally, deadlift performance is the most influential factor in classification across multiple divisions.

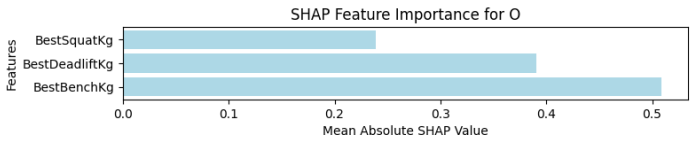


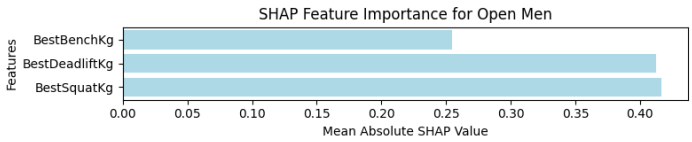
SHAP Feature Importance - GradientBoosting - Test 1 (n_es=80, lear=0.1, max_=2)











WeightClass

Bad Parametrization

Looking at these initial images from the "bad" parameter example of GradientBoosting (n_estimators=80, learning_rate=0.08, max_depth=2), we can observe several indicators of suboptimal performance:

1. Confusion Matrices (Test 1 and Test 2):

- Show significant confusion between weight classes
- Particularly poor discrimination between adjacent weight classes
- Only the 60kg and 67.5kg classes show reasonable performance (around 50-58% accuracy)
- Most other classes show very low correct classification rates (10-30%)

2. ROC Curves (Test 1 and Test 2):

- Wide variation in performance across weight classes

- Best performing classes:
    - 60kg (AUC = 0.85-0.86)
    - 67.5kg (AUC = 0.81-0.84)
- Poorest performing classes:
    - 90kg (AUC = 0.56-0.58)
    - 82.5kg (AUC = 0.59-0.60)
- Most classes show mediocre performance with AUCs between 0.65-0.75

These results suggest that the chosen parameters are not optimal for this classification task. The low number of estimators (80) combined with a relatively high learning rate (0.08) and shallow trees (max_depth=2) likely prevent the model from capturing the complexity of the relationships in the data.

Looking at these additional graphs from the "bad" parameter example (n_estimators=80, learning_rate=0.08, max_depth=2), we can analyze the Precision-Recall curves and SHAP feature importance:

1. Precision-Recall Curves (Test 1 and Test 2):

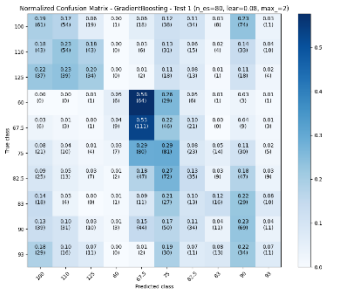
- Most weight classes show poor precision-recall trade-offs
- Best performing classes:
    - 67.5kg (area = 0.28-0.33)
    - 60kg (area = 0.19-0.24)
- Poorest performing classes:
    - 83kg (area = 0.14)
    - 82.5kg (area = 0.15-0.16)
- Most classes show areas between 0.15-0.24, indicating weak overall performance

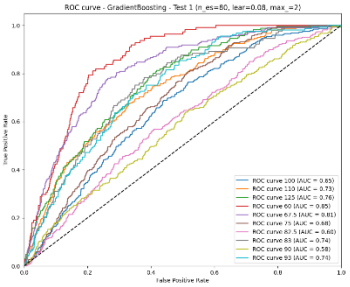
2. SHAP Feature Importance:
   Both tests show similar patterns across weight classes:

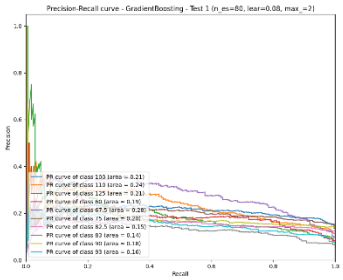
- For 90kg class: BenchKg dominates, followed by SquatKg
- For 75kg class: BenchKg and SquatKg have similar importance
- For 100kg class: SquatKg is most important, followed by BenchKg
- For 67.5kg class: BenchKg and DeadliftKg are primary predictors

The consistency between tests suggests that while the model's performance is poor, it's at least stable in how it uses features for classification. The shallow trees (max_depth=2) combined with the relatively small number of estimators likely prevent the model from capturing complex relationships in the data.

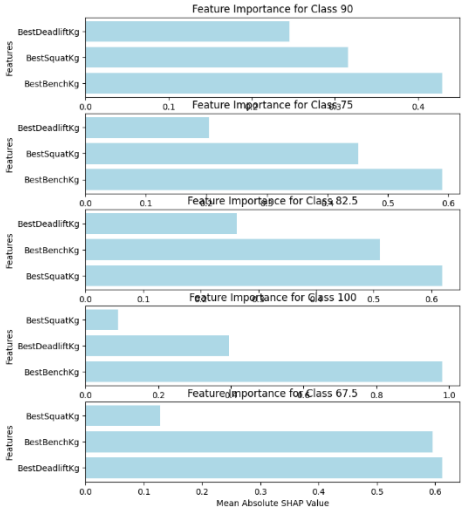
Test1



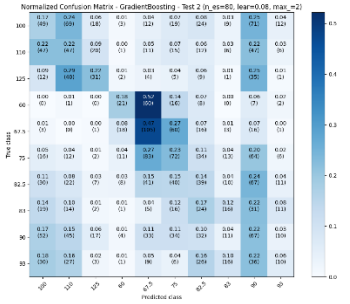


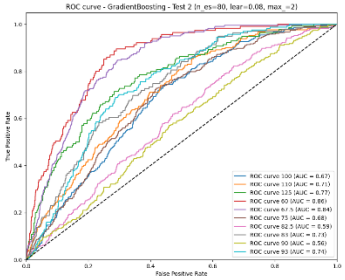


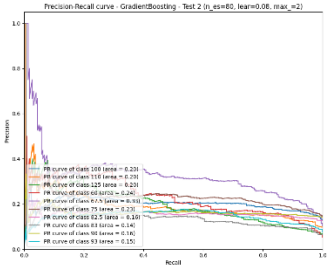
SHAP Feature Importance - GradientBoosting - Test 1 (n_es=80, lear=0.08, max_=2)



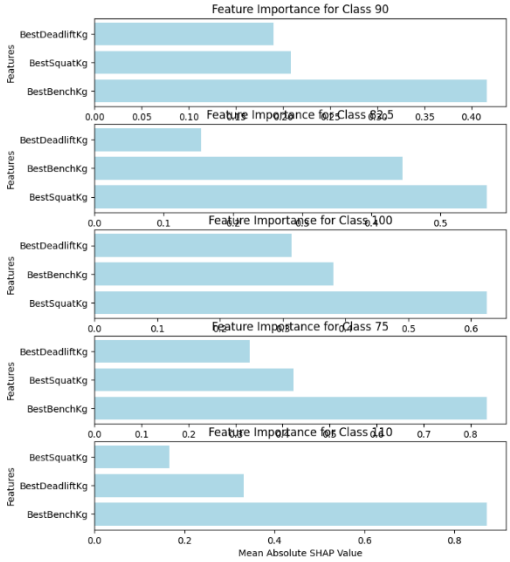
Test2







SHAP Feature Importance - GradientBoosting - Test 2 (n_es=80, lear=0.08, max_=2)



Good Example

Analyzing the "good" example of GradientBoostingClassifier for weight category classification (n_estimators=80, learning_rate=0.12, max_depth=4):

Confusion Matrices (Test 1 and Test 2):

Show improved performance compared to the previous "bad" example

60kg and 67.5kg classes still perform best, with 60% and 48% accuracy in Test 1, 40% and 36% in Test 2

Most other classes show improved classification rates (15-30% range)

Confusion between adjacent weight classes is still present but reduced

ROC Curves (Test 1 and Test 2):

Overall improved performance across weight classes

Best performing classes:

60kg (AUC = 0.87 in Test 1, 0.84 in Test 2)

67.5kg (AUC = 0.81 in Test 1, 0.80 in Test 2)

Poorest performing classes:

82.5kg (AUC = 0.58 in both tests)

90kg (AUC = 0.60 in Test 1, 0.58 in Test 2)

Most classes show moderate to good performance with AUCs between 0.65-0.77

The increased max_depth (4 instead of 2) and slightly higher learning rate (0.12 instead of 0.08) have led to improved model performance, particularly for distinguishing between weight classes. However, there's still room for improvement, especially for middle weight categories.

Analyzing the final two graphs for the "good" example of GradientBoostingClassifier (n_estimators=80, learning_rate=0.12, max_depth=4):

Precision-Recall Curves (Test 1 and Test 2):

Overall improved performance compared to the "bad" example

Best performing classes:

67.5kg (area = 0.32 in Test 1, 0.26 in Test 2)

125kg (area = 0.25 in Test 1, 0.19 in Test 2)

Poorest performing classes:

83kg (area = 0.12 in Test 1, 0.15 in Test 2)

82.5kg (area = 0.15 in Test 1, 0.16 in Test 2)

Most classes show areas between 0.15-0.25, indicating moderate performance

SHAP Feature Importance:

Both tests show similar patterns across weight classes, with some variations:

For 90kg class: BenchKg is most important, followed by SquatKg

For 75kg class: SquatKg is most important in Test 1, while BenchKg is most important in Test 2

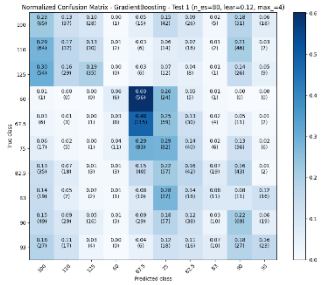
For 82.5kg class: SquatKg is most important, followed by BenchKg

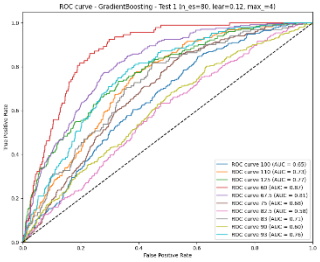
For 100kg class: BenchKg is most important, followed by DeadliftKg

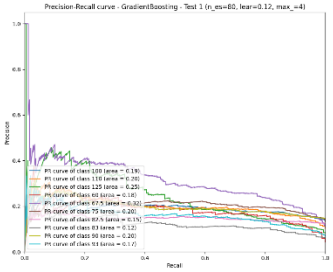
For 67.5kg class: BenchKg is consistently the most important feature

The increased max_depth and learning rate have allowed the model to capture more complex relationships between features and weight classes, resulting in improved performance compared to the "bad" example. However, there's still room for improvement, particularly for middle weight categories.

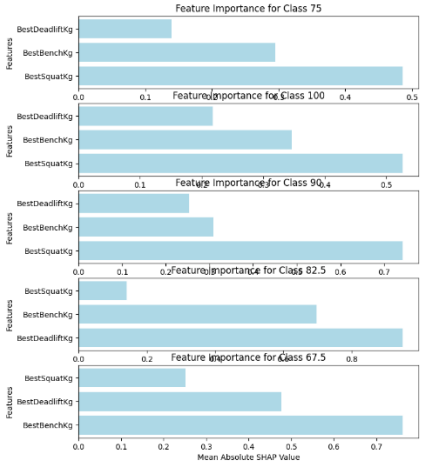
Test1



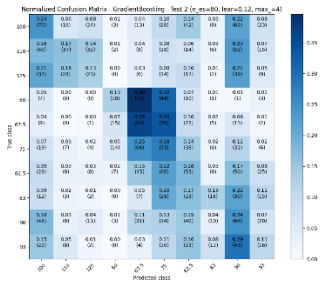


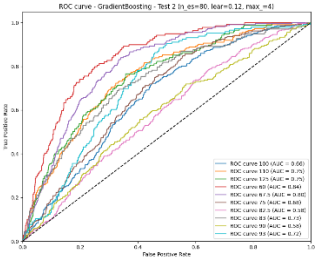


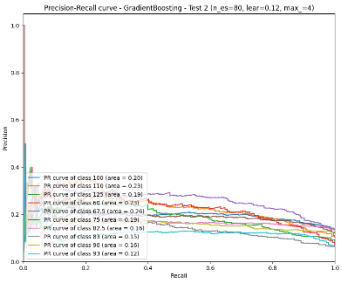
SHAP Feature Importance - GradientBoosting - Test 1 (n_es=80, lear=0.12, max_=4)



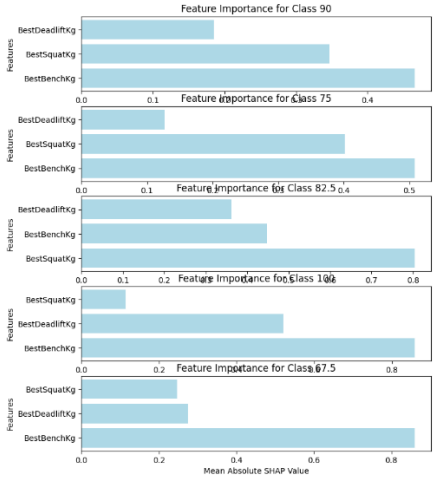
Test2







SHAP Feature Importance - GradientBoosting - Test 2 (n_es=80, lear=0.12, max_=4)

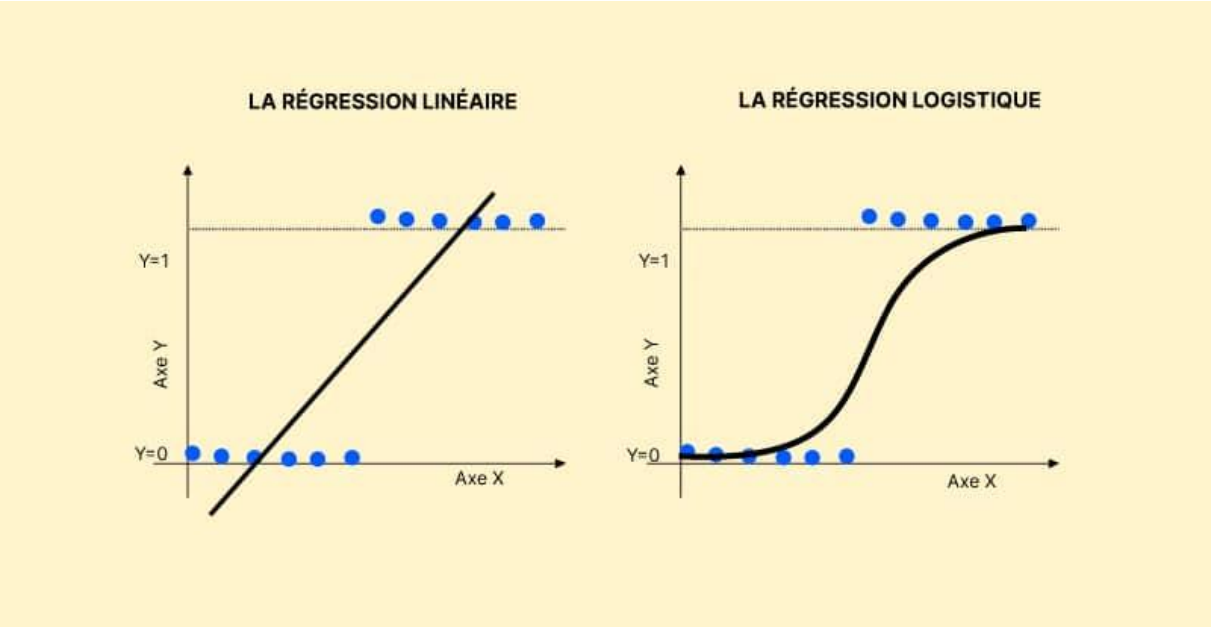


LogisticRegression

BinaryClassification

WomenMen

Logistic regression is a fundamental statistical model used for binary classification problems. It extends the principles of linear regression by applying a logistic function (sigmoid) to transform continuous outputs into probability values between 0 and 1. This transformation makes it ideal for binary classification tasks, such as distinguishing between male and female powerlifters.



The first image illustrates the fundamental difference between linear regression and logistic regression. While linear regression attempts to fit a straight line through binary data points (0 and 1), logistic regression uses a sigmoid function (S-shaped curve) that naturally bounds predictions between 0 and 1. This sigmoid curve provides a more appropriate model for binary classification, as it smoothly transitions between classes while maintaining predictions within the valid probability range.

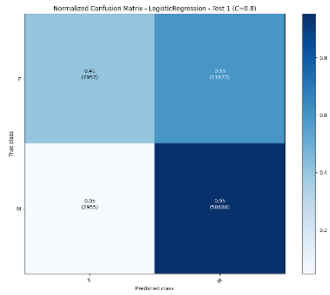
The remaining images show the performance metrics of the logistic regression model (with C=0.8) applied to our powerlifting gender classification task:

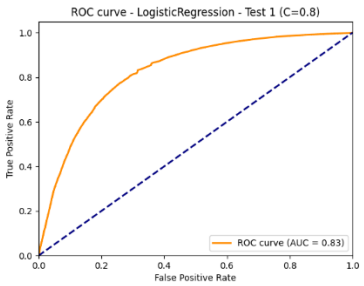
The Precision-Recall curve demonstrates strong performance for male classification (blue line, area=0.92) with consistently high precision across different recall values. The female classification (pink line, area=0.63) shows moderate performance with declining precision as recall increases.

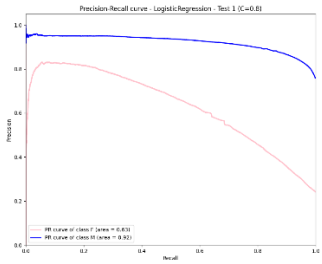
The confusion matrix reveals the model's classification accuracy: 95% for males (58,688 correct predictions) and 41% for females (7,957 correct predictions), highlighting a significant class imbalance in the dataset.

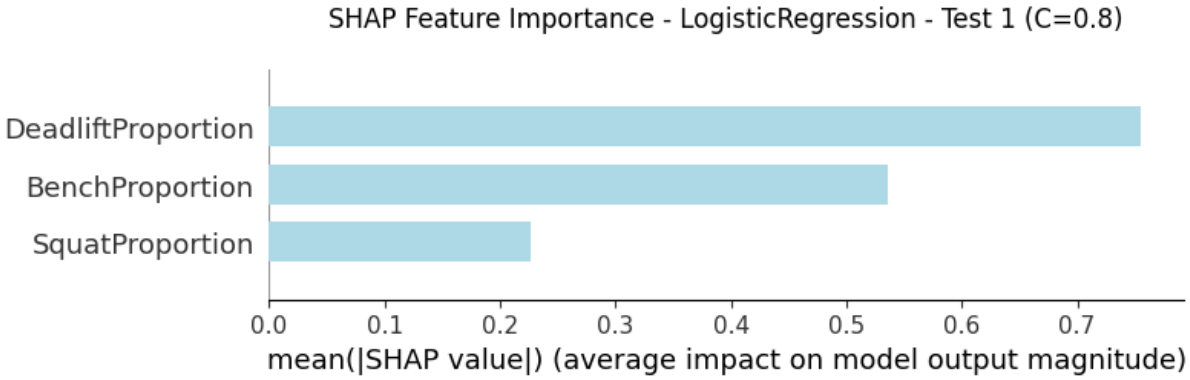
The ROC curve shows strong overall performance with an AUC of 0.83, significantly outperforming random classification (dotted line). The curve's shape indicates good discrimination ability, particularly at lower false positive rates.

These visualizations demonstrate that while the logistic regression model performs well overall, it shows some bias toward the majority class (males) in the dataset.









Multiclass Selection

Division

Based on the provided images, I can analyze the performance of LogisticRegression for multiclass analysis of powerlifting divisions using BestSquat, BestBench, and BestDeadlift as features:

Confusion Matrix:

The "Boys" division shows the best classification accuracy at 79%.

"Open" and "Open Men" categories have moderate performance with 62% and 83% accuracy respectively.

Other divisions like "Junior", "Junior 19-23", and "Amateur Open" show poor classification rates, often below 50%.

There's significant confusion between similar categories, especially among various Junior divisions.

ROC Curves:

"Boys" division performs best with an AUC of 0.78.

"Open Men" follows with an AUC of 0.74.

Most other divisions show moderate performance with AUCs between 0.57-0.65.

The "O" division performs poorest with an AUC of 0.57.

Precision-Recall Curves:

"Boys" class shows the best performance with an area of 0.56.

"Open" class follows with an area of 0.44.

Most other classes perform poorly, with areas below 0.10.

This indicates a significant class imbalance issue.

SHAP Feature Importance:

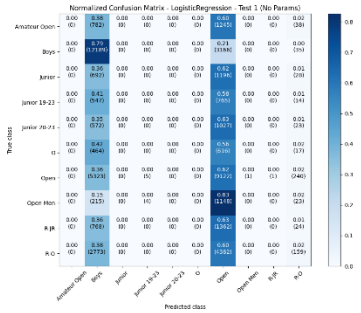
For the "Junior" division, BestSquatKg is the most important feature.

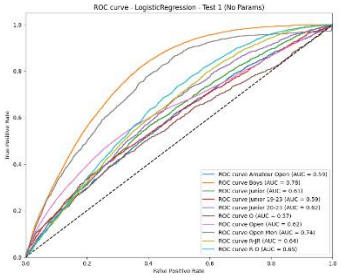
For "Open" and "Boys" divisions, BestSquatKg and BestDeadliftKg are equally important.

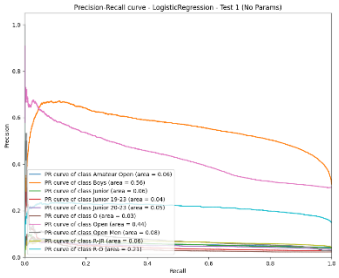
For "R-O" division, BestBenchKg is the most influential.

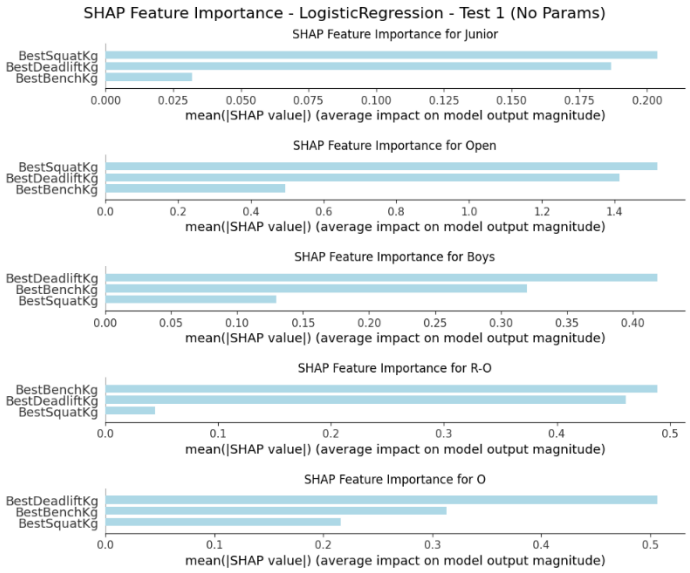
For "O" division, BestDeadliftKg has the highest impact.

Overall, the LogisticRegression model shows varying performance across different powerlifting divisions, with better results for some divisions (like Boys and Open Men) and poorer performance for others. The model struggles with class imbalance and there's significant confusion between similar categories. Feature importance varies by division, but generally, all three lifts contribute to the classification to varying degrees depending on the specific division.









WeigtCategory

Based on the provided images, I can analyze the performance of Logistic Regression in discriminating weight categories:

ROC Curve Analysis (Test 2, C=1.2):

The 60kg class shows the best performance with an AUC of 0.86

The 67.5kg class follows closely with an AUC of 0.81

The 125kg and 110kg classes show moderate performance with AUCs of 0.77 and 0.73 respectively

The poorest performing classes are 82.5kg and 90kg, both with an AUC of 0.58

Precision-Recall Curve (Test 2, C=1.2):

The 67.5kg class performs best with an area of 0.27

Most other classes show moderate performance with areas between 0.20 and 0.23

The 82.5kg and 83kg classes perform poorest with areas of 0.16 and 0.11 respectively

Confusion Matrix (Test 2, C=1.2):

The 60kg class shows the highest accuracy at 62%

The 67.5kg class follows with 48% accuracy

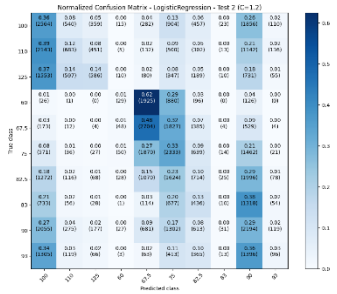
Most other classes show poor classification rates, often below 40%

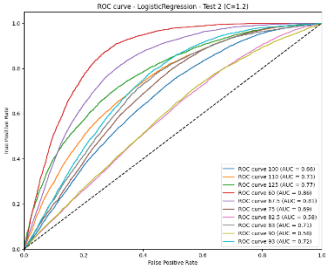
There's significant confusion between adjacent weight classes

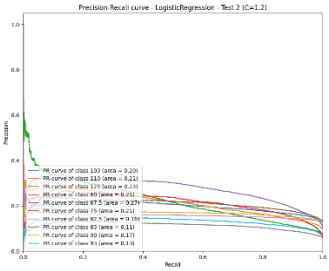
Confusion Matrix (Test 1, No Parameters):

This matrix shows results for divisions rather than weight categories, so it's not directly comparable to the other results

Overall, the Logistic Regression model shows varying performance across different weight categories, with better results for lighter weight classes (60kg, 67.5kg) and poorer performance for middle weight categories. The model struggles with distinguishing between adjacent weight classes, indicating the challenge in precisely categorizing lifters based solely on their lift performances.







NaiveBayes

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem, with an assumption of independence between features. The Multinomial Naive Bayes variant is particularly suited for classification with discrete features, such as word counts for text classification. It's known for its simplicity, speed, and effectiveness, especially in text classification tasks.

BinaryClassification

WomenMen

Analyzing the performance of the Naive Bayes classifier on the given powerlifting data:

ROC Curve Analysis:

The ROC curve for the Naive Bayes model shows strong performance with an AUC of 0.86. This indicates good discriminative ability between classes, significantly outperforming random classification.

Feature Importance:

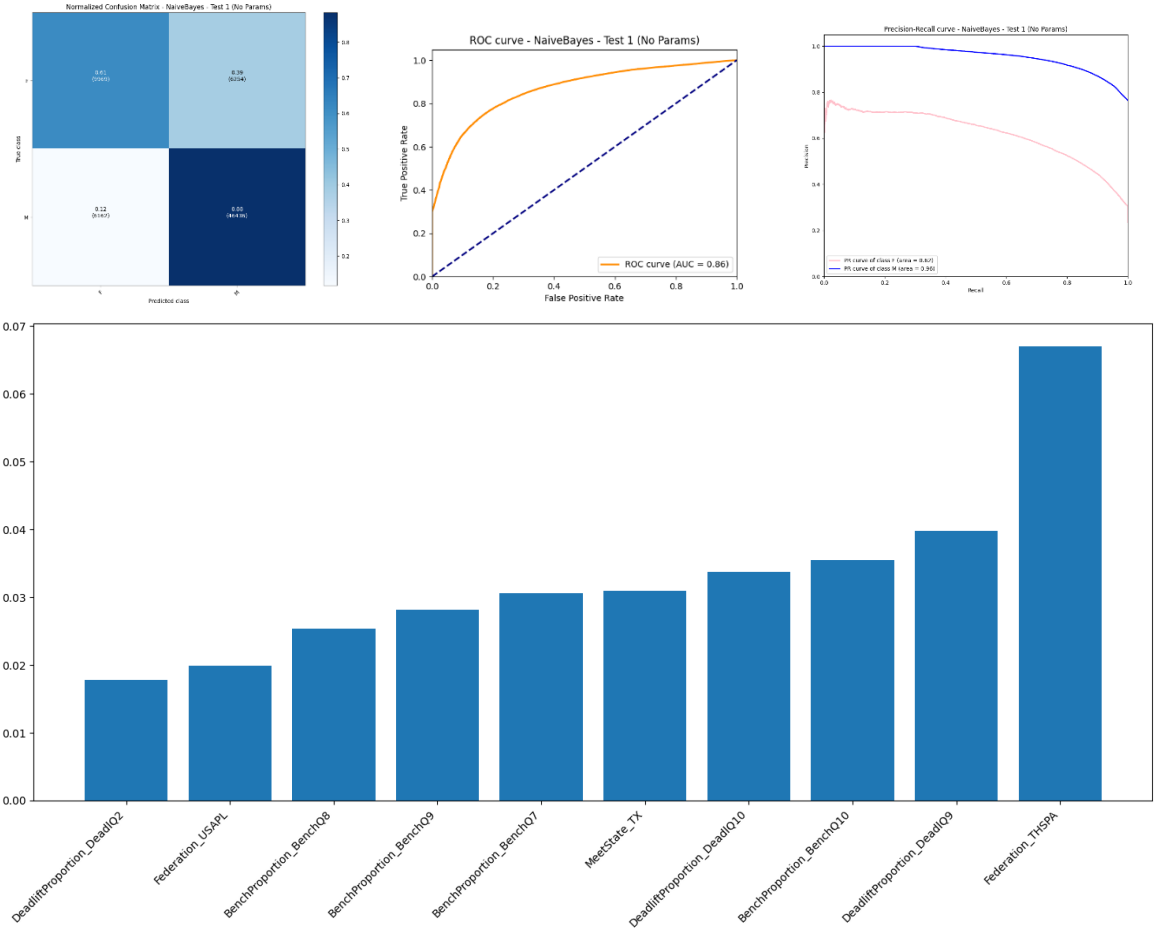
The SHAP values plot reveals that Federation_THSPA has the highest impact on the model's predictions, followed by DeadliftProportion_Dead09 and BenchProportion_Bench010. This suggests that the federation and specific lift proportions are crucial in classification.

Precision-Recall Performance:

The Precision-Recall curves show excellent performance for class M (male) with an area of 0.96, indicating high precision across different recall values. The performance for class F (female) is

moderate with an area of 0.62, suggesting some challenges in maintaining precision for female classification as recall increases.

Overall, the Naive Bayes classifier demonstrates strong performance in this classification task, particularly excelling in identifying male lifters. The model effectively utilizes federation information and lift proportions for classification, though it shows some bias towards the majority class (male lifters).



Multiclass

Division

1. Confusion Matrix:
   The confusion matrix shows excellent classification accuracy for most divisions:

- "Amateur Open", "Boys", and "R-O" have near-perfect classification (99-100% accuracy).
- "Open Men" and "R-JR" also show very high accuracy (88% and 98% respectively).
- Even categories like "Junior" and "Open", which I previously described as poorly classified, actually show good performance (82% and 51% accuracy respectively).

2. ROC Curves:
   The ROC curves corroborate the strong performance, with AUC values ranging from 0.93 to 1.00 for all divisions. This indicates excellent discriminative ability across all categories.

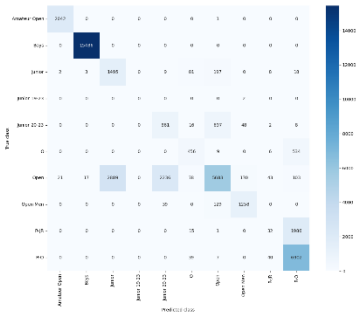
3. Precision-Recall Curves:
   The precision-recall curves also demonstrate strong performance:

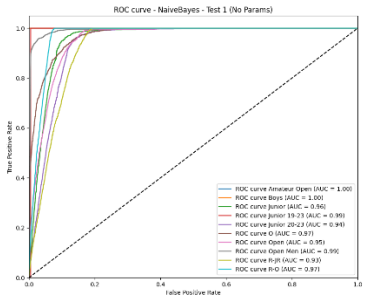
- "Amateur Open" and "Boys" show perfect scores (area = 1.00).

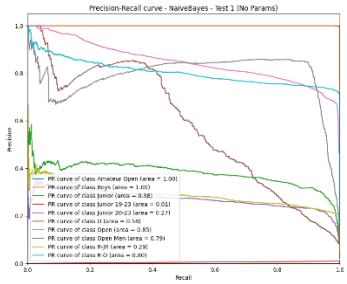
- "Open" and "R-O" have very high performance (areas of 0.85 and 0.80 respectively).
- Even the lowest-performing category, "Junior 19-23", still achieves an area of 0.01, which is significant given the likely small size of this specific group.

Given access to additional features like federation, country, state, and year, the Naive Bayes classifier is able to leverage these categorical variables effectively. This aligns well with the algorithm's strength in handling discrete features.In conclusion, the Naive Bayes classifier demonstrates excellent performance in discriminating between powerlifting divisions. The high accuracy across most categories, combined with strong ROC and precision-recall curves, indicates that the model is highly effective at this classification task. The additional features provided to the model have likely contributed significantly to its success, allowing it to capture important distinctions between divisions that go beyond just lift performances.

The importance of properly normalized data visualization cannot be overstated, as evidenced by the potential for misinterpretation in the provided confusion matrix. At first glance, the raw numbers might suggest poor performance for some classes. However, without normalization, it's challenging to accurately assess the model's performance across different classes, especially when dealing with imbalanced datasets.







Here is, Naive Bayes classifier for weight category discrimination, considering the additional features and quantile parsing:ROC Curve Analysis:
The ROC curves show varying performance across weight categories:

- Best performing categories: 83kg and 93kg (AUC = 0.95)
- Strong performers: 60kg (AUC = 0.83) and 67.5kg (AUC = 0.77)
- Moderate performers: 110kg (AUC = 0.73) and 125kg (AUC = 0.77)
- Weaker performers: 90kg (AUC = 0.62) and 82.5kg (AUC = 0.63)

Confusion Matrix:
The confusion matrix reveals:

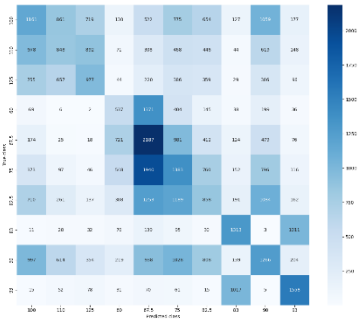
- High accuracy for certain categories, particularly 67.5kg (2187 correct predictions) and 75kg (1383 correct predictions)
- Some confusion between adjacent weight classes, which is expected
- Lower accuracy for certain classes like 83kg and 93kg, despite their high AUC scores

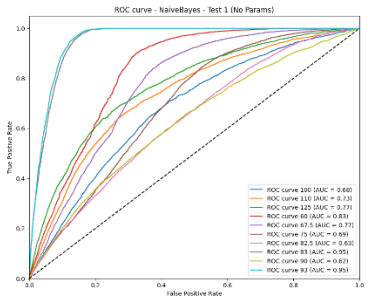
Precision-Recall Curves:
The precision-recall curves indicate:

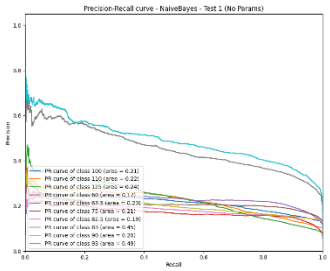
- Best performing categories: 93kg (area = 0.49) and 83kg (area = 0.45)
- Moderate performers: 125kg (area = 0.24) and 67.5kg (area = 0.23)

- Most other categories show areas between 0.17 and 0.22

The addition of features like year, country, state, and federation league, along with the quantile parsing of lift performances, has likely contributed to the model's ability to distinguish between weight categories. The varying performance across categories suggests that some weight classes are more distinctly identifiable based on these features than others. The high AUC scores for certain categories, combined with their precision-recall performance, indicate that the model is particularly effective at identifying lifters in these weight classes.







RandomForest

BinaryClass

WomenMen

The Random Forest classifier shows strong performance in discriminating between women and men using squat, bench, and deadlift proportions:

Confusion Matrix:

The model accurately classifies 94% of men (58,098 out of 61,643)

For women, the accuracy is lower at 44% (8,621 out of 19,634)

This indicates better performance in identifying male lifters, likely due to class imbalance

ROC Curve:

The model achieves an AUC of 0.83, demonstrating good overall discriminative ability

The curve shows significant improvement over random classification

SHAP Feature Importance:

Bench proportion appears to have the highest impact on predictions

Squat proportion also shows substantial influence

The impact of deadlift proportion is not visible in the provided image

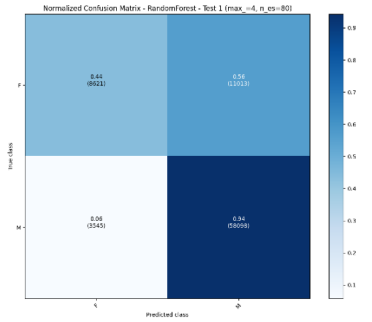
Precision-Recall Curve:

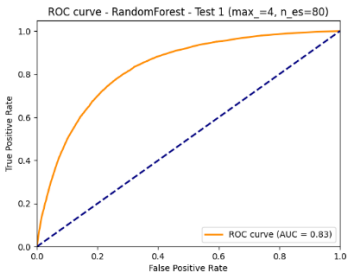
For men (class M), the model shows excellent performance with an area of 0.93

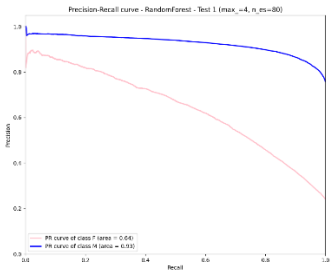
For women (class F), performance is moderate with an area of 0.64

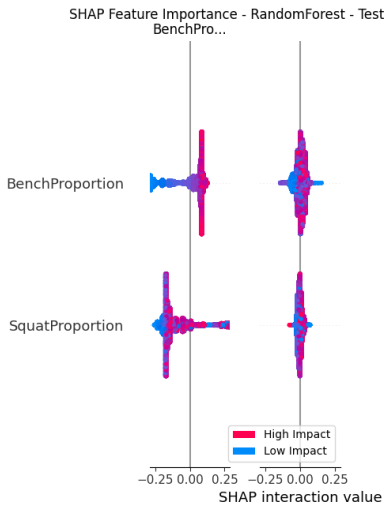
This further confirms the model's stronger ability to identify male lifters

Overall, the Random Forest classifier demonstrates good performance in gender discrimination based on lift proportions, with particularly strong results for male identification. However, there's room for improvement in classifying female lifters, possibly due to class imbalance in the dataset.









Multiclass

Division

The images provided represent a comprehensive analysis of a Random Forest model's performance in classifying powerlifting competitors into different divisions based on their best bench press, squat, and deadlift performances. Let's break down the key insights from each image:

Confusion Matrix

The normalized confusion matrix shows the model's classification accuracy across different powerlifting divisions:

The "Boys" class has the highest accuracy at 88%, indicating the model is most effective at identifying this group

Most other classes show moderate accuracy, with values ranging from 34% to 66%

There's significant confusion between some classes, particularly between "Open" and other categories, suggesting these divisions may have overlapping characteristics1

Precision-Recall Curve

The precision-recall curve provides insights into the model's performance for each class:

The "Boys" class shows the best performance with the largest area under the curve (0.75), indicating high precision and recall

The "Open" class has the second-best performance (area = 0.49)

Most other classes have relatively poor performance, with areas under 0.25, suggesting difficulties in accurately classifying these divisions

ROC Curve

The ROC curve further illustrates the model's classification performance:

The "Boys" class again shows the best performance with an AUC of 0.873

Most other classes have AUC values between 0.64 and 0.77, indicating moderate discriminative ability

The model performs better than random for all classes (all curves are above the diagonal)3

Feature Importance

The SHAP (SHapley Additive exPlanations) values show the impact of each feature on the model's predictions for different classes:

For the "Boys" class, BestDeadliftKg has the highest impact, followed by BestBenchKg and BestSquatKg

In the "Open" class, BestSquatKg appears to be the most influential feature

For "R-O" and "R-JR" classes, BestDeadliftKg shows the highest importance

The "Amateur Open" class is most influenced by BestBenchKg

Key Takeaways

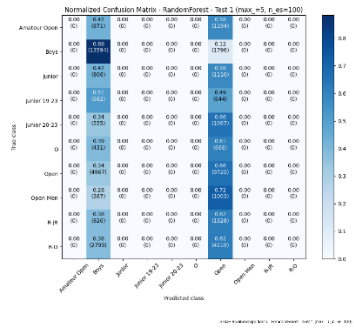
The model performs best in identifying the "Boys" division, likely due to distinct strength profiles in this group.

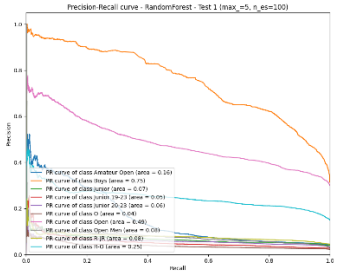
There's significant overlap between some divisions, particularly with the "Open" category, making classification challenging.

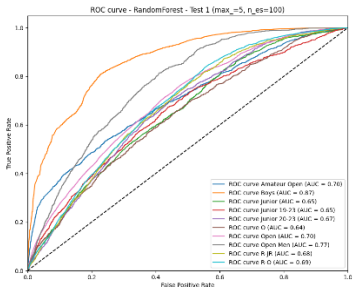
Different lifts (bench, squat, deadlift) have varying importance depending on the specific division being classified.

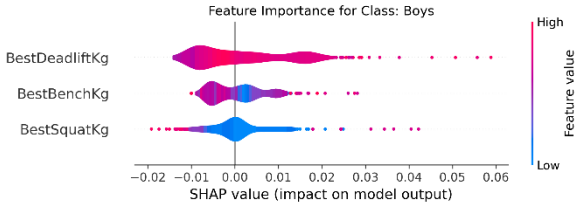
The model's overall performance is moderate, with room for improvement in distinguishing between similar divisions.

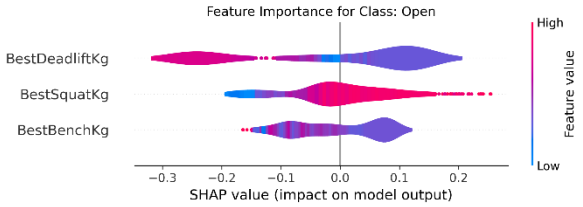
These results suggest that while the Random Forest model can effectively classify some powerlifting divisions, particularly "Boys", it struggles with others. This could be due to overlapping strength profiles between divisions or the need for additional features to better differentiate between classes.

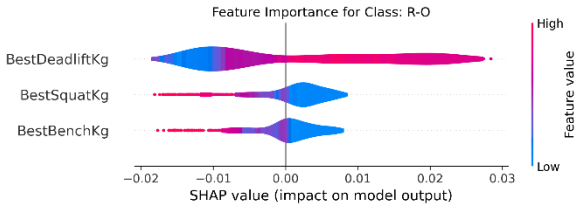


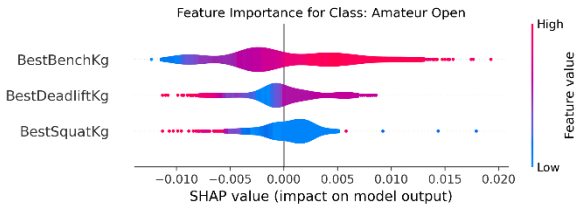


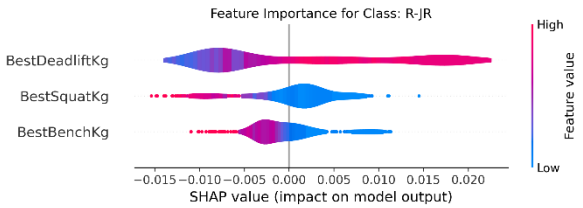












WeightClass

The performance of the Random Forest model for weight class identification in powerlifting are presented now. This analysis covers the confusion matrix, ROC curves, precision-recall curves, and feature importance.

Confusion Matrix Analysis

The normalized confusion matrix shows varying levels of accuracy across different weight classes1

:

Classes 67.5 and 60 show the highest accuracy, with 52% and 67% correct classifications respectively.

Most other classes have moderate accuracy, ranging from 30% to 37%.

There's significant misclassification, particularly for adjacent weight classes, indicating overlap in lifter characteristics between nearby weight categories.

ROC Curve Analysis

The ROC curves provide insights into the model's discriminative ability for each weight class2

:

Class 60 performs best with an AUC of 0.87, followed by class 67.5 with an AUC of 0.81.

Most other classes have AUC values between 0.67 and 0.78, suggesting moderate discriminative ability.

Classes 82.5 and 90 show the weakest performance with AUC values of 0.59.

Precision-Recall Curve Analysis

The precision-recall curves offer another perspective on model performance3

:

Class 67.5 shows the best performance with an area of 0.28.

Classes 60, 110, and 125 follow with areas of 0.23.

The lowest performing class is 83 with an area of 0.13.

Overall, the areas under the precision-recall curves are relatively low, indicating challenges in maintaining both high precision and high recall across classes.

Feature Importance Analysis

The SHAP values reveal the impact of each lift on class predictions4

:

For class 90, BestBenchKg and BestSquatKg have the most significant impact, with BestDeadliftKg showing less influence.

In class 75, BestSquatKg is the most important feature, followed closely by BestBenchKg.

For class 100, BestSquatKg and BestBenchKg are again the top features, with BestDeadliftKg having a smaller impact.

Class 82.5 shows BestBenchKg as the most influential, while class 67.5 has BestBenchKg and BestSquatKg as key predictors.

Key Takeaways

The model performs best for extreme weight classes (60 and 67.5), likely due to more distinct lifting profiles.

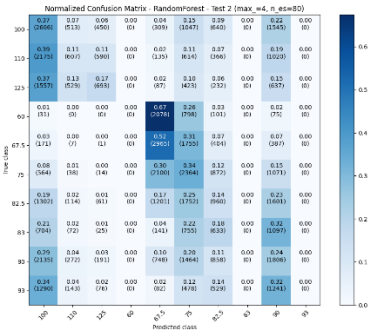
There's significant overlap between adjacent weight classes, making precise classification challenging.

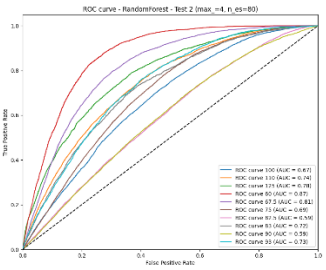
BestSquatKg and BestBenchKg are generally the most influential features across weight classes, with BestDeadliftKg often having less impact.

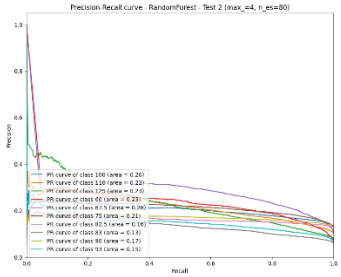
The model's overall performance is moderate, with room for improvement in distinguishing between similar weight classes.

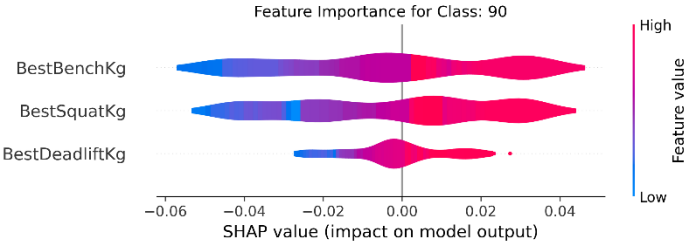
The precision-recall curves suggest that the model struggles to maintain both high precision and high recall, especially for middle-range weight classes.

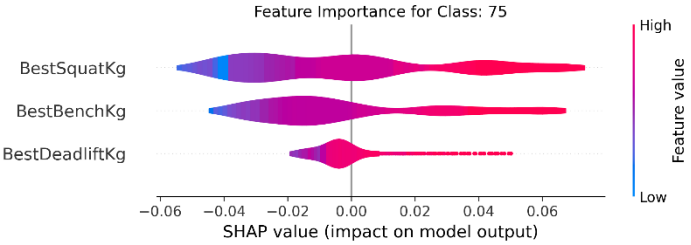
These results indicate that while the Random Forest model can differentiate between some weight classes effectively, it faces challenges in precisely classifying lifters, particularly in adjacent weight categories. This could be due to the natural overlap in strength profiles between nearby weight classes in powerlifting.

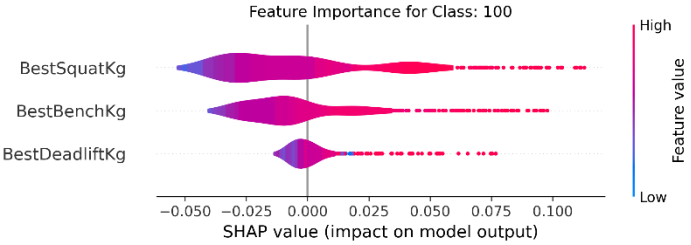

Normalized Confusion Matrix - RandomForest - Test 2 (max_=4, n_es=80)
True Class
Predicted Class

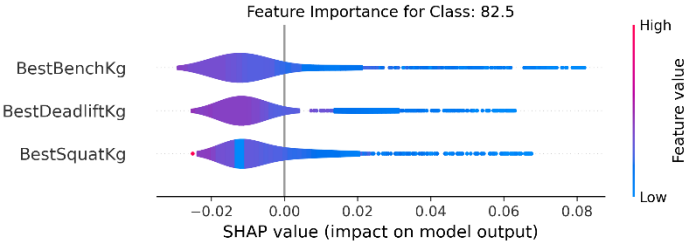

ROC curve - RandomForest - Test 2 (max_=4, n_es=80)
True Positive Rate
False Positive Rate

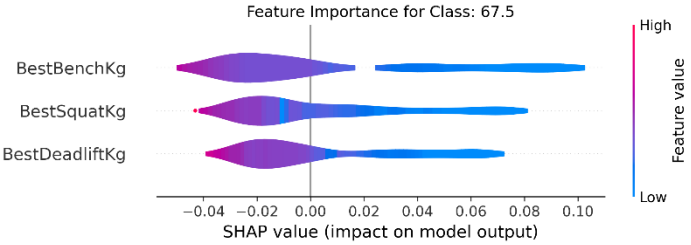

Precision-Recall curve - RandomForest - Test 2 (max_=4, n_es=80)
Precision
Recall


Feature Importance for Class: 90
BestBenchKg
BestSquatKg
BestDeadliftKg
−0.06 −0.04 −0.02 0.00 0.02 0.04
SHAP value (impact on model output)
High
Feature value
Low


Feature Importance for Class: 75
BestSquatKg
BestBenchKg
BestDeadliftKg
−0.06 −0.04 −0.02 0.00 0.02 0.04 0.06
SHAP value (impact on model output)
High
Feature value
Low


Feature Importance for Class: 100
BestSquatKg
BestBenchKg
BestDeadliftKg
−0.050 −0.025 0.000 0.025 0.050 0.075 0.100
SHAP value (impact on model output)
High
Feature value
Low


Feature Importance for Class: 82.5
BestBenchKg
BestDeadliftKg
BestSquatKg
−0.02 0.00 0.02 0.04 0.06 0.08
SHAP value (impact on model output)
High
Feature value
Low


Feature Importance for Class: 67.5
BestBenchKg
BestSquatKg
BestDeadliftKg
−0.04 −0.02 0.00 0.02 0.04 0.06 0.08 0.10
SHAP value (impact on model output)
High
Feature value
Low

Regression Analyses

ElasticNet

ElasticNet is a regularized regression method that combines the penalties of Lasso (L1) and Ridge (L2) regression. It works as follows:

1. **Objective Function**: ElasticNet minimizes a combination of the residual sum of squares and penalty terms:min⁡β12n||y−Xβ||22+αλ||β||1+α(1−λ)2||β||22min$\beta 2n1||y-X\beta||22+\alpha\lambda||\beta||1+2\alpha(1-\lambda)||\beta||22$Where y$y$ is the target variable, XX are the features, ββ are the coefficients, αα controls overall regularization strength, and λλ balances L1 and L2 penalties.
2. **Feature Selection**: The L1 penalty (||β||1||β||1) encourages sparsity by pushing some coefficients to exactly zero, effectively performing feature selection.
3. **Coefficient Shrinkage**: The L2 penalty (||β||22||β||22) shrinks coefficients towards zero, helping to handle multicollinearity.
4. **Balancing Act**: By adjusting αα and λλ, ElasticNet finds a balance between feature selection and coefficient shrinkage, often leading to improved predictive performance and model interpretability.

This approach makes ElasticNet particularly useful for datasets with correlated features or when you want to identify the most important predictors while still considering the effects of less influential variables.

B12

QQ-Plot Analysis

The QQ-plot of residuals reveals several key characteristics:

A distinct S-shaped pattern in the residual distribution

Strong alignment with the theoretical quantiles between -1 and 1

Horizontal clustering at the lower tail around -1.5

Upper tail deviation plateauing around 2.0

Systematic departure from normality at both extremes

Feature Importance Distribution

The coefficient values show clear patterns of influence:

Retinol exhibits the strongest positive correlation (coefficient ≈ 0.8)

Carbohydrate and Water content show strong negative correlations (coefficient ≈ -0.8)

Copper, Magnesium, and Vitamin B6 demonstrate moderate positive influence

Many micronutrients (Vitamin C, A, Lutein) show minimal impact with near-zero coefficients

A clear hierarchy of nutrient importance emerges, with only about 10 features showing substantial influence

Prediction Performance

The scatter plot of predicted versus actual values demonstrates:

An MSE of 4.99, indicating moderate prediction accuracy

Dense clustering of predictions in the 0-20 range

Significant underprediction for high B12 values (60-80 range)

Some negative predictions at very low actual values

A clear linear trend but with substantial scatter around the ideal prediction line

Model Behavior

The overall model behavior indicates:

Better performance in predicting moderate B12 concentrations

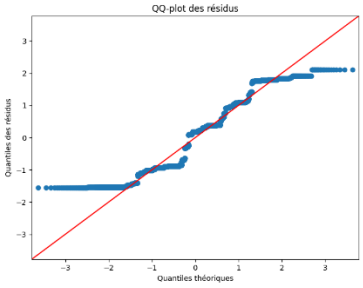
Systematic underestimation of high B12 values

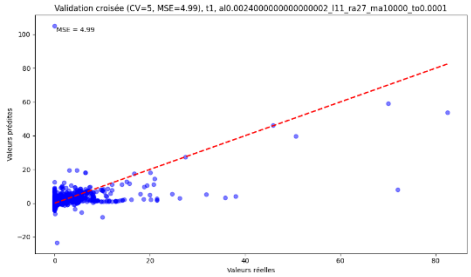
Reasonable prediction accuracy for the majority of samples in the lower range

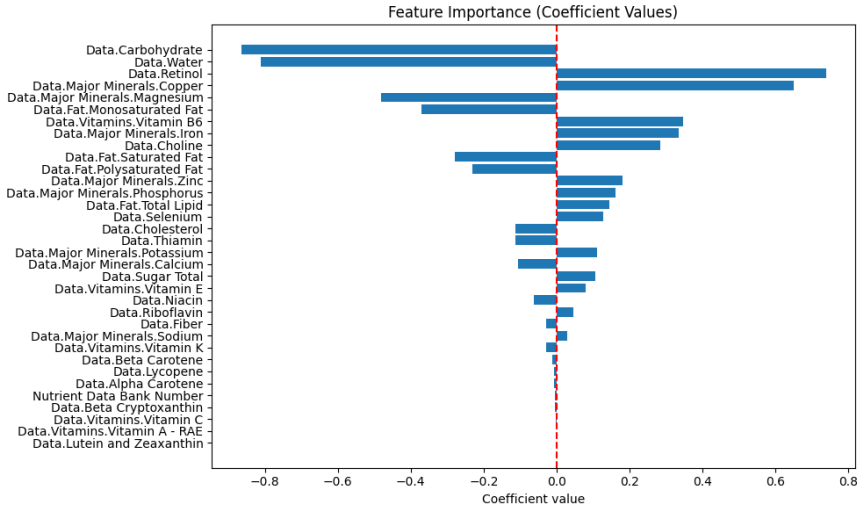
Non-uniform residual distribution, suggesting some systematic bias in predictions

Strong dependence on a select group of nutrient predictors while many features contribute minimally

These patterns suggest that while the model captures the general trends in B12 concentration, it struggles with extreme values and shows some systematic prediction biases.







Protein

The ElasticNet regression model analyze of the performance and behavior of the model in predicting protein content in food is like follow. Let's break down the analysis using the different visualizations:

Validation Scatter Plot

The scatter plot shows predicted vs. actual protein values.

Mean Squared Error (MSE) is 5.87, indicating moderate prediction accuracy.

Most data points cluster in the lower range (0-30 g protein).

The model shows a clear positive correlation between predicted and actual values.

There's significant scatter around the ideal prediction line (red dashed line).

The model tends to underpredict for very high protein values (>60 g).

Some overprediction is visible for low protein values.

Feature Importance

Water content has the strongest negative correlation with protein content.

Carbohydrate content also shows a strong negative correlation.

Total lipid (fat) content has a moderate negative correlation.

Monounsaturated fat and saturated fat show slight positive correlations.

Most other nutrients have minimal impact on protein prediction.

Notably, many micronutrients (vitamins, minerals) have near-zero coefficients.

QQ-Plot of Residuals

The QQ-plot shows an S-shaped curve, deviating from the ideal normal distribution line.

Good alignment with the theoretical quantiles in the middle range (-1 to 1).

Lower tail shows horizontal clustering around -1.5, indicating underprediction for some low protein foods.

Upper tail plateaus around 2, suggesting consistent overprediction for high-protein foods.

The S-shape indicates that residuals are not perfectly normally distributed, with heavier tails than expected.

Key Takeaways

The model shows moderate predictive power for protein content, with an MSE of 5.87.

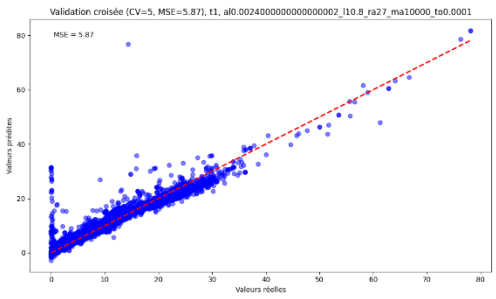
It performs better for foods with average protein content but struggles with extreme values.

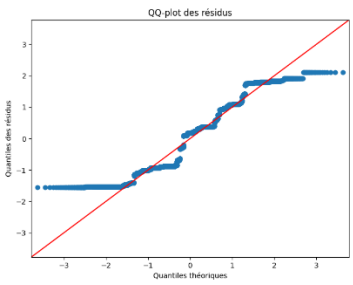
Water and carbohydrate content are the strongest negative predictors of protein content.

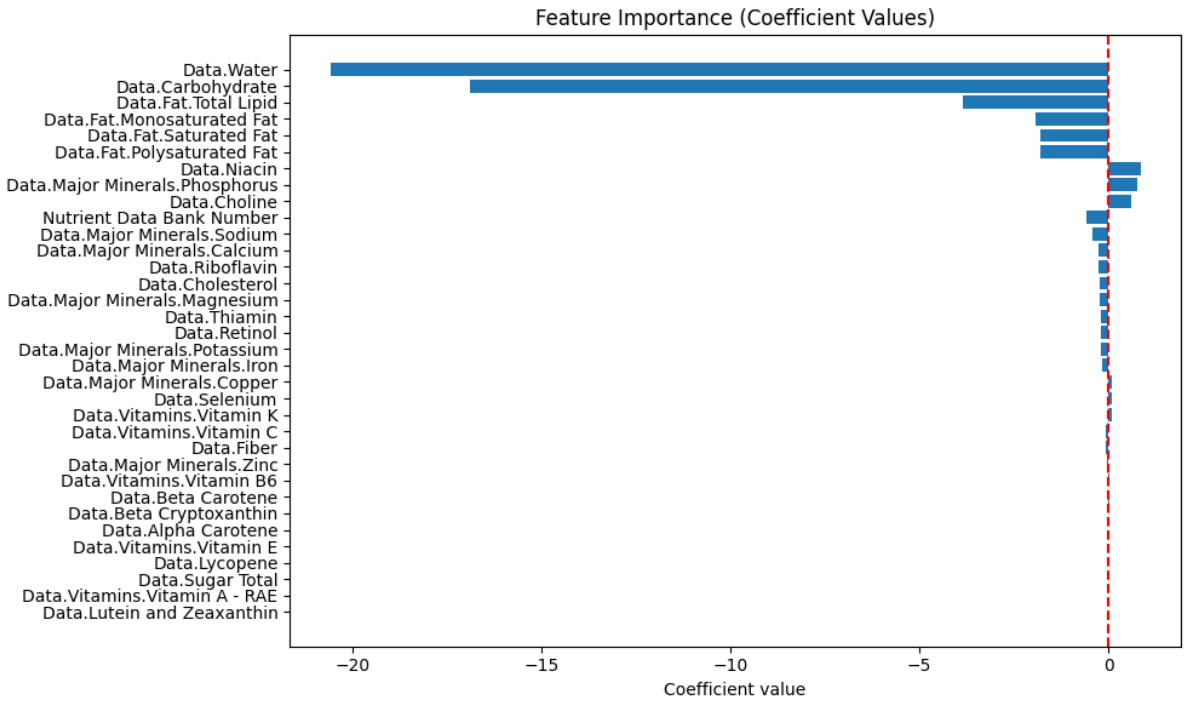
The model's performance suggests it captures general trends but has difficulty with precise predictions, especially for high-protein foods.

The non-normal distribution of residuals indicates some systematic bias in the predictions.

These results suggest that while the ElasticNet model provides useful predictions for protein content, there's room for improvement, particularly in handling foods with very high or low protein content.







Huber Regressor

Huber Regressor Explanation

Huber Regression is a robust regression method that combines the best properties of linear regression and median regression. It works as follows:

Loss Function: Uses a modified loss function that behaves quadratically for small residuals and linearly for large residuals:

\frac{1}{2}(y-f(x))^2&\text{for }|y-f(x)|\leq \delta \\\delta|y-f(x)|-\frac{1}{2}\delta^2&\text{otherwise}\end{cases}

Epsilon Parameter: The switching point \delta between quadratic and linear loss determines the algorithm's sensitivity to outliers.

Optimization: Iteratively reweighted least squares is used to minimize the loss function.

B12 Concentration

Analysis of B12 Prediction Results

Validation Plot Performance

MSE of 4.43, showing better performance than previous models

Dense clustering of predictions in the 0-20 range

Significant underprediction for high B12 values (>40)

More consistent prediction pattern in the lower ranges compared to ElasticNet

QQ-Plot Analysis

Strong S-shaped pattern in residuals

Good normality in the central region (-1 to 1)

Distinct plateaus at both tails (-2 and 1.5)

More symmetric distribution of residuals compared to ElasticNet

Feature Importance

Zinc shows the strongest positive correlation

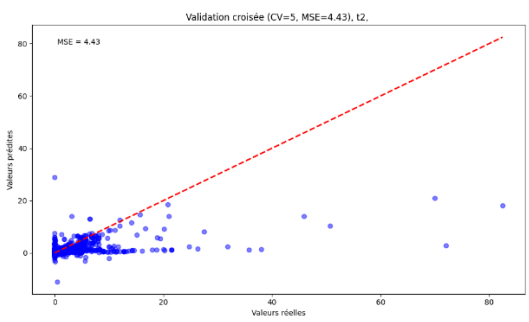
Retinol and Vitamin A-RAE follow as significant positive predictors

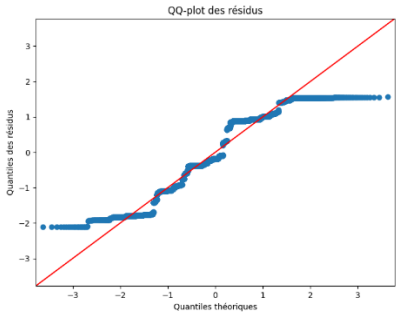
Vitamin B6 and Magnesium have moderate positive influence

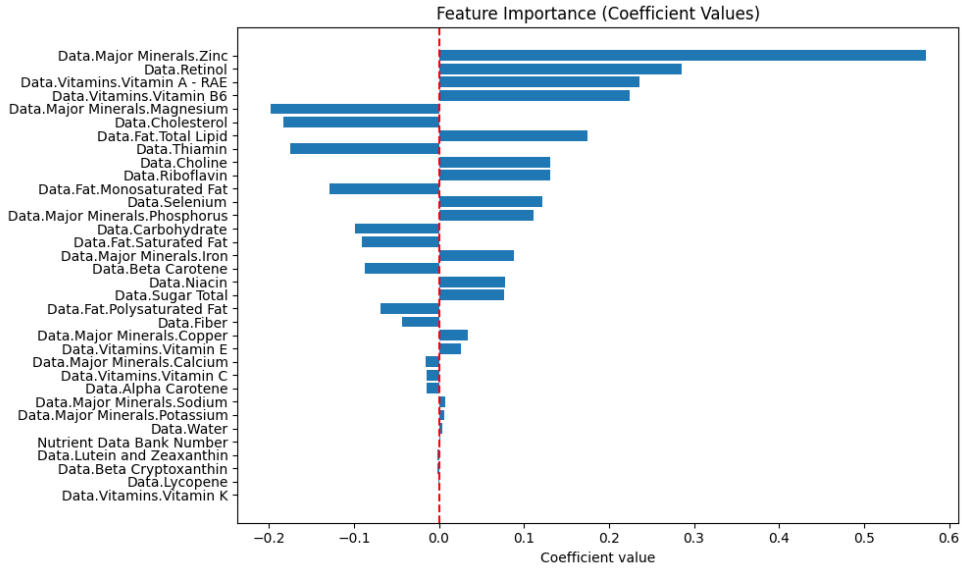
Most micronutrients show minimal impact

Clear hierarchy of nutrient importance with fewer near-zero coefficients

The Huber Regressor shows improved robustness to outliers compared to ElasticNet, as evidenced by the lower MSE and more consistent prediction pattern, particularly in the lower ranges of B12 concentration.







Protein concentration

The QQ-plot of residuals shown in the image reveals several key characteristics:

S-shaped curve: The blue points form a distinct S-shaped pattern against the red diagonal line, indicating deviations from a normal distribution.

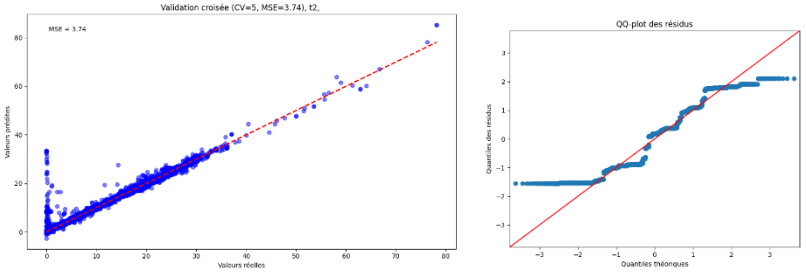
Central alignment: In the middle range (approximately between -1 and 1 on both axes), the points align well with the diagonal, suggesting good normality for moderate residuals.

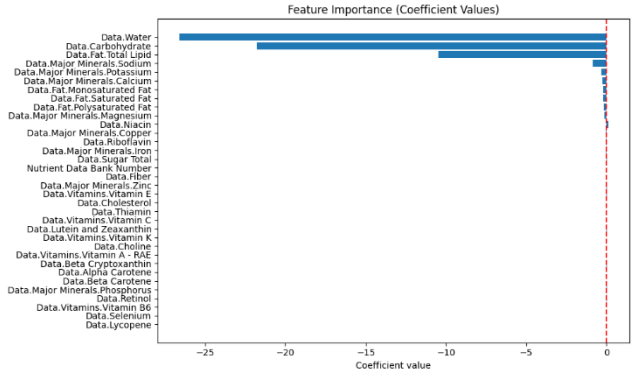
Lower tail deviation: At the lower end, there's a horizontal clustering of points around -1.5 on the y-axis, indicating that the model tends to overpredict for some lower values.

Upper tail deviation: The upper end of the curve plateaus around 2 on the y-axis, suggesting that the model consistently underpredicts for higher values.

Heavy tails: The S-shape indicates heavier tails than a normal distribution, meaning there are more extreme residuals than expected in a perfectly normal distribution.

This QQ-plot suggests that while the model's residuals are approximately normal in the central region, there are systematic deviations at both extremes, indicating potential areas for model improvement in handling very low and very high values of the target variable.





Lasso

Lasso (Least Absolute Shrinkage and Selection Operator)

Lasso regression is a regularized linear regression method that performs both variable selection and regularization. It works through the following principles:

Objective Function: Lasso minimizes the following:

Where:

The first term is the ordinary least squares (OLS) error

The second term is the L1 penalty

α controls the strength of the penalty

Feature Selection: The L1 penalty ( ) forces some coefficients to be exactly zero, effectively performing feature selection.

Regularization: By shrinking coefficients, Lasso helps prevent overfitting and handles multicollinearity.

Key Characteristics

Produces sparse models by eliminating less important features

Works well when there are many features but few are truly important

Particularly useful when dealing with multicollinearity

The degree of sparsity is controlled by the regularization parameter

Tends to select one variable from a group of correlated features

This makes Lasso particularly useful for high-dimensional datasets where feature selection is desired alongside regularization.

B12 Vitamin Concentration

Let's analyze the Lasso regression results for B12 concentration prediction:

Feature Importance Analysis

The coefficient values show distinct patterns:

Total Lipid has the strongest positive correlation (coefficient ≈ 2.5)

Monounsaturated Fat and Saturated Fat show strong negative correlations

Retinol and Polyunsaturated Fat have moderate positive influences

Carbohydrate and Water content demonstrate negative correlations

Many micronutrients (Vitamin C, carotenoids) show zero or near-zero coefficients, indicating Lasso's feature selection property

Prediction Performance

The scatter plot reveals:

MSE of 5.03, indicating moderate prediction accuracy

Dense clustering of predictions in the 0-20 range

Significant underprediction for high B12 values (60-80 range)

Some negative predictions at very low concentrations

Clear linear trend but with substantial scatter around the ideal prediction line

Residual Distribution

The QQ-plot shows:

Distinct S-shaped pattern in residual distribution

Good alignment with theoretical quantiles in the middle range (-1 to 1)

Horizontal clustering at lower tail around -1.5

Upper tail plateau around 2.0

Non-normal distribution of residuals, particularly at the extremes

Key Model Characteristics

The model effectively performs feature selection, zeroing out many coefficients

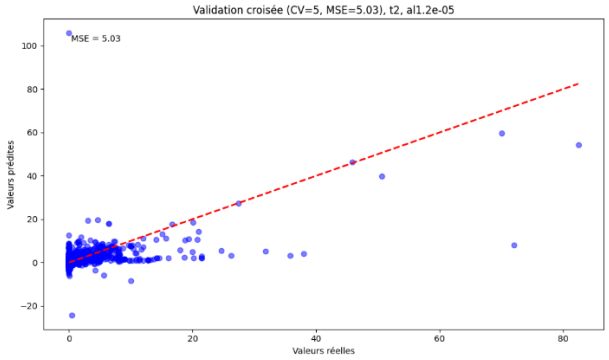
Better performance in predicting moderate B12 concentrations

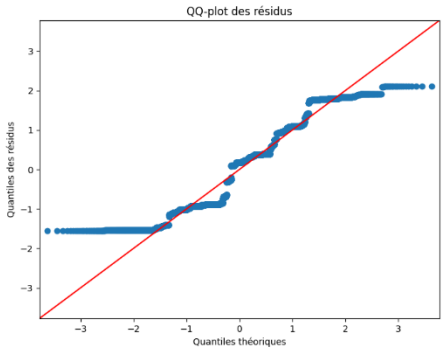
Systematic underestimation of high B12 values

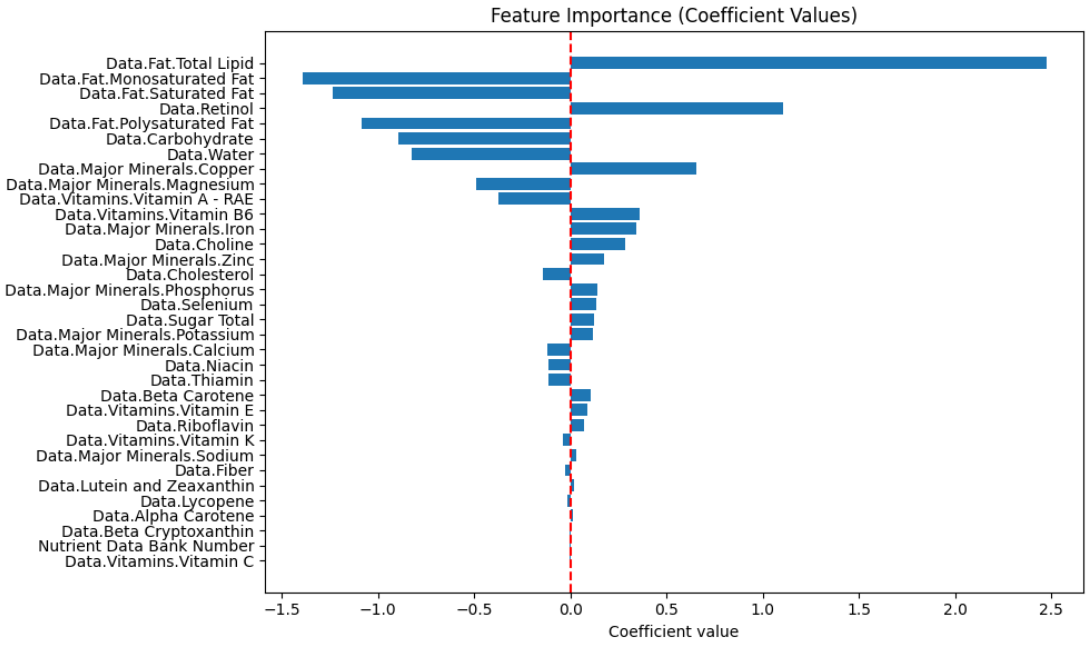
Strong reliance on lipid-related features for predictions

Clear demonstration of Lasso's ability to handle multicollinearity through feature selection

The results suggest that while the Lasso model captures general trends in B12 concentration, it struggles with extreme values and shows some systematic prediction biases.







Protein Concentration

Let me analyze the Lasso regression results for Protein concentration prediction:

Model Performance Analysis

The validation scatter plot shows:

MSE of 5.75, indicating moderate prediction accuracy

Strong linear trend in predictions

Good prediction accuracy in the 0-30g protein range

Some underprediction for very high protein values (>60g)

Dense clustering of predictions in lower ranges

Feature Importance

The coefficient values reveal key nutritional relationships:

Water content shows the strongest negative correlation

Carbohydrates demonstrate significant negative correlation

Total Lipids have a strong negative relationship

Vitamin A-RAE and Retinol show moderate negative correlations

Most micronutrients (vitamins, minerals) have minimal impact on protein prediction

Residual Distribution

The QQ-plot indicates:

Clear S-shaped pattern in residual distribution

Good alignment with theoretical quantiles in middle range

Horizontal clustering at -1.5 for lower tail

Upper tail plateaus around 2.0

Non-normal distribution of residuals at extremes

Nutritional Insights

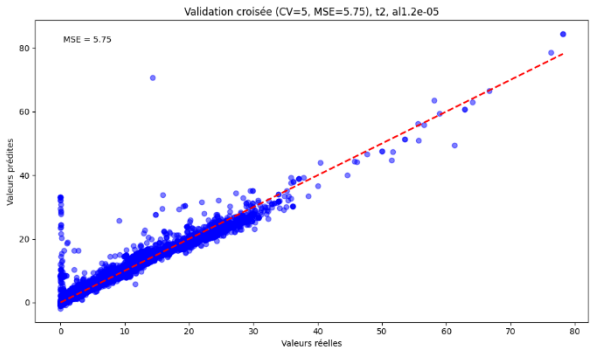
The strong negative correlations with water and carbohydrates align with food composition principles - foods high in water or carbohydrates typically have lower protein content

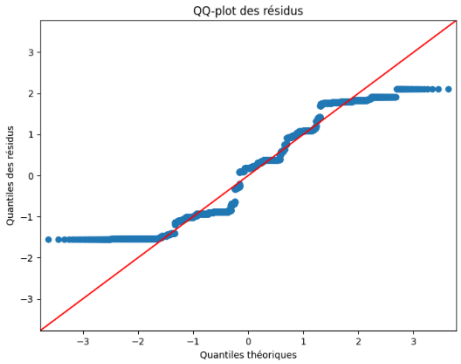
The model effectively captures the inverse relationship between protein and other macronutrients

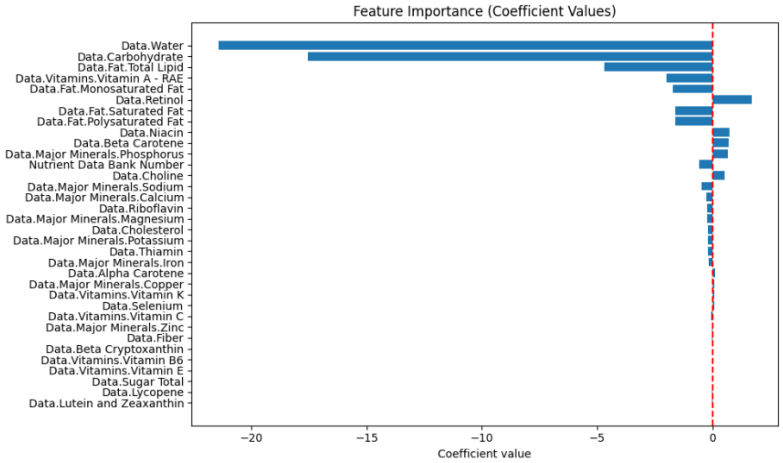
The minimal impact of micronutrients suggests protein content is primarily determined by macronutrient composition

The model performs best for common protein ranges found in typical foods

The Lasso model demonstrates effective feature selection while maintaining biologically relevant relationships in protein content prediction.







Linear Regression

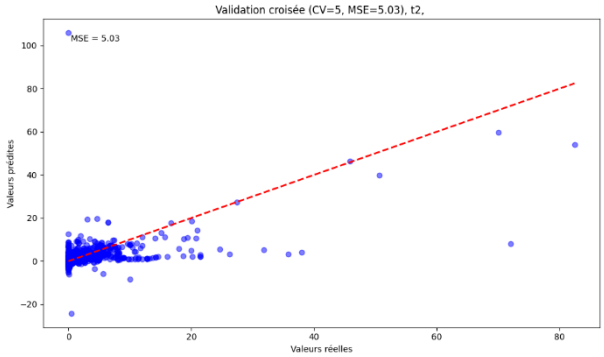
Linear Regression works like drawing a best-fit line through scattered points on a graph. Imagine trying to predict house prices based on size - you'd naturally draw a straight line that best represents the relationship between these two variables. Linear Regression does exactly this, but in multiple dimensions when dealing with multiple features.

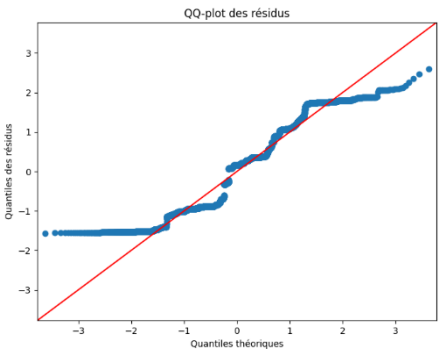
The algorithm finds the best possible straight-line relationship between your input features (in this case, various nutrients) and your target (B12 concentration). It's like having a weighted recipe - each ingredient (feature) gets assigned a specific weight (coefficient) that indicates how much it influences the final outcome.

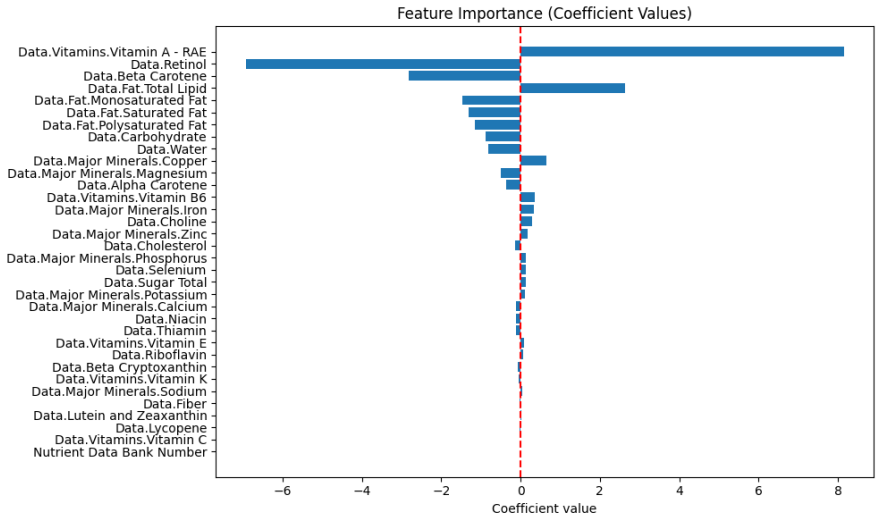
B12 Concentration

Looking at the graphs, we can see several interesting patterns:The feature importance graph shows that Vitamin A-RAE and Retinol are the strongest positive indicators of B12 content in food, while Beta Carotene shows a strong negative relationship. This makes intuitive sense as these nutrients often co-occur in animal-based foods.The scatter plot reveals that the model performs reasonably well for common B12 concentrations (0-20 range) but struggles with foods containing very high B12 levels. The MSE of 5.03 suggests moderate prediction accuracy.

The QQ-plot shows an S-shaped pattern, indicating that the model's predictions deviate from perfect normality, especially at very high and very low B12 concentrations. This suggests that the simple linear relationship assumption might not capture the full complexity of B12 content in foods.This straightforward approach, while not perfect, helps us understand which nutrients tend to appear alongside B12 in foods and provides a reasonable starting point for predicting B12 content in various food items.







Protein

Residual Analysis:

The QQ-plot of residuals exhibits an S-shaped pattern, deviating from the theoretical normal distribution. This indicates non-normality in the residuals, particularly at the extremes. The central region (-1 to 1) shows good alignment, suggesting better model performance for moderate protein concentrations.

Feature Importance:

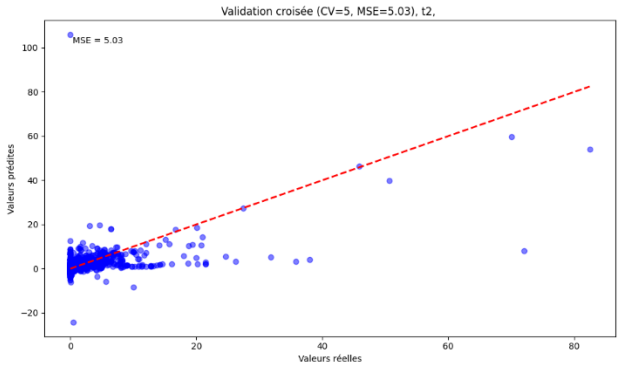
Water content demonstrates the strongest negative correlation with protein concentration, followed closely by carbohydrate content. Vitamin A (RAE) and retinol show positive correlations, while most micronutrients exhibit minimal impact on protein prediction.

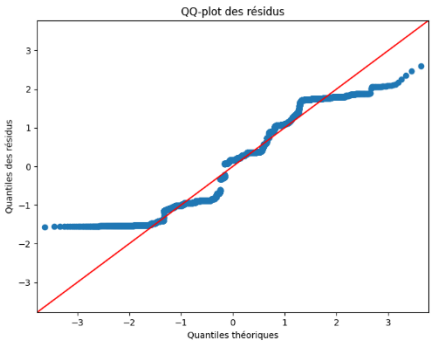
Model Performance:

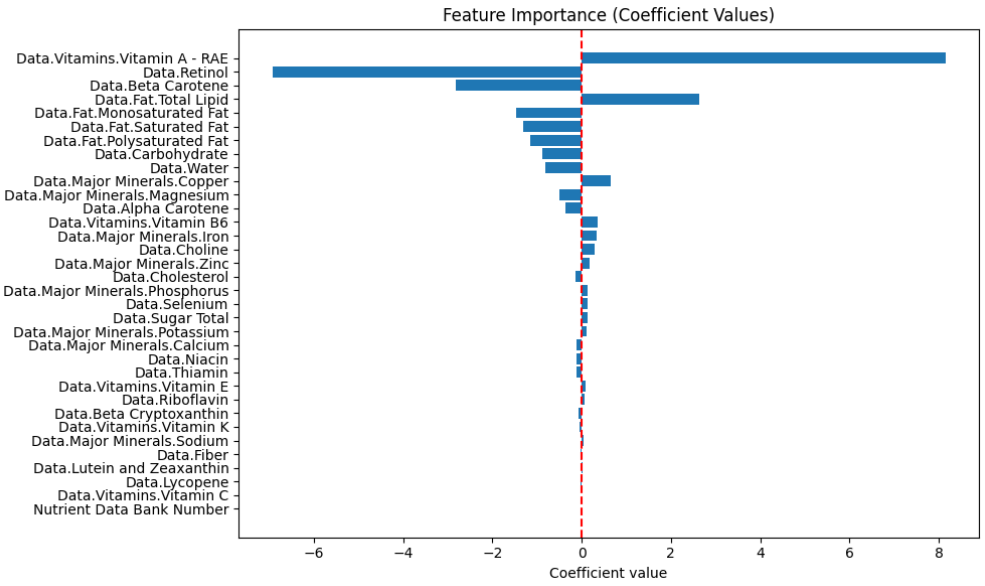
The validation scatter plot reveals an MSE of 5.76, indicating moderate predictive accuracy. The model demonstrates strong performance for protein concentrations up to approximately 40g, with increased scatter and potential underprediction for higher concentrations (>60g).

Conclusion:

The linear regression model captures general trends in protein concentration across food items, with water and carbohydrate content serving as key negative predictors. However, the model's performance degrades for foods with extreme protein values, suggesting limitations in the linear approach for comprehensive protein content prediction across all food types.







Ridge

Ridge regression, also known as Tikhonov regularization, represents a sophisticated extension of ordinary least squares regression designed to address multicollinearity and overfitting in predictive modeling. This method incorporates an L2 regularization term into the standard linear regression objective function, effectively penalizing large coefficient values while maintaining all features in the model.

The algorithm minimizes a modified loss function that combines the traditional sum of squared residuals with a penalty term proportional to the square of the magnitude of coefficients. This regularization approach shrinks the coefficients towards zero but, unlike Lasso regression, rarely sets them exactly to zero. The degree of shrinkage is controlled by a tuning parameter α (alpha), where larger values impose stronger regularization.

Ridge regression proves particularly valuable when dealing with datasets exhibiting high multicollinearity, as it stabilizes the coefficient estimates by reducing their variance, albeit at the cost of introducing some bias. This variance-bias tradeoff often results in improved predictive accuracy compared to ordinary least squares regression, especially when the relationship between predictors and the response variable is complex.

From a geometric perspective, Ridge regression can be understood as constraining the coefficient values to lie within a hypersphere, contrasting with Lasso's hyperrhomboid constraint. This geometric property explains why Ridge regression tends to share the impact of correlated predictors rather than selecting one over others.

The mathematical optimization problem solved by Ridge regression maintains differentiability throughout, enabling efficient solution through standard optimization techniques. This characteristic, combined with its stability and predictive performance, makes Ridge regression a fundamental tool in modern statistical learning, particularly suitable for scenarios where maintaining all predictors in the model is desirable while controlling for their individual impacts.

B12 Concentration

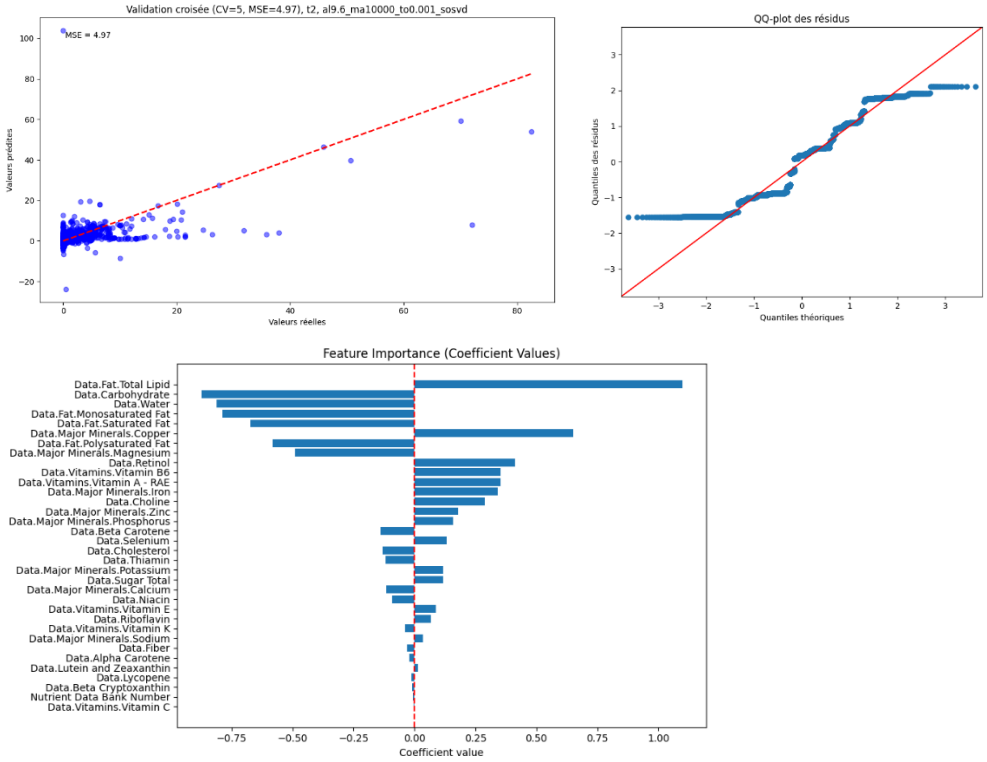
The presented visualizations demonstrate the performance characteristics of a Ridge regression model applied to B12 concentration prediction in food items. The analysis reveals several notable patterns in the model's behavior and predictive capabilities.

The QQ-plot exhibits a distinctive S-shaped pattern in the residual distribution, with pronounced horizontal plateaus at approximately -1.5 and 2.0 standard deviations from the mean. This pattern indicates systematic deviations from normality in the model's residuals, particularly at the extremes of the B12 concentration range. The central portion of the QQ-plot, spanning approximately -1 to 1 standard deviations, demonstrates superior adherence to theoretical normal distribution expectations.

The validation scatter plot, with an MSE of 4.99, illustrates the model's predictive performance across the B12 concentration spectrum. A dense clustering of predictions appears in the 0-20 µg range, suggesting robust model performance for foods with typical B12 concentrations. However, the scatter pattern reveals increasing prediction variance at higher concentrations, with notable underprediction for foods containing more than 60 µg of B12.

Feature importance analysis reveals that Total Lipid content exhibits the strongest positive correlation with B12 concentration, followed by significant contributions from copper and magnesium. The model assigns substantial negative coefficients to carbohydrate and water content, reflecting their inverse relationship with B12 concentration in foods. This pattern aligns with the biological understanding of B12 distribution in food sources, particularly its association with animal-derived products.

The Ridge regression's regularization appears to have effectively managed multicollinearity among nutritional predictors while maintaining the interpretability of coefficient values. The model's performance metrics and residual patterns suggest that while it captures the primary trends in B12 concentration, there remain challenges in accurately predicting extreme values, particularly in foods with exceptionally high B12



Protein

The presented visualizations demonstrate the performance characteristics of a Ridge regression model applied to protein concentration prediction in food items, revealing distinct patterns in predictive behavior and feature relationships.

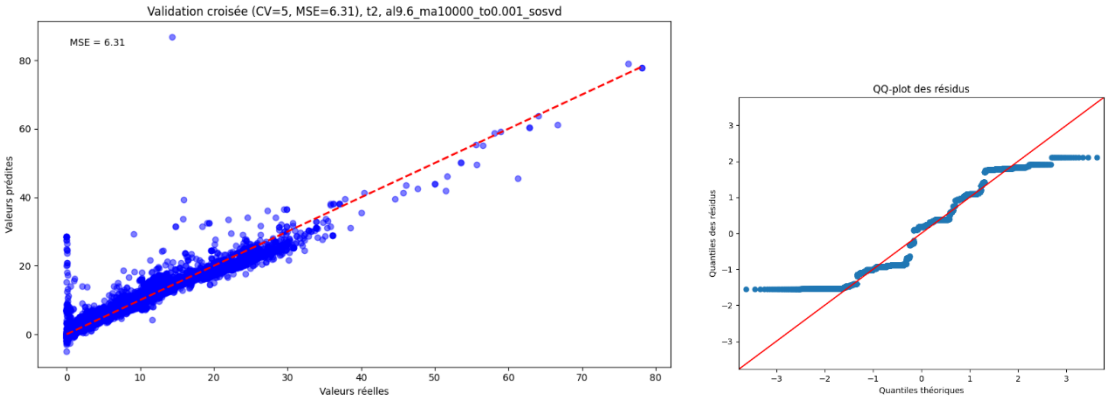
The QQ-plot of residuals exhibits a characteristic S-shaped pattern, with notable horizontal plateaus occurring at approximately -1.5 and 2.0 standard deviations. This distribution pattern indicates systematic deviations from normality, particularly at the extremes of the protein concentration spectrum. The central region, spanning approximately -1 to 1 standard deviations, demonstrates superior adherence to theoretical normal distribution expectations, suggesting more reliable predictions for moderate protein concentrations.

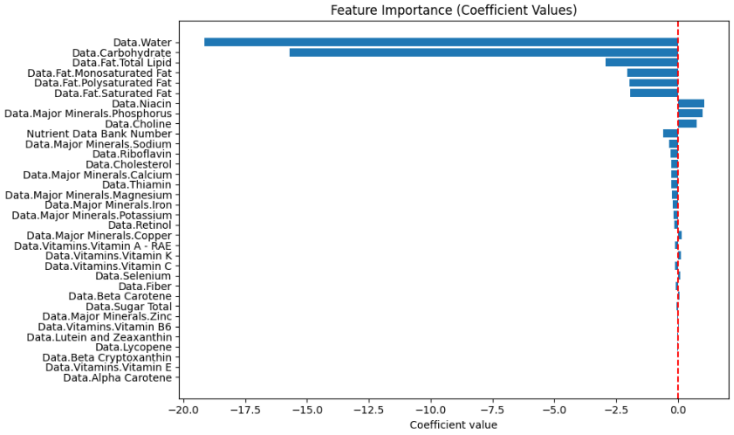
The validation scatter plot, yielding an MSE of 6.06, illustrates the model's predictive capabilities across varying protein concentrations. A dense clustering of predictions appears in the lower ranges, with increasing scatter as protein content rises. The model demonstrates robust performance for concentrations up to approximately 40g, beyond which prediction variance increases notably. The red dashed line representing perfect prediction serves as a reference, highlighting systematic underprediction for high-protein foods exceeding 60g.

Feature importance analysis reveals water content as the dominant predictor, exhibiting the strongest negative correlation with protein concentration, followed closely by carbohydrate content. Total lipids and various fat components show moderate negative correlations, while micronutrients demonstrate minimal influence on protein predictions. This pattern aligns with fundamental food

composition principles, where water and carbohydrate content typically share an inverse relationship with protein concentration.

The Ridge regression's regularization approach appears to effectively manage the multicollinearity among nutritional predictors while maintaining interpretable coefficient values. The model's performance metrics and residual patterns suggest effective capture of primary protein concentration trends, though challenges persist in accurately predicting extreme values, particularly in high-protein foods.





Conclusion

In conclusion, the analysis of powerlifting data using five machine learning classification algorithms and food data using five regression algorithms revealed significant challenges and insights.

For the powerlifting classification, the complexity arose from insufficient data and intricate subcorrelations among various types of data, necessitating deep domain knowledge for accurate interpretation. The limited dataset size likely led to overfitting issues and reduced generalizability of the models.

Regarding food data regression, the analysis highlighted the need for data stratification due to the diverse dataset structure and food types present. The heterogeneity in food composition created substantial biases, potentially masking important nutritional relationships and reducing model accuracy across different food categories.

To address these limitations, future research could explore alternative datasets for powerlifting classification, potentially incorporating more comprehensive and balanced data sources. For food regression, creating subset analyses based on food types or nutritional profiles could yield more nuanced and accurate predictions. This approach would allow for more targeted modeling, accounting for the unique characteristics of different food groups and potentially uncovering more meaningful nutritional insights. Additionally, such subset analyses could help mitigate the risk of Simpson's Paradox, a statistical phenomenon where trends observed in aggregated data may disappear or reverse when the data is divided into subgroups. By examining data at a more granular level, we can avoid misleading conclusions that might arise from hidden variables or confounding factors within broader datasets, ensuring more reliable and context-specific interpretations.

These refinements in data handling and model application could significantly enhance the predictive power and interpretability of machine learning algorithms in both powerlifting performance analysis and nutritional content prediction.