Multiclass classification of body performance based on LightGBM

Ziniu Lu

Carey Business School, Johns Hopkins University, Washington DC, 20036, United States

zlu50@jhu.edu

Abstract. The fitness industry is transforming due to the Covid-19 pandemic, and the demand for personal training is growing. The absence of a standardized and accurate body performance measuring framework can lead to misjudgement of the client’s physical condition and risks of injuries or dissatisfaction with the service. This paper applies LightGBM on a dataset containing 13393 instances from Kaggle.com for body performance level classification analysis. The model is built with 12 predictors, such as age, gender, body fat percentage, sit and bend forward distance, and a target variable class, representing the body performance level of a person. The accuracy score obtained from the confusion matrix is used as the key metric for model performance. A grid search on hyperparameters including regularization term, maximum number of leaves, and maximum depth of the tree is conducted for parameter tuning. The outcome of the model gives insight into the importance of various features to body performance and helps fitness trainers to better understand their clients’ physical conditions. Results of this research show that sit and bend forward distance and sit ups counts are the most important factors of body performance. The quality of the classifier solution makes it a useful decision aid for personal trainers to access their client’s physical condition.

Keywords: body performance, multiclass classification, LightGBM, Shapley value.

1. Introduction

Responding to the Covid-19 pandemic and controlling its spread, quarantines, curfews, and stay-at-home orders were conducted across multiple nations. As a result of fear and anxiety caused by the difficult situation, people are becoming more health-conscious [1]. The fitness industry is transforming, extending its services beyond opening physical gyms and selling memberships to offering virtual classes and fitness apps. Due to the increasing health consciousness and easy accessibility of new technologies, the target customer groups of the fitness industry are expanding to those with insufficient physical training experience. Since customers range from teenagers to seniors, and their health conditions differ, developing personalized training plans based on each individual's body performance is crucial to avoid injuries and provide better services.

To stay competitive in the fitness industry, gyms, personal training studios, and virtual apps must provide good services to raise customer satisfaction. According to the Importance-Performance Analysis conducted by Zourladani, Kesoglou, and Tsourela, professional knowledge, skills recognition, and customized programs are among the most critical factors of customer satisfaction, indicating the importance of providing good, personalized training programs [2]. A common practice of assessing a customer's body performance level involves taking a few physical tests, and the training plan is made according to the trainer's knowledge and experience. However, due to trainers' various backgrounds, common injuries are reported during training sessions [3]. A more rigorous analysis of body performance levels is needed to develop better personal training plans.

The measurement of body performance is complex, involving a variety of factors ranging from physical attributes to mental conditions, and there does not exist one standardized evaluation criteria. Gender, age, and anthropometric indices are often used to analyze physical performance, with age negatively correlated to performance levels [4]. Common power tests for athletes involve sprint running, lactate threshold power, jump height/distance, and cycle ergometer, and the reliability of such tests is investigated using statistical measurement [5]. Besides power and body build, studies from a nutritional perspective focus on the positive effects of vitamins and minerals on body performance [6]. Furthermore, non-physical factors such as psychological health and genetics are also shown to be related to body performance [7, 8]. While there is a lot of research on body performance measurements targeting older people, medical patients, and professional athletes, developing a generalized quick test applicable to regular people of all age groups is rarely studied. This paper aims to study the impact of several factors, including age, physical attributes, and power tests, on body performance level classification, utilizing machine learning approaches. Factors to be studied are chosen from those which are significant from previous studies, and which can be conveniently gathered by personal trainers.

Due to the complexity of classifying body performance and the large number of factors that could influence the output, using machine learning methods is an appropriate approach. While simple linear regression provides great interpretability, it might not reach high accuracy due to its inability to capture non-linear relationships between factors and the target variable. Pan and Feng researched building a performance prediction model for sports athletes, and the result showed that the support vector machines, and the particle swarm optimization algorithm outperformed the simple linear regression model [9]. However, advanced machine learning algorithms often need more transparency to make it easier to figure out how each factor contributes to the outcome. Explainable artificial intelligence (XAI), such as Shapley Additive explanation (SHAP), global attribution mappings (GAMs), and local interpretable model-agnostic explanations (LIME), are developed to provide the expansibility of black box models. This paper adopts some XAI methods to help trainers better understand their clients' body performance.

The objective of this paper is to examine the relationship between various physical factors and the body performance level and build a multi-class classification model to help personal trainers better understand their customers' physical abilities. The LightGBM model was picked from a horse race of 8 classifiers based on the key metric of the confusion matrix. A grid search was conducted for hyperparameter tuning to further improve the LightGBM model performance on the validation dataset comparing the chosen benchmark. Due to the low-interpretable nature of the LightGBM model, this paper used Shapley values to illustrate the contribution of each feature to the model’s outcome. The quality of the classifier solution makes it useful as a decision aid for personal trainers to build appropriate training plans for clients.

2. Methodology

2.1. Data source

This paper uses the body performance dataset from Kaggle.com. The dataset contains 13393 instances and 12 parameters including age, gender, height, weight, and some test scores. The dataset is already pre-processed and filtered using raw data from Korea Sports Promotion Foundation. There is no missing value in the dataset, and duplicates are dropped.

2.2. Variable description

Table 1 shows the variable name and the value range of the 12 variables from the dataset.

Table 1. Name, symbol and value range of variables.

Name	Symbol	Range
age	AGE	[20,64]
gender	GEN	{F, M}
height_cm	H	[125,193.8]
weight_kg	W	[26.3,138.1]
body fat_%	BF	[3,78.4]
diastolic (blood pressure)	DBP	[6,156.2]
systolic (blood pressure)	SBP	[14,201]
gripForce	GF	[0,70.5]
sit and bend forward_cm	SBF	[-25,213]
sit ups counts	SU	[0,80]
broad jump_cm	BJ	[0,303]
class	CLASS	{A, B, C, D}

The class variable is the target variable this paper tries to classify using the other variables. Class A indicates the best body performance, and class D indicates the worst body performance. All four classes are balanced.

2.3. Tukey's method for outlier detection

An outlier is a data point that has extreme values and is significantly different from other observations. It can reduce the performance of machine learning algorithms due to its disproportionate influence. Tukey's method uses quartiles to detect outliers in the dataset.

$$Lower\ Bound = Q1 - (1.5 * IQR) \tag{1}$$

$$Upper\ Bound = Q3 + (1.5 * IQR) \tag{2}$$

Where Q1 is the 1st quartile, Q3 is the 3rd quartile. The interquartile range (IQR) can be calculated by Q3 – Q1. Any observations with data not within the lower and upper bound will be treated as outliers and removed.

2.4. Light gradient-boosting machine (LightGBM)

LightGBM is a gradient boosting framework with tree-based learning algorithms. It combines the Gradient Boosting Decision Tree (GBDT) with techniques of Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bunding (EFB) to increase the speed of the algorithm [10]. GBDT is a boosted ensemble with base learner decision trees which iteratively focuses attention on the records with the greatest errors by fitting the decision tree on the residuals error of the previous one [11].

2.5. Hyperparameter tuning for LightGBM

The regularization parameter (*lambda_l1*), maximum number of leaves (*num_leaves*) and maximum depth of the tree (*max_depth*) are some of the most important hyperparameters of the LightGBM model. A higher penalty value can significantly reduce the overfitting issue of the model but also increase the risk of underfitting the data. The maximum number of leaves and depth of each decision tree controls the overall complexity of the model and is associated with the training time of the model and the overfitting issue.

A grid-search with 5-fold cross-validation using accuracy for evaluation is applied for hyperparameter tuning. Table 2 shows the value ranges and step sizes of hyperparameters used in the grid search. The range of lambda_l1 is between 0 and 50 with step size of 1. The range of num_leaves is between 10 to 1000 with step size of 10. The range of max_depth is between 4 to 10 with step size of 2.

Table 2. Hyperparameters grid search.

Distance (m)	Value Range	Step
lambda_l1	[0,50]	1
num_leaves	[20,800]	20
max_depth	[4,10]	1

2.6. *Shapley value*

Shapley Value comes from cooperative game theory, which measures each player's contribution to the payout of a coalitional game [12]. It can be used as an Explainable Artificial Intelligence tool to help understand the importance of each feature by viewing the model as the game, the model's output as the payout of the game, and the model's variables as players.

$$\phi_i(v) = \frac{1}{|N|}\sum_{S\subseteq N\backslash\{i\}} \binom{|N|-1}{|S|}^{-1} (v(S\cup\{i\}) - v(S)) \tag{3}$$

N is the set of all features from the machine learning model, and S is a subset of N. v is a function which gives the model output for any subset of features.

3. Results and discussion

3.1. *Descriptive analysis*

Based on Figure 1, the target variable of the multi-class classification model, class, is well balanced. There are around 3500 instances for each of the four classes. A indicates the best body performance level, and D indicates the worst body performance level.

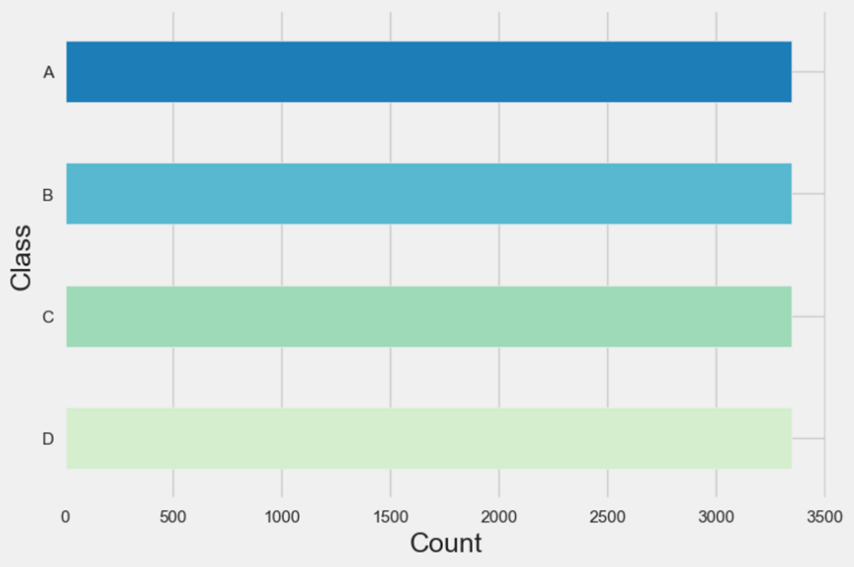

Figure 1. Bar plot of target variable class.

The average body fat percentage among all instances is 23.24. The distribution of body fat percentage varies among four classes, indicating that it is an important factor of a person's body performance level. As shown in Figure 2, the body fat percentage distribution of class D has a higher mean compared to that distribution of class A and class B. Also, observations with the largest body fat percentage (>45%) are all classified as class D.

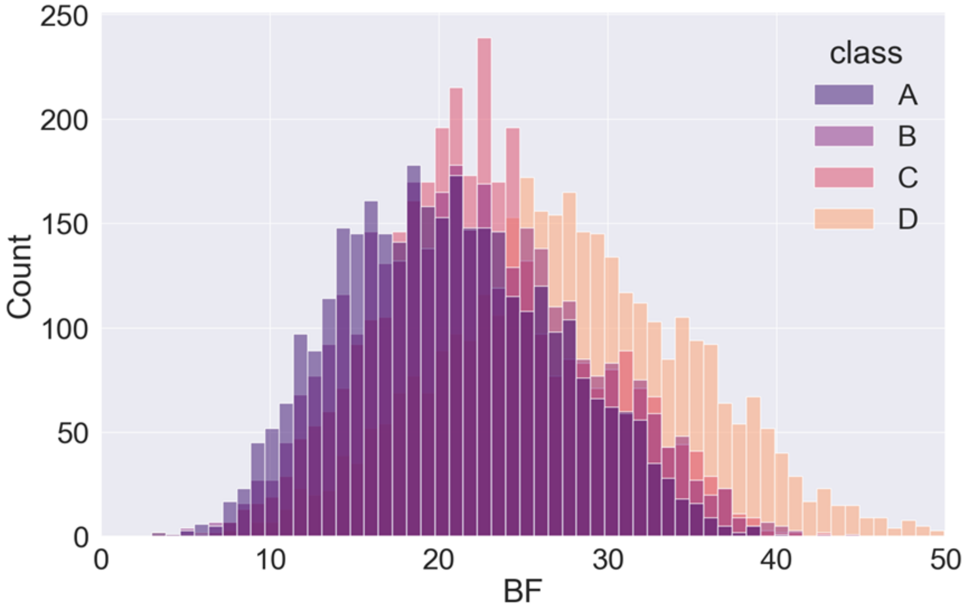

Figure 2. Histogram of body fat percentage by class.

Similar to the body fat percentage, a significant difference can be observed for the distribution of sit and bend forward distance among different classes. Class A has the highest mean sit and bend forward distance (21.39), which is approximately three times the mean sits and bend forward distance for Class D (7.59). As shown in Figure 3, the mean sits and bend forward distance is higher for instances with better body performance levels. Also, only instances with class D have negative distances.

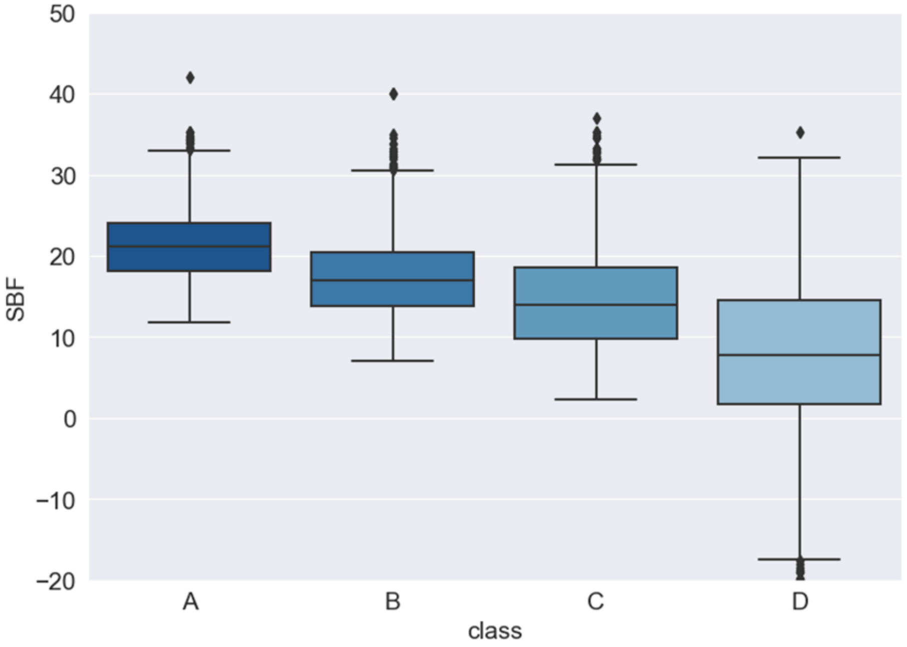

Figure 3. Histogram of SBF by class.

The physical test scores, including sit up counts and grip force scores, are designed to measure a person's physical strength which is closely related to overall body performance level. As shown in Figure 4, instances of class A tend to have higher GF and SU values. Instances of class D tend to have lower GF and SU values. It indicates that people with better body performance are more likely to have better grip force and sit up counting scores.

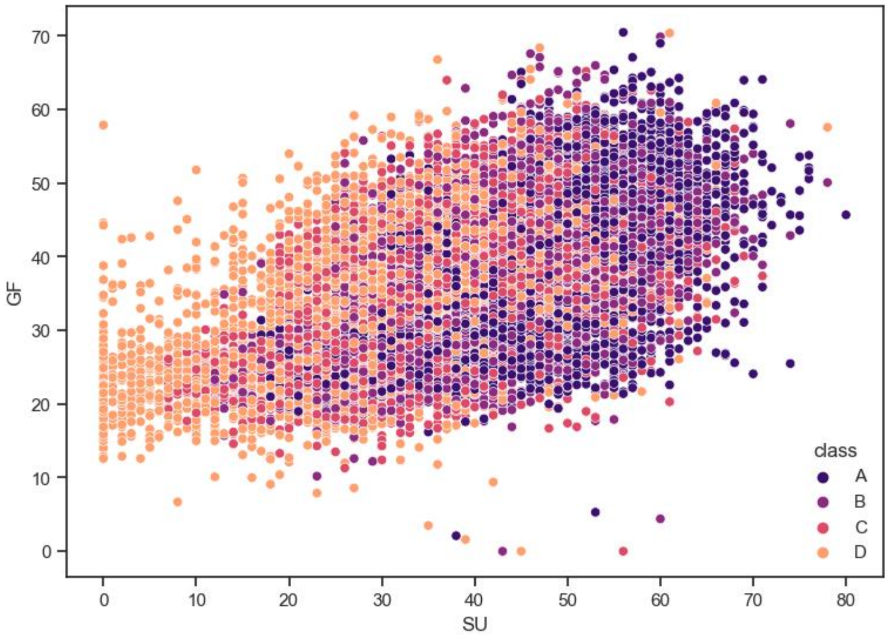

Figure 4. Scatterplot of GF and SU.

3.2. Inferential analysis

The dataset is divided into 80% for model training and 20% for validation based on random resampling. The training dataset contains 10712 instances, and the validation dataset contains 2679 instances. All four target body performance classes are balanced in both datasets after splitting. The standardization is first applied to the training dataset. The standardized training dataset has a mean of -3.84 * 10-16 and a standard deviation of 1.00, and these values are used to standardize the validation dataset.

The key performance metric used is the accuracy rate which is commonly used for measuring performance of multi-class classification model. It shows the fraction of the data that is correctly predicted by the model. A higher accuracy indicates a higher probability that the model gives the correct prediction on a random observation. However, accuracy only measures the general performance of the model, and the confusion matrix is needed to identify the model performance on each individual class.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

By using the 5-fold cross validation grid search on hyperparameters, a LightGBM model with *lambda_l1* = 0, *max_depth* = 10, and *num_leaves* = 40 gives the highest accuracy. By fitting the LightGBM model with the obtained hyperparameter values on the training dataset, the model gives a 0.745 accuracy on the validation dataset. The confusion matrix is also obtained.

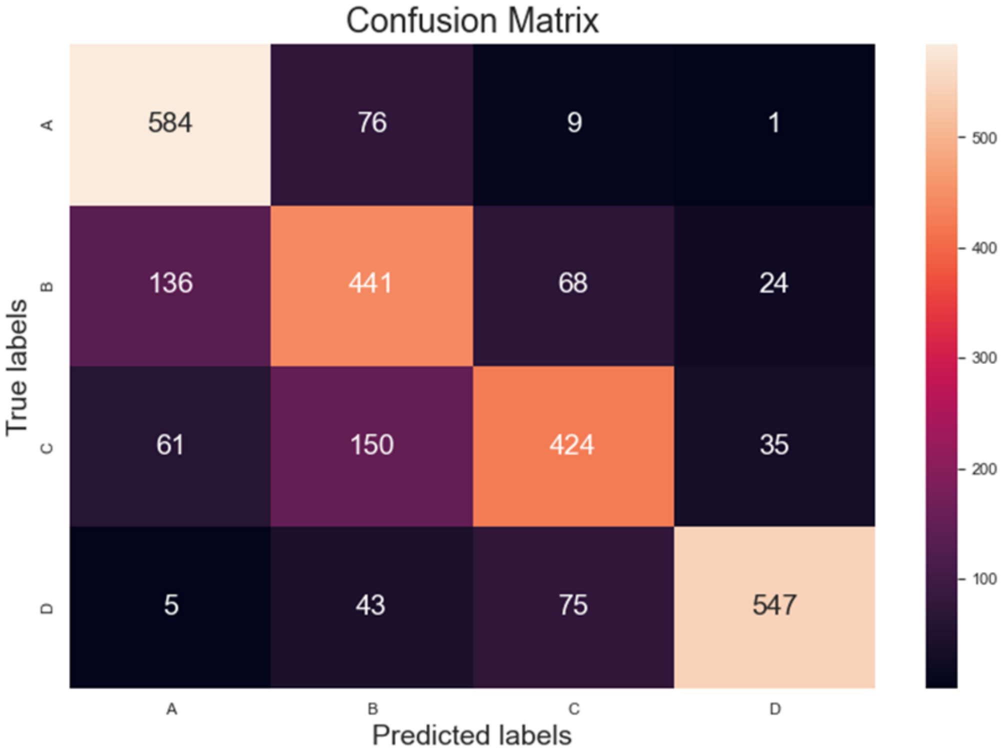

Figure 5. Confusion matrix of model outcomes.

Figure 5 shows that the model gives more accurate predictions for class A and class D comparing to class B and C. One possible explanation for the predicting accuracy difference is that since class A and class D are the two extreme classes representing people with the best and worst body performances, they are easier to distinguish and identify.

A benchmarked decision tree model with max_depth = 10 is used to compare with the LightGBM model. Using the same training and validation dataset, the benchmark model has a 0.585 accuracy on the validation dataset. The LightBGM model significantly improves the predicting accuracy by 0.16.

A learning curve plots the model's performance on the training and validation datasets over experience and is used to detect underfitting/overfitting issues. Underfitting occurs when the model cannot learn from the dataset and thus generates large loss and low predicting power. On the contrary, overfitting occurs when the model is fitting too well on the training dataset and fails to be generalized to other data. The learning curve of the trained LightBGM model is shown below. Figure 6 shows no underfitting/overfitting issues, since the accuracy for both training and validation are relatively high, and both curves are converging over experience.

This paper uses the Shapley Value to access the feature importance of the LightBGM model. A large absolute value of Shapley Value means that the feature is contributing more to the model outcomes and is thus more important.

Figure 7 shows that sit and bend forward distance, sit-ups count, and age have the largest mean absolute Shapley values and thus contribute the most to the model's outcome. Diastolic and systolic blood pressures have the lowest Shapley values and contribute the least to the model's outcome. This paper also plots Shapley values for specific classes separately.

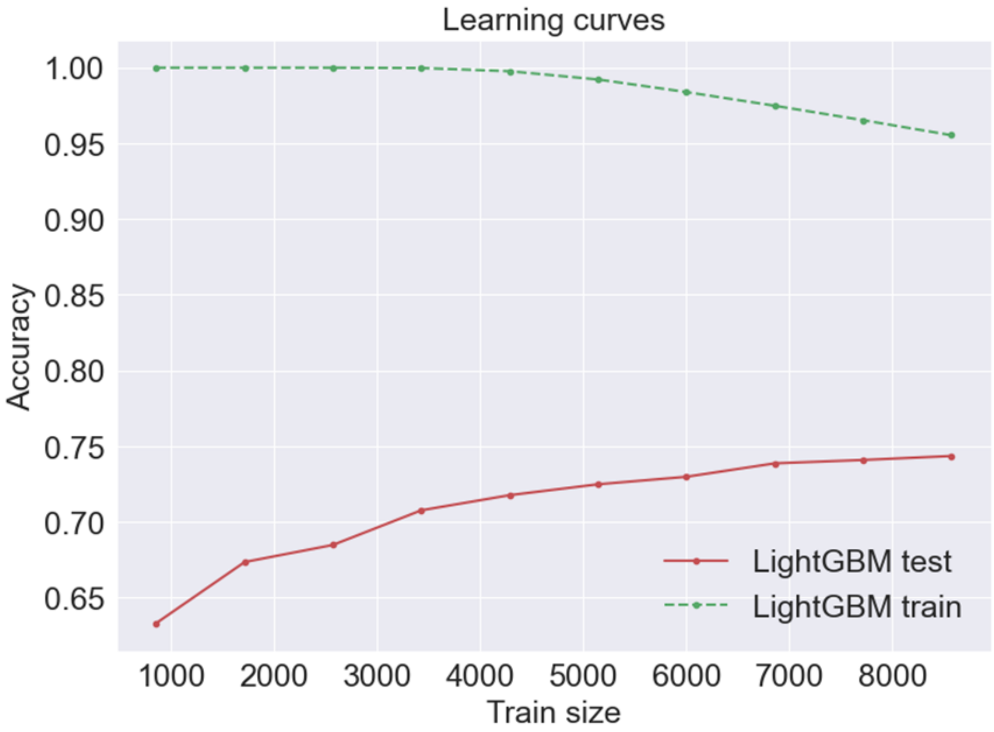

Figure 6. Learning curve with accuracy.

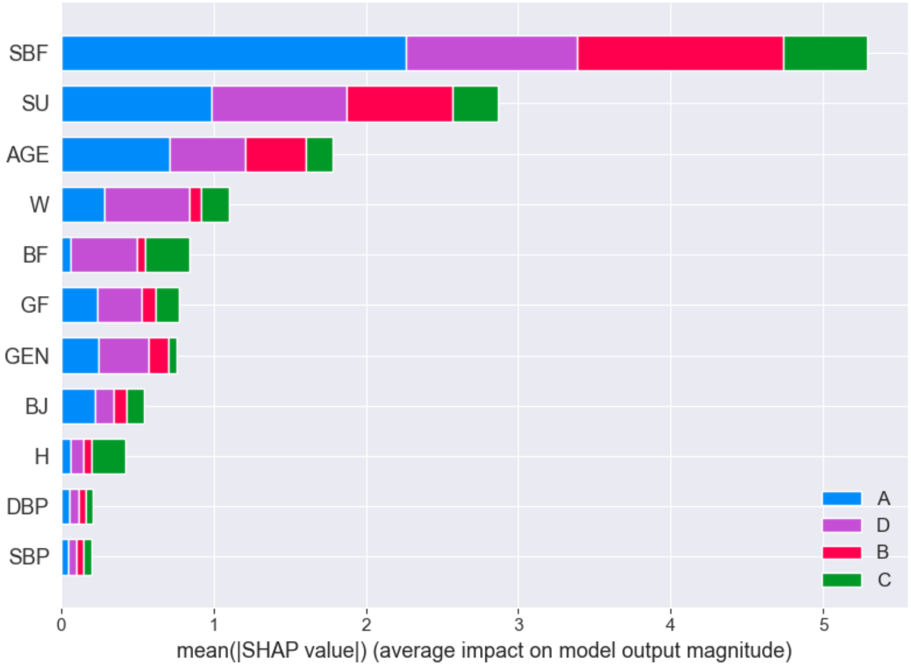

Figure 7. SHAP values of each feature.

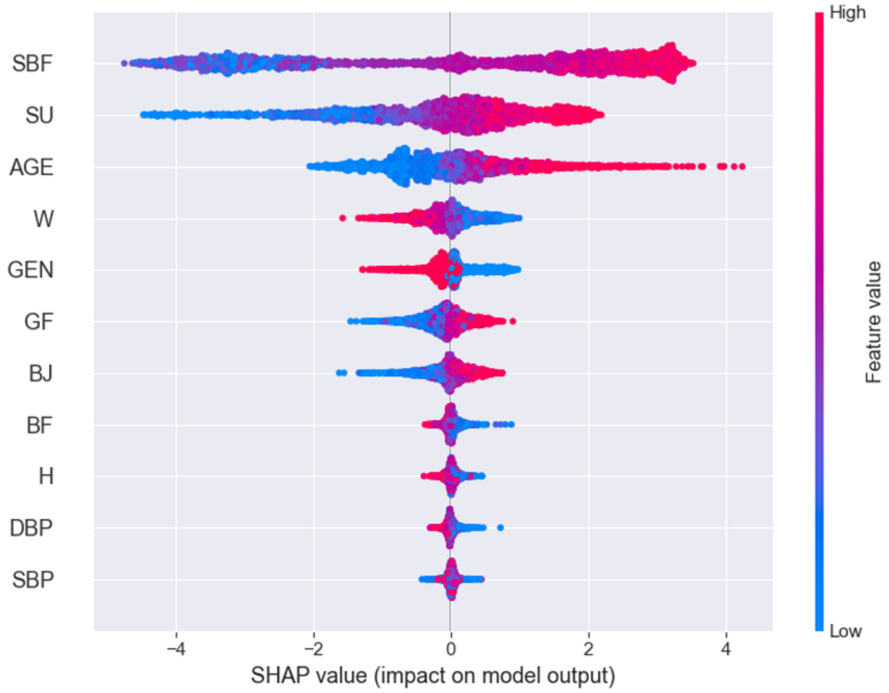

Figure 8. SHAP values for Class A.

Figure 8 shows that sit and bend forward distance, sit-up count, and age are the most important features and are positively correlated with the probability of being classified as Class A. A person with high values in these three features is more likely to have good body performance. On the contrary, a male with high body weight is less likely to have good body performance.

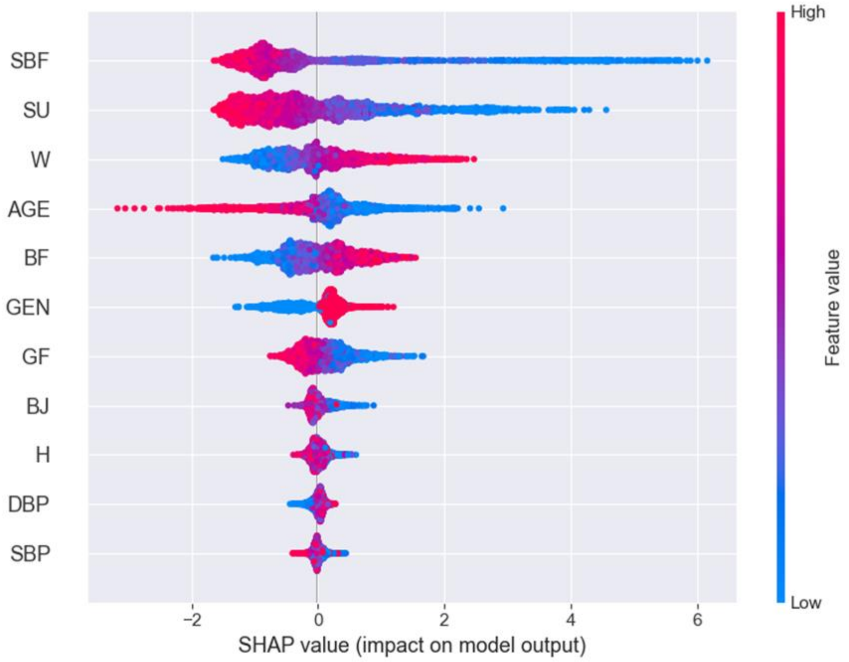

Figure 9. SHAP values for Class D.

Figure 9 shows that sit and bend forward distance, sit-up count, and weight are the most important features for predicting Class D. Low Sit and bend forward distance and low sit-up count increases the probability of a person being classified as Class D, which means bad body performance level. A person with large weight is more likely to have low body performance level as well.

4. Conclusion

This paper aims to help businesses in the fitness industry to better understand their customers’ physical conditions by studying the relationship between various physical features, physical test scores, and body performance levels. Using data from Kaggle, a multiclass classification analysis was conducted with body performance level as the target variable and key metrics obtained from the confusion matrix. Descriptive analysis and a horserace of 8 classifiers were performed, and the LightGBM model was selected due to its high accuracy and fast speed. A grid search for hyperparameters was performed to improve the accuracy of the LightGBM model to 0.745, making it a useful aid for personal trainers to build training plans. The Shapley values of variables revealed that sit and bend forward distance and sit up counts are positively correlated with body performance and are the most important features in predicting body performance levels.

References

[1] Aksoy N C, Kabadayi E T and Alan A K 2021 An unintended consequence of covid-19: healthy nutrition. *Appetite*, 166.

[2] Athanasia Z, Vasiliki K and Maria T 2020 An importance-performance analysis of personal training studios and gyms service quality. *International Journal of Progressive Sciences and Technologies*, 22(1), 401-11.

[3] Waryasz G R, Daniels A H, Gil J A, Suric V and Eberson C P 2016 Personal trainer demographics, current practice trends and common trainee injuries. *Orthop Rev (Pavia)*, 8(3).

[4] Samson M M, Meeuwsen I B, Crowe A, Dessens J A, Duursma S A and Verhaar H J 2000 Relationships between physical performance measures, age, height and bodyweight in healthy adults. *Age and Ageing*, 29(3).

[5] Hopkins W G, Schabort E J and Hawley J A 2001 Reliability of power in physical performance tests. *Sports Med*, 31, 211-34.

[6] Lukaski H C 2004 Vitamin and mineral status: effects on physical performance. *Nutrition*, 20(7-8), 632-44.

[7] Plante T G and Rodin J 1990 Physical fitness and enhanced psychological health. *Current Psychology*, 9, 3-24.

[8] Thompson W R and Binder-Macleod S A 2006 Association of genetic factors with selected measures of physical performance. *Physical Therapy*, 86(4), 585-91.

[9] Zhu P and Sun F 2019 Sports athletes' performance prediction model based on machine learning algorithm. *Advances in Intelligent Systems and Computing*, 1017.

[10] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu T Y 2017 LightGBM: a highly efficient gradient boosting decision tree. *31st International Conference on Neural Information Processing System*.

[11] Friedman J H 2001 Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-232.

[12] Shapley L 1953 A value for n-person games. *Contributions to the Theory of Games II*, 307-17.