Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 15, No. 3, 2025

A Study of a Prediction Framework Incorporating Random Forests and Gradient Boosting Trees

Yining Zhang[1, *], Yimeng Wang[2]

[1]Guangdong University of Foreign Studies, Guangzhou, China
[2]Macao Polytechnic University, Macao, China
[*]Corresponding author: 1455862781@qq.com

Abstract: This study constructs a hybrid prediction framework integrating Random Forest regression and gradient boosting tree. Firstly, the study uses Random Forest regression to mine the nonlinear associations between multidimensional features through self-sampling and feature random selection mechanisms to achieve the classification and regression tasks, and builds in feature importance analysis to assess feature contributions. Second, logistic regression and linear regression are used to provide statistical explanations for binary classification problems and multivariate interactions, respectively, to enhance the interpretability of the model. In addition, this study constructs a gradient boosting tree model and combines it with SHAP value analysis, iterative fitting of residuals to improve prediction accuracy, and hyperparameter tuning to further optimise model performance. This framework enhances the analysis and prediction ability of multivariate system data through multi-model collaboration, demonstrates good stability and generalisation ability, and provides a reusable technical paradigm for similar studies.

Keywords: Random Forest Model, Regression Model, Gradient Boosting Trees Model, SHAP Model.

1. Introduction

In multivariate system analysis, it is difficult for a single algorithm to take into account the nonlinear modelling and interpretability needs of complex features. With the continuous increase of data dimensions, how to efficiently mine the potential associations among features and improve the generalisation ability of prediction models becomes a research challenge. Focusing on the modelling challenges of multivariate systems, this study proposes a hybrid prediction framework integrating Random Forest regression and gradient boosting tree, aiming to break through the limitations of traditional methods through the mechanism of multi-model synergy [1]. Firstly, the study systematically captures the nonlinear associations among multidimensional features with the help of the self-sampling and feature random selection mechanism of the Random Forest model, so as to realise the classification and regression tasks of high-dimensional data. Secondly, the logistic regression model is used to complete the binary prediction task, and the linear regression model combined with One-Hot coding is used to analyse the effects of multivariate interactions, quantify the contribution of each feature to the results, and present the effects of the additional features and the ranking of regression coefficients through visualization [2]. In addition, the gradient boosting tree model optimises the prediction accuracy by iterative fitting of residuals, and quantifies the importance of features by combining with SHAP value analysis to further improve prediction accuracy and transparency. This framework provides a solution for multivariate system prediction and decision-making with both accuracy and interpretability [3].

2. Medal Count Predictions by Country

2.1. Medal distribution prediction

(1) Feature selection

The prediction is mainly based on the following aspects, including the trend of historical medal counts, the host country effect, and the number of events, so as to provide high-quality input variables for the model and improve the accuracy of prediction.

a) Historical performance

Analyze the performance trend of each country in the past few Olympic Games, such as the year-on-year change in the number of medals. Calculate the average number of medals and the total number of medals for each country over the last few years.

b) Host country effect

Host countries often have more resources to invest in sports competitions and usually have higher medal counts.

c) Number of events

The number of medals is closely related to the number of events participated in. Extract the number and type of events participated by each country as an important input variable.

(2) Construction and training of the Random Forest model

In the medal table prediction, there are multiple features (historical results, host country effect, number of events, etc.), but not every feature has a significant impact on the number of medals. The model has built-in feature importance analysis to evaluate the contribution of each feature by splitting the gain of the node.

The host country effect can be modeled by adding host country features (such as Host_2028) and geographic proximity features. It can flexibly identify the strength of the host country effect and its interaction with historical results.

Assume that an observation value X_i is not zero and is contained in the leaf node l (x, θ), and its weight ω_i(x, θ) can be expressed as:

$$\omega_i(x,\theta) = \frac{l\{X_i \in R_l(x,\theta)\}}{l\{j : X_j \in R_l(x,\theta)\}}, (i = 1, 2, \ldots, n) \quad (1)$$

The weight $\omega_i(x)$ of each observation $Y_i \in (1, 2, ..., n)$ can be represented by taking the mean of the weights $\omega_i(x, \theta_t)$ (t=1, 2, ..., k) of the decision tree:

$$w_i(x) = \frac{1}{k}\sum_{i=1}^{k} w_i(x,\theta)Y \tag{2}$$

In cases involving large amounts of data such as multiple Olympic Games, the Random Forest model divides the data into training and test sets, and evaluates the model performance by calculating the mean square error (MSE) and R^2. The training process can be completed quickly, saving time and improving efficiency. Based on the gradient boosting algorithm, this paper focused on samples with large prediction errors to improve the prediction accuracy of countries with fewer medals.

(3) The Solution of model

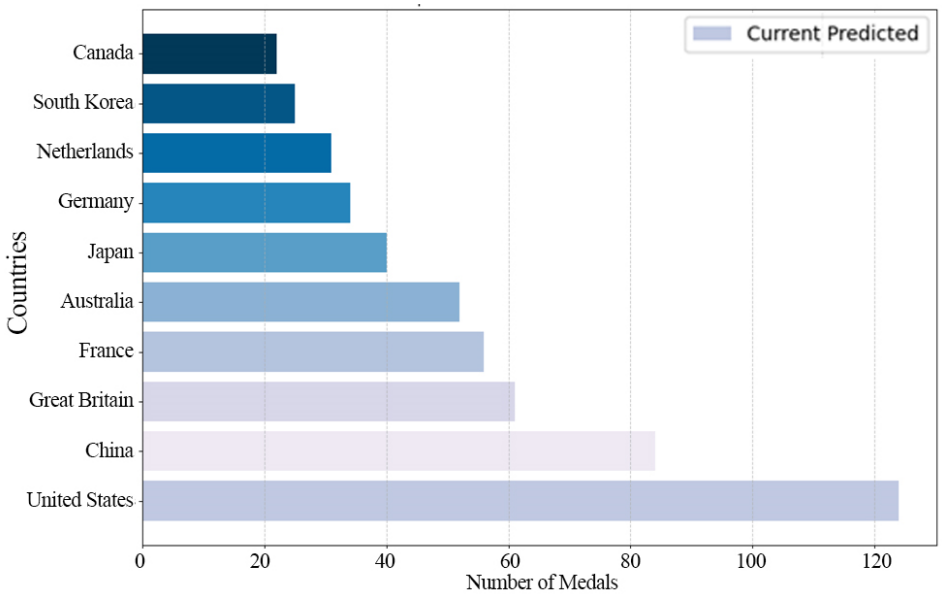

Figure 1. Prediction of the top ten countries in Los Angeles Olympic

Figure 1 shows that the USA and China continue to lead in the number of gold medals and the total number of medals, followed by Great Britian and France. There is little fluctuation in the number of medals won by the strongest countries (standard deviation <5 per cent), and some small and medium-sized countries may slip in the rankings as a result of the downsizing of their participation.

2.2. First-Time winners' prediction

(1) Feature selection

The performance of a logistic regression model depends on the quality and relevance of the input features. Therefore, the goal of feature selection is to pick out variables that are highly correlated with the probability of winning a gold medal, while avoiding redundant or irrelevant variables to improve the performance and interpretability of the model [4].

This paper selected the following three features to build the model:

a) Num_Events

The more events a country participates in, the more sports it competes in and the more likely it is to win a medal. This feature directly reflects a country's participation in the Olympics and is an important feature directly related to winning a medal.

b) Num_Athletes

The more athletes a country participates in, the more it invests in the Olympics and the more likely it is to win a medal.

c) Nearby_Host

This is a binary variable (0 or 1). Countries that are close to the host country tend to perform better in the Olympics. Possible reasons include: Geographical advantage (short travel time). The influence of the host country on neighboring countries. Cultural or economic ties.

This feature reflects the potential impact of geographic and social factors on Olympic performance.

(2) Construction and training of the logistic regression model:

Step 1: Train the logistic regression model.

Use Num_Athletes, Num_Events, Nearby_Host as features (X) and Gold as the target variable (y) for training.

The data is split into training and test sets by train_test_split.

Step 2: Smooth the predicted probabilities.

The probability values output by the prediction model may be close to 0 or 1, resulting in an insufficiently refined model. Use the smoothed_probability function to adjust these extreme values to a more appropriate range.

Step 3: Select the best threshold using Youden's J statistic.

The best threshold is selected by calculating fpr (false positive rate) and tpr (true positive rate), as well as their difference. This threshold is used to determine whether a country will win the award.

Using this threshold, classify the test set data and output a classification report.

2.3. The Solution of model

From the results, countries with higher economic level and rich sports resources are more likely to achieve gold medal victories in uncharted sports, reflecting the positive cycle of 'resource accumulation - programme expansion'. On the other hand, countries with weak infrastructures and a single sport are less likely to win gold medals for the first time, and need to improve their competitiveness by expanding the scale of participation, optimising the layout of the sport or international cooperation. Combined with the feature selection in the previous section (number of participants,

number of events, geographic proximity), the high probability countries usually have the characteristics of ‘many events and large athletes’, which verifies the reasonableness of the input features of the model.

3. The Relationship Between Events and Medals

3.1. Model training

Linear regression is used to analyze project settings, the importance of each sport, and the strategic impact of the host country. It has the significant advantages of intuitive results, strong interpretability, and support for decision optimization:

a) Quantified impact

Linear regression can quantify the specific contribution of each sport and event to the number of medals, analyze the regression coefficient of each variable, and clearly understand which events have the greatest impact on the number of medals. By comparing the medal contributions of different countries in various events, identify which sports have a strategic position in each country. Analyze the direct impact of the host country's choice of new events or focused events on the overall medal distribution.

b) Support multivariate analysis

The model can simultaneously consider the interaction between multiple variables such as sports, events, and host country identity, so as to comprehensively analyze how these factors jointly affect the medal distribution.

c) One-Hot encoding of sports and events, extracting regression coefficients and visualization.

3.2. Visualization

Data visualization presents numbers and statistical information in a graphical way, which allows people to quickly and intuitively understand complex data distribution and trends. In this question, this paper mainly generates the following two graphs:

a) Total medals of newly added events of host countries

Shows the total medals won by each host country in newly added events. Analyze the performance of host countries in newly added events and observe whether there is a trend that host countries increase their total medals through newly added events.

b) Regression coefficient of linear regression model

Process sports and events through One-Hot encoding, fit linear regression model, extract and sort regression coefficients, and then display them in bar charts. Quantify the impact of different events and events on the number of medals, analyze which events are most important for medal distribution, and the impact of event settings on the results.

Through visualization, data analysis has changed from a simple "number accumulation" to a clearer, more vivid and more explanatory one, which can not only reveal hidden information, but also help decision makers make effective strategic adjustments quickly.

3.3. The Solution of Model

As shown in Figure 2, the number of medals won by the host country is generally higher than that of the non-host country, for example, the number of medals won by the United States as the host country is more than 2,000, which shows that the home field advantage has a significant effect on the new items, and at the same time, it also reflects that the host country is able to optimise the distribution of its own medals by virtue of the right to set up its own items in the tournament and make a reasonable layout.

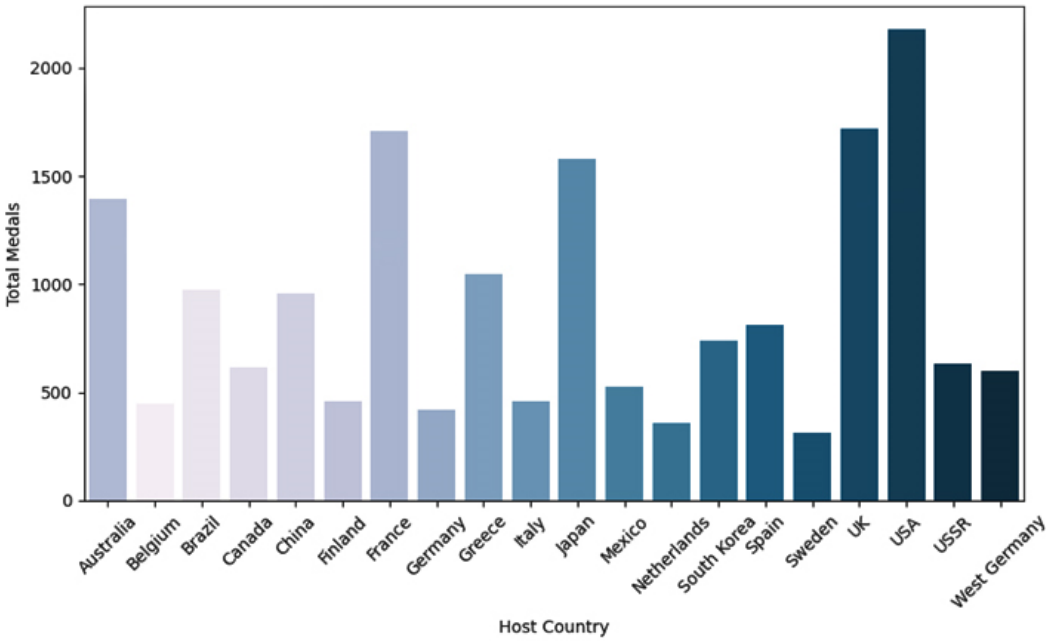

Figure 2. Total medals of newly added events of host countries

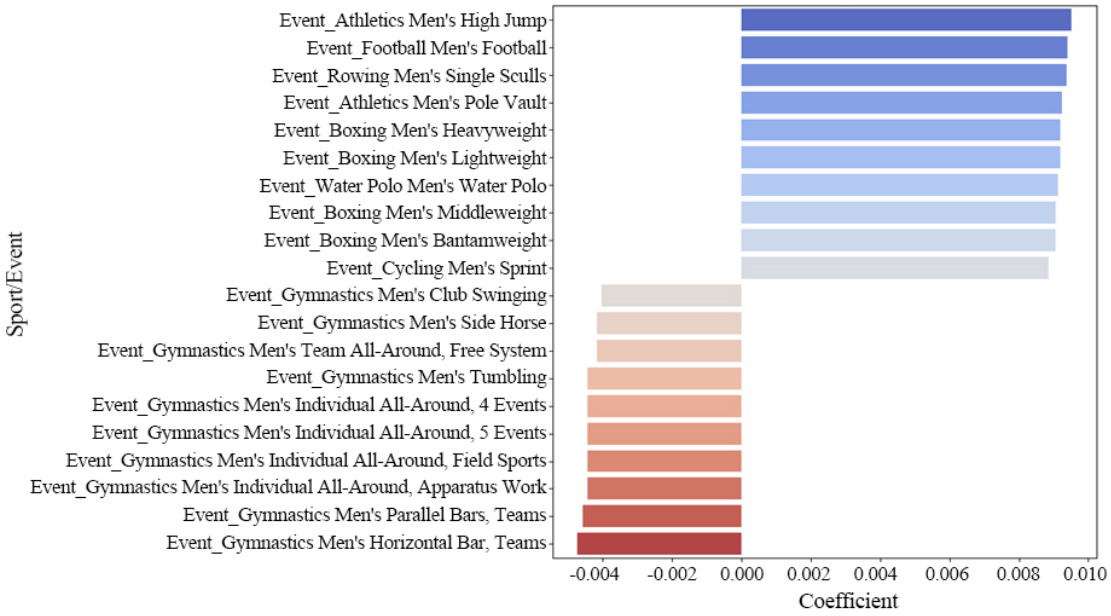

Figure 3. Top 10 and bottom 10 regression coefficients

Figure 3 shows that the top 10 high contributing sports are mostly team sports and sports with high popularity, such as football, basketball, etc., and their regression coefficients are concentrated in the range of 0.008 - 0.01; while the bottom 10 low contributing sports are mostly more technical or niche sports, such as Artistic Gymnastics, Modern Pentathlon, etc., whose regression coefficients are negative. This visualisation shows the strategic value of different events in terms of medal winning, and provides a valuable reference for decision makers in resource allocation, such as prioritizing investment in high coefficient events when resources are limited.

4. "Great Coach" Analysis

4.1. SHAP model description

SHAP is a method for interpreting machine learning model predictions based on Shapley values from game theory. It evaluates each feature's contribution to the model's output, improving model interpretability and transparency. SHAP helps explain both the overall behavior of the model (global interpretation) and the results of individual predictions (local interpretation). By enhancing model trust, detecting biases, and improving clarity, SHAP is particularly useful with black-box models. The core of SHAP is the Shapley value, which is calculated as:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (3)$$

SHAP works in the following steps:

a) Model Training: Train the machine learning model.

b) Shapley Value Calculation: For each sample, calculate the Shapley value of each feature to measure its contribution to the prediction.

c) Local Interpretation: For a single prediction, SHAP explains how each feature affects the outcome.

d) Global Interpretation: Aggregate Shapley values across all samples to identify the most important features.

e) Visualization: Use visual tools like bar charts, dependency charts, and force diagrams to show the impact of each feature on the model's output.

4.2. The establishment of Gradient Boosting Tree

(1) Feature engineering construction

a) Country Codes: Converts country codes (NOCs) into numeric data. Using the astype ('category').cat.codes method, a unique integer code is assigned to each country, helping the machine learning model process these categorical data.

b) Coach Information: Each athlete is randomly assigned a coach (assuming there are four coaches), and each coach has randomly generated years of experience and performance ratings. These features help capture the potential impact of coaches on athletes' performance. Specifically:

c) Coach Assignment: The np.random.choice function is used to randomly assign a coach to each athlete, simulating the impact of different coaches on athlete performance.

d) Coaching Experience: Each coach is assigned a random number of years of experience ranging from 1 to 20 years.

e) Coaching Performance: Each athlete is randomly assigned a coaching performance score between 0 and 1, indicating the coach's performance (e.g., based on the performance of the team they lead).

(2) Interpretable modeling based on Gradient Boosted Trees

Gradient Boosting Regression is a regression model based on the Gradient Boosting Trees (GBT) method, which uses stepwise addition for modeling. Its core idea is to correct the pre-diction error of the previous round of the model through each round of training, and gradually improve the prediction results[5].

In each round of training, the model fits the residuals (i.e., the gap between the true value and the predicted value) of the previous round of modeling: The initial model uses simple estimates to make predictions.

In round m, a new regression tree is trained to predict these errors based on the residuals from the previous round, and the model is adjusted.

$$F_m(x) = F_{m-1}(x) + \alpha \cdot h_m(x) \quad (4)$$

The specific solution steps are as follows:

a) Initialize the model: set up the initial model, which is usually a simple constant prediction (e.g., mean value).

b) Calculate residuals: Calculate the difference between the true value and the current model prediction, and get the residuals.

c) Fit residuals: train a decision tree to fit these residuals to get.

d) Update the model: add the weighted predictions of the new tree to the current model to get the updated predictions.

e) Iterate: Repeat the above process until the maximum number of iterations is reached or the residuals are less than a certain threshold.

In addition, in order to improve the interpretability of the model, the gradient boosting tree is combined with the SHAP model. The Gradient Boosting Tree can automatically learn complex nonlinear relationships in the data, which is suitable for dealing with high-dimensional data and complex feature interactions. SHAP helps to explain black-box models such as gradient boosting tree by calculating the contribution degree for each feature, which makes the model's decision-making process clearer and enhances the interpretability [6].

4.3. The solution of model

The results of hyperparameter optimization are shown in Table.1.

Table 1. Hyper-Parameter optimization results

Parameters	Optimal Hyperparameters for Gold Medal Count Prediction	Optimal Hyperparameters for Total Medals Prediction
n_estimators	1	0.9
learning_rate	100	100
max_depth	2	5
min_samples_split	1	2
min_samples_leaf	5	3
subsample	0.05	0.1

After hyperparameter optimization search, the MSE of gold medal count prediction is 0.0472 and MAE is 0.0956; the MSE of total medal count prediction is 0.1229 and MAE is 0.2484.

In order to better clarify the influence of coaching, the SHAP model was combined and analyzed, and the results are shown in Figure 4 and Figure 5. It can be found that:

a) Coach performance is a very important feature in the medal prediction model. SHAP analysis shows that better performing coaches can significantly increase the number of medals of athletes, especially the number of gold medals and total medals. Therefore, in practical applications, coach performance can be used as a more critical feature to optimize the model.

b) Although years of coaching experience may theoretically affect athlete performance, SHAP value analysis suggests that the actual performance of the coach (rather than years of experience) has a greater impact on the number of medals won by the athlete.

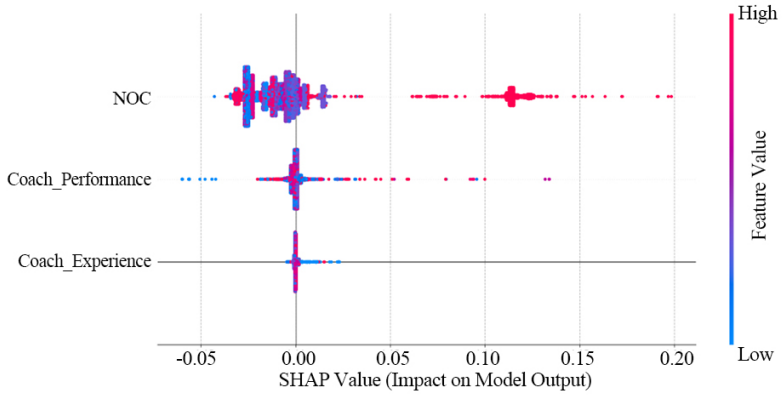

Figure 4. Analysis of the number of gold medals

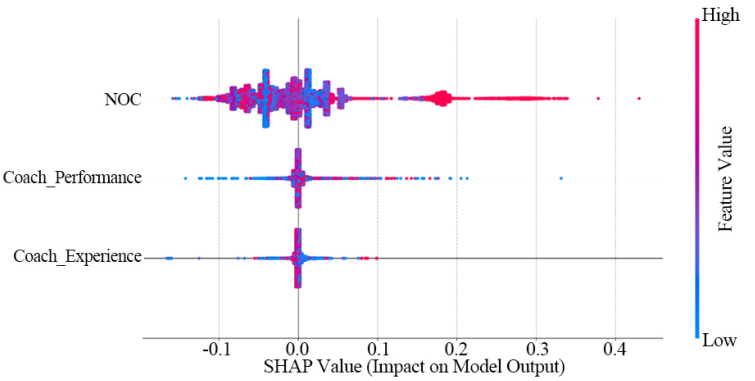

Figure 5. Analysis of total medals

By improving the performance of coaches, the total number of medals predicted for certain countries increases significantly, suggesting that improving the level of coaching has a positive impact on the number of medals won by these countries. This suggests that the role of "great coaches" in medal counts is critical, especially in countries that need to improve. However, for some other countries, improving coaching performance has a relatively small impact on the total number of medals, perhaps because their athletes are already highly competitive or because other factors (e.g., athlete talent, training facilities, etc.) dominate their medal performance.

5. Conclusion

The hybrid prediction framework integrating random forest regression and gradient boosting tree constructed in this study effectively solves the modelling and analysis challenges of complex features in multivariate systems. Firstly, the study achieves efficient capture of multidimensional feature nonlinear associations based on the random forest model through self-sampling and feature random selection mechanism, and its built-in feature importance analysis provides a quantitative basis for variable screening. Secondly, logistic regression and linear regression enhance the statistical explanatory ability of the model from the perspective of dichotomous prediction and multivariate interaction, respectively, and clearly present the differences in feature contribution through One-Hot coding and visualisation techniques. In addition, this study constructs a gradient boosting tree combined with SHAP value analysis, which significantly improves the prediction accuracy by iterative fitting of residuals, and realises the transparency of the decision-making process of the model by quantifying the importance of features. This framework not only breaks through the limitations of a single algorithm in terms of accuracy and interpretability, but also provides a reusable technical paradigm for predictive analyses of multivariate systems through multi-dimensional feature engineering and algorithmic complementarity. Future research can further explore the fusion path of dynamic features and multi-source data, continuously optimise the model parameter settings, and promote the expansion of the framework in a wider range of applications.

References

[1] Li Hongda. Research on land cover classification of Sentinel-2 multi-seasonal data based on gradient boosting tree and random forest[D]. Qinghai Normal University, 2021.

[2] Zhou Yunhao, Yang Baojie, Liu Dan, et al. Modelling and simulation of predictive analy-sis of power engineering data based on random forest algorithm [J]. Electronic Design Engineering, 2024, 32 (04): 103-106+111.

[3] Guo Yanhao, Do Jie, Xiang Zilin, et al. Evaluation of landslide susceptibility of Wenchuan co-seismic landslide based on gradient boosting decision tree and random forest with optimised negative sample sampling strategy [J]. Geoscience Bulletin, 2024, 43 (03): 251-265.

[4] Zou Hang, Jiang Yunlu. A review of methods for selecting robust variables for high-dimensional linear regression models [J]. Applied Probability Statistics, 2024, 40 (01): 157-181.

[5] Gong Yue, Luo Xiaoqin, Wang Dianhai, et al. Gradient boosting regression tree-based travel time prediction for urban roads [J]. Journal of Zhejiang University (Engineering Edition), 2018, 52 (03): 453-460.

[6] Nie Hu, Wu Xiaoyan. A study of factors influencing depression combining gradient boosting tree algorithm and interpretable machine learning model SHAP [J]. Data Analysis and Knowledge Discovery, 2024, 8 (03): 41-52.