Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 5, No. 2, 2025

Construction and Application of a Hybrid Predictive Model Incorporating Random Forest and Multiple Linear Regression

Zhicai Zhang[1, *], Yunzhao Zhou[2], Zhicheng Ye[3]

[1]School of Mechanics and Engineering Sciences, Hohai University, Nanjing, China

[2]School of Hydrology and Water Resources, Hohai University, Nanjing, China

[3]School of Mathematics, Hohai University, Nanjing, China

[*] **Corresponding author**: Zhicai Zhang(Email: 3262621659@qq.com)

Abstract: In this study, a prediction framework integrating random forest and multiple linear regression is constructed to focus on the quantitative analysis and prediction of multidimensional features. Firstly, a random forest decision tree is constructed based on the Gini index, and the classification and regression tasks are achieved by ranking the importance of features, completing the model parameter setting and error validation, so as to achieve the modelling and prediction of the non-linear relationship of multi-dimensional features. In addition, the study introduces independent variables such as dichotomous variables, occupancy category indicators and capacity values, fits a multiple linear regression model relying on the least squares method, and tests for multiple covariances through variance-inflated factor tests to quantify the extent to which specific factors influence the results. The framework enhances the stability and generalisation ability of multivariate system modelling through multi-level model collaboration and data processing optimisation, and provides a scalable technical paradigm for related fields.

Keywords: Random Forest Model, Multiple Linear Regression, Gini Index, Feature Engineering.

1. Introduction

In the field of data-driven modelling and decision-making, complex feature resolution and accurate prediction of multivariate systems are the core challenges. Existing single model architectures often fail to balance nonlinear feature mining and linear effect quantification, resulting in insufficient prediction accuracy and explanatory power [1]. In order to break through this limitation, this study proposes a hybrid prediction framework integrating random forest and multivariate linear regression [2], aiming to achieve full coverage from nonlinear pattern recognition to linear effect quantification through the synergy of multilevel algorithms.

The study first relies on the decision tree integration property of the random forest model, constructs a feature screening mechanism for classification and regression tasks based on the Gini index, and achieves nonlinear modelling and prediction of complex data distributions through adaptive mining of higher-order correlations among multidimensional features [3]. In addition, independent variables such as binary variables, percentage indicators and capacity values are introduced, and multiple linear regression models are fitted by the least squares method to quantify the direction and degree of the influence of specific factors on the prediction results and enhance the model interpretability [4]. The innovation of this study lies in the construction of a cross-algorithm collaborative modelling paradigm, which provides a new path to solve the prediction and interpretation challenges in high-dimensional data scenarios. The framework can be widely used in social science, engineering and technology, etc., to help the scientific and intelligent development of decision-making in related fields through data-driven multilevel analysis.

2. Medal count based on random forest model

2.1. Predictions of Olympic medals by country

2.1.1. Random forest model interpretation

Random Forest model is an advanced predictive model, mainly used to complete the task of classification and regression of a large number of features, after the classification of a large number of features to build a large number of "decision trees", each decision tree is constructed by random sampling of the data for training (constantly splitting the nodes), there is no correlation between the tree and the tree [5]. Each tree makes independent predictions, which are eventually merged to improve the accuracy and stability of the predictions. Compared to general prediction models such as ARIMA model, LSTM model, linear regression model performs better in dealing with large scale datasets and high dimensional data.

The steps for applying the Random Forest model are roughly as follows:

a) Selection of different eigenquantities.

b) Implementation of classification of features by MATLAB, filtering features by importance.

c) Evaluate the node purity of the decision tree based on the Gini index with the following equation:

$$\mathrm{Gini}(\Omega) = 1 - \sum_{i=1}^{k} p_i^2 \tag{1}$$

d) Final prediction: if the prediction belongs to a classification problem, the predictions of the decision tree are voted on the following equation:

$$\hat{y} = \text{ majority } \{y_1, y_2, y_3, \ldots, y_N\} \tag{2}$$

If the prediction is a regression problem, the predictions from the decision tree are averaged.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i \tag{3}$$

2.1.2. Illustration and quantification of different traits

Firstly, the number of medals for each participating country in the 2028 Olympic Games in Los Angeles is predicted and prediction intervals are given for all outcomes. It was necessary to select the various characteristics that may affect the number of medals a country will win, and the following characteristics were identified:

a) Statistics on the weight of dominant items

This paper defines a dominant item as one in which there is a country with a clear advantage, therefore, a non-dominant item, i.e. one in which there is no country with a clear advantage, is more random. Obviously, the share of each country in the dominant sports will greatly affect the number of medals, especially gold medals, that country will end up with. This paper filtered out the possible advantageous items of different countries, for example, the strong advantage of the United States in track and field and swimming two sports. At the same time using Excel's summing and other functions to calculate the corresponding project different countries in the previous Olympic Games to win the proportion of awards, such as the United States in 2024 in the swimming program accounted for the proportion of 0.37.2024 countries accounted for the proportion of dominant projects. Figure 1 shows the proportion of dominant sports in each country, taking athletics as an example.

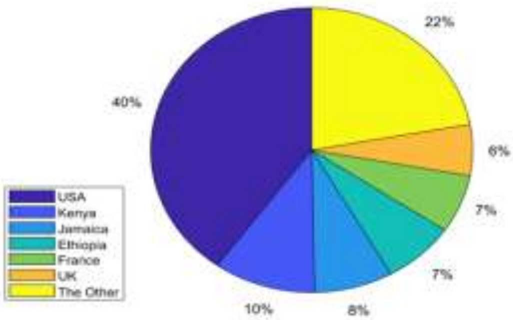

Figure 1. Athletics dominant countries percentage

b) Host country advantages

This paper has noticed that every year the host country has a very large advantage in participating in the Olympic Games, and to measure this advantage this paper has introduced K (the coefficient of the host country's advantage), where K is calculated by the formula.

$$K = \frac{X_1}{X_2} \tag{4}$$

Where X_1, X_2 are the number of medals won by the host country in the previous and current year, respectively.

By looking up the data, this paper derived the K values for all years. As shown in Figure 2, the dispersion of K values are more concentrated. From this, this paper decided to use least squares to fit on the curve.

$$Y = \beta_0 + \beta_1 X \tag{5}$$

$$\beta_1 = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_i - \bar{X}\right)^2} \tag{6}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \tag{7}$$

$\bar{X}$ and $\bar{Y}$ are the mean values of x and y, respectively, β_1 is the slope and β_0 is the intercept.

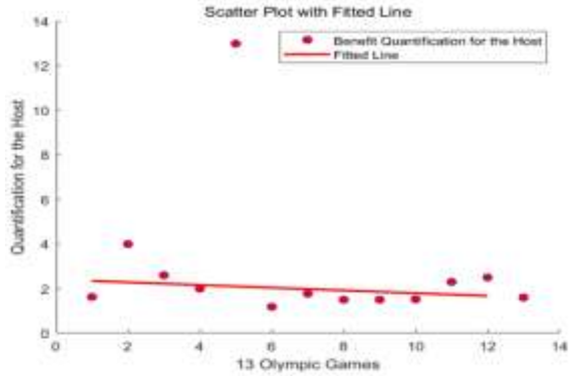

Figure 2. Host country advantages

c) Overall number of competitive events statistics

A review of the relevant literature shows that the number of overall competitive events at previous Olympic Games also has an impact on the final number of gold and medals. It is possible to derive the number of competitive events as well as the overall number of events in previous Olympic Games from 1996-2024.

d) Statistics on the number of participating athletes

Experience has shown that the number of athletes participating in previous Olympic Games. The number of athletes will also be associated with the final medal data. According to the athletes of the Summer Olympic Games, this paper first filtered the data from 1996-2024 based on the time, and then filtered the data based on the country to get the statistics of the number of athletes who participated in each country in the past.

2.1.3. The Practical Application of Model

After explicit statistical quantification of each feature quantity, regarding the implementation of decision tree, this paper selected the best feature as the dividing feature of the current node based on the Gini index (i.e., whichever feature brings the most magnitude of information change is selected for classification), and then based on the selected feature, the dataset is divided into several subsets based on the difference of feature values. Subtrees are constructed recursively for each subset until all subsets are correctly categorized. The decision tree is also pruned to avoid overfitting. Combine multiple decision trees together, each time the dataset is selected randomly with put back, and at the same time some of the features are randomly selected as inputs, i.e., creating a random forest.

The related program was written to perform the screening of random forest features, and the specific results are shown in Figure 3. The number of leaves was determined to be 5, and the number of decision trees to be 100.

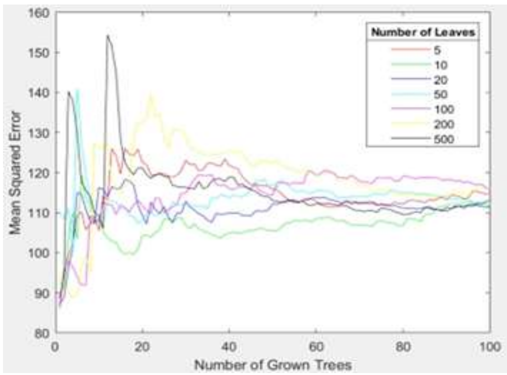

Figure 3. Decision tree-leaf node relationship diagram

After these tasks are completed, the Random Forest model is finally obtained for the number of gold medals and the total number of medals for each country in 2028, and the visualization results are shown in Figure 4 and Figure 5.

The predictions obtained are broadly characterized and possible corresponding explanations are given.

a) Individual European countries such as Great Britain and Germany will pick up in medal counts.

b) Although East Asian countries and many European countries are predicted to see a decline in their performance, the overall thickness of their performance is still guaranteed, thanks to a sufficient economic base to ensure investment in sports.

c) As the host country, the United States of America has a clear advantage in the number of gold medals and the total number of medals. This can be attributed to the fact that the host country has set up programs that are closer to its own sports training situation, and has a better correspondence between the athletes' reserves and the programs.

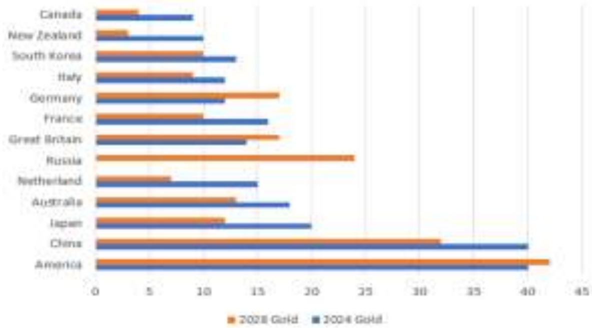

Figure 4. Gold medal count predictions for each country

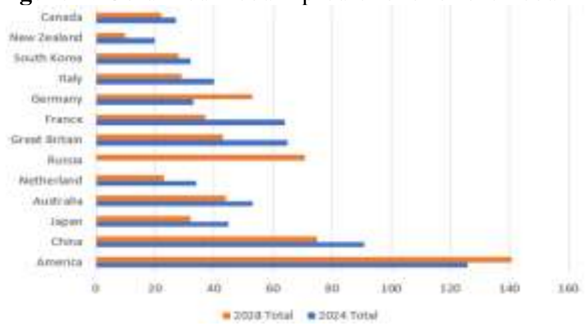

Figure 5. Total medal count projections by country

Also based on the random forest regression forecasts, this paper obtained forecast intervals for the corresponding countries as shown in Table.1.

Table 1. Gold and Overall Medal Prediction Ranges

Country	Gold Forecast Interval	Overall Medal Prediction Ranges
United States of America	(35, 47)	(106, 140)
China	(24, 40)	(69, 81)
Russia	(19, 29)	(50, 60)
Germany	(15, 19)	(51, 55)
Great Britain	(14, 29)	(28, 73)
Japan	(10, 21)	(20, 52)
Australia	(9, 17)	(27, 64)
France	(9, 11)	(33, 41)
Italy	(8, 11)	(28, 30)
South Korea	(7, 12)	(25, 31)

2.2. Projections for first-time award winners

In order to predict the country that will win its first ever medal at the 2028 Olympic Games in Los Angeles, this paper still used the Random Forest model for our predictions, however, since the number of medals for countries that have never won a medal is zero, this paper no longer choosed either the number of gold medals or the number of medals as a criterion for our predictions.

In the previous section, this paper has quantified the impact of the number of athletes participating on the results of the competition, and for countries that have never won a medal, this paper considered the number of participants to be the only criterion for the chances of winning a medal for the first time. Therefore, this paper has filtered out the following countries from the list of countries that have never won a medal based on the number of participants: Bolivia, Nepal, Maldives, Niger, Congo Brazzaville, Togo, Chad and Bangladesh. So this paper predicted the odds of winning based on the ratio of the number of people who chose athletics and swimming. The odds of these eight countries winning for the first time are shown on Figure 6.

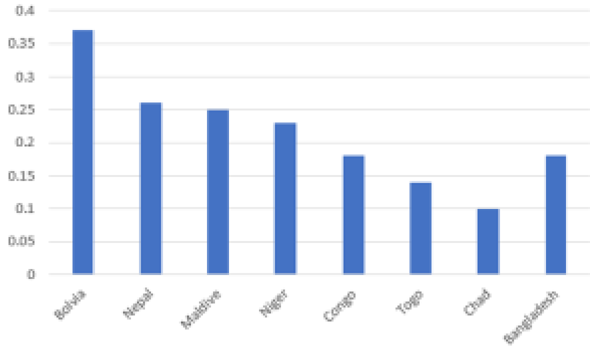

Figure 6 Breakthrough first medal national probability chart

Based on the image, it can be concluded that Bolivia is most likely to win its first ever gold medal in the 2028 Los Angeles Olympics, followed by Nepal, Maldives and Niger respectively.

2.3. Relationship between host country's choice of events and medals

2.3.1. Relationship Between Number of Program and Medals

Based on our 3 classification of the features and the extraction of the data, this paper can get heat map as in Figure 7.

It is easy to see that the number of items set up in the Olympic Games has a strong and negative correlation with the number of gold medals, and the correlation with the total number of medals is not obvious. It can also be seen from the significance of the characteristics that the number of items set up in the Olympic Games has a significant effect on the number of gold medals, and the effect on the total number of medals is not obvious.

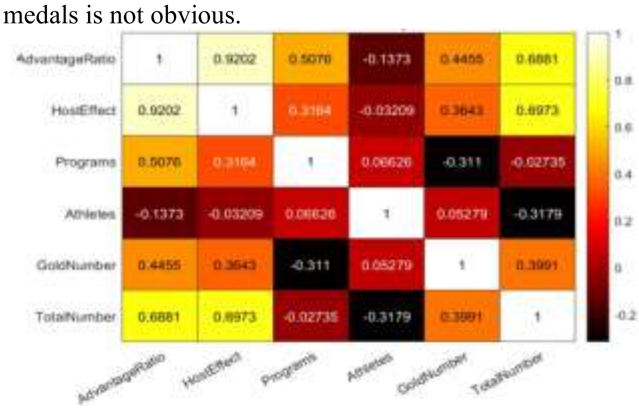

Figure 7. Host country advantages

2.3.2. Relationship Between Program Types and Medals

The dominant events show a strong positive correlation with both gold medals and the total number of medals, and it can also be seen from the significance of the characteristics that the dominant events have a significant impact on both the gold and the total number of medals. The percentage of dominant sports in each country up to 2024 has been given in Fig. 1 and Fig. 2. The choice of the type of sport by the host country is often directly linked to the performance of the competition. For example, in the 2028 Olympic Games in Los Angeles, weightlifting was cancelled, which directly led to a significant reduction in China's gold and medal totals.

3. The effect of great coach based on the multiple linear regression model

3.1. The definition of great coach effect

The 'great coach effect' refers to the impact that a coach with a high level of experience and a track record of success has on a country's performance in a particular sport when he or she coaches in more than one country. And it has the opposite effect on it. For example, a coach may have unique training methods, team management, tactical choices or experience to enhance an athlete's performance, thereby helping the country win more medals.

On the other hand, since the acquisition of outstanding athletes' ability requires excel- lent athletic talent, in individual sports, although the correct training program of great coaches still helps to improve athletes' strength, the effect of great coaches in group sports is undoubtedly greater compared with that of group sports for the cooperation between groups, the planning of tactics, and the selection of athletes suitable for the team. So in order to analyze the role of great coaches more accurately, this paper chose to analyze the great coaching effect in group sports.

3.2. Practical Application of Model

Firstly, this paper identified the factors to be included in the linear regression analysis along with the "great coach effect", and after combining the context of the problem and filtering the data, this paper determined the dependent variable in the model to be the total number of medals won by each country in the event Y.

This paper identified the independent variables as the following three factors:

a) X_1 represents "Great Coach Effect". X_1 is a binary variable, i.e., performance is quantified as 1 in the presence of that coach and 0 in the absence of that coach.

b) The number of events in which the country participates as a percentage of the over-all number of Olympic entries, X_2, reflects the importance the country attaches to this event.

c) X_3 represents the level of competence of the country's sports teams.

In gymnastics the main expression is the distribution of the number of gold, silver and bronze medals, which this paper quantifies as a competence value by calculating its weighted average, with gold medals counting for 0.5, silver medals for 0.3, and bronze medals for 0.2; in volleyball this paper introduce a computational formula between the ranked place z and the competence value of the sports team:

$$X_3 = 10 - z \tag{8}$$

In order to prevent the problem of multicollinearity between independent variables, this paper consider the value of VIF (Variance Inflation Factor). VIF is an important indicator of multicollinearity. It indicates the extent to which a particular independent variable is correlated with other independent variables. Generally, if the VIF value of a particular independent variable is greater than 10, it indicates that there maybe a more serious problem of covariance. The greater the VIF value, the higher the redundancy of that independent variable in the model.

$$VIF_i = \frac{1}{1 - R_i^2} \tag{9}$$

After analysing the data through the program, it was found that the VIF values of all the factors were less than 10, i.e., the independent variables were not correlated with each other two by two, which is inline with the application of the multiple linear regression model. For example, the value of the percentage of the number of events was 1.739. Based on the data set, it can be very easy to get that during the coaching period of gymnastics coach Béla Károlyi, there were:

Total number of medals in Romanian gymnastics:

$$Y_1 = 4.6474X_1 + 3.8007X_2 + 1.7415X_3 - 52.888 \tag{10}$$

Total US Gymnastics medals:

$$Y_2 = -0.053X_1 + 0.0051X_2 + 3.3143X_3 - 0.0096 \tag{11}$$

Similarly, it is possible to get volleyball coach Lang Ping's coaching period with: China volleyball total medals:

$$Y_3 = -0.0333X_1 + 0.0021X_2 + 1.0833X_3 \tag{12}$$

Total medals in USA volleyball:

$$Y_4 = 0.2388X_1 + 0.0672X_2 + 2.0896X_3 - 0.1343 \tag{13}$$

According to the regression coefficients, it can see that the great coaching effect exists, especially during gymnastics coach Béla Károlyi's tenure in Romania, where the influence of the coach is not negligible or even dominant; however, the great coaching effect is not evident in the data from the United States and China.

Based on the data, three countries were selected with the aim of winning gold medals at the 2028 Olympic Games and identified sports in which investment in 'great coaches' should be considered: Italian women's volleyball, Australian women's water polo, and Argentine hockey.

4. Conclusion

In this study, a prediction model framework integrating Random Forest and multiple linear regression is constructed to achieve integrated analysis of complex features for multivariate systems. Firstly, the study constructs a decision tree integration through Gini index, and makes use of self-sampling and feature random selection mechanism to effectively capture the nonlinear association between multidimensional features and improve the modelling efficiency and prediction accuracy. In addition, the study relies on the multiple linear regression model to introduce independent variables such as binary variables and percentage indicators, fit the linear relationship based on the least squares method, and test the multicollinearity through the variance expansion factor, so as to systematically quantify the influence paths and effect strengths of specific factors on the prediction results, and to make up for the shortcomings of the machine learning model's interpretability. The predictive model framework breaks through the limitations of a single algorithm and achieves a balance between prediction accuracy and explanation ability through cross-model collaboration, providing a reusable technical paradigm for similar high-dimensional data scenarios. In the future, the model parameters can be further optimised, and the fusion of dynamic features and multi-source data can be explored to improve the timeliness and applicability of data-driven prediction.

References

[1] Qin Yanjiao. Comparative study of house price prediction models based on multiple linear regression and random forest algorithm [J]. Modern Information Technology, 2024, 8 (22): 127-131.

[2] Chen Junqing, Li Yi, Wang Bin, et al. Prediction of extreme temperature events in the Yellow River Basin based on stepwise multiple linear regression and random forest model [J]. Journal of Natural Hazards, 2024, 33 (01): 74-88.

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

[4] Zhang Lina, Jiang Zhicheng, Liu Dayong, et al. Correlation flow data processing method based on K-means cluster analysis and multiple linear regression [J]. Petroleum Pipes and Instruments, 2024, 10 (01): 52-56+62.

[5] Zhang Kunbin, Chen Yuming, Wu Keshou, et al. A study on random forest classification algorithm driven by granular vectors [J]. Computer Engineering and Applications, 2024, 60 (03): 148-156.

[6] Lee, J., Min, K., The Effect of Economic Power and Population Size on the Athlete's Size and Olympic Achievements, Korean Society of Sport Policy, 11(2), 97-109, 2013.