Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 1, No. 1, 2026

Olympic Medal Prediction Method Based on Random Forest and Multilayer Perceptron Models

Liuxi Zhang *

North China Electric Power University (Baoding Campus), Baoding, China
***Corresponding author:** 3190837832@qq.com

Abstract: This paper constructs a medal count prediction model based on historical Olympic data, utilizing the Random Forest algorithm to model nonlinear relationships within complex datasets. First, the raw data undergoes feature selection and correlation analysis to extract key input features from variables such as athlete potential, host country factors, number of events, historical medal records, and gender. The statistical correlations among these features are verified using a correlation matrix. Building on this foundation, multiple training subsets are generated using Bootstrap sampling. A Random Forest prediction model is constructed through a multi-decision tree ensemble mechanism, and model parameters are optimized by adjusting the number of nodes and decision trees. To validate the model's performance, multi-layer perceptron, multiple linear regression, and support vector regression models are also constructed for comparative experiments. Comprehensive evaluation is conducted using metrics such as mean squared error, root mean square error, mean absolute error, and coefficient of determination. The experimental results indicate that the Random Forest model outperforms the other models in terms of prediction accuracy and stability. It effectively captures the nonlinear characteristics within complex data and achieves accurate predictions of medal distributions, demonstrating good generalization ability and practical value.

Keywords: Random Forest; Multilayer Perceptron; Medal Prediction.

1. Introduction

In complex data environments, how to effectively predict future outcomes using information from multiple sources has become a key research focus in the fields of data analysis and intelligent computing. As the scale and structural complexity of data continue to increase, traditional modeling methods based on single linear relationships often struggle to produce stable and accurate predictions when dealing with data characterized by multiple variables and pronounced nonlinear features. Consequently, machine learning models capable of simultaneously handling multidimensional features and capturing complex relationships have gradually emerged as essential tools in data prediction research.

In related research, linear regression models have been widely applied to data relationship analysis due to their simple structure and interpretability; however, their reliance on the assumption of linearity between variables limits the model's adaptability to data with complex structures. Meanwhile, although neural network models possess strong nonlinear fitting capabilities, they are prone to training instability or difficulties in parameter tuning when sample sizes are limited or feature structures are complex. Furthermore, methods such as Support Vector Regression (SVR) offer certain advantages when handling high-dimensional features, but the selection of model parameters and kernel functions often significantly impacts predictive performance.

To address these issues, this paper proposes a prediction model framework centered on Random Forests, which effectively captures complex data relationships through an ensemble learning mechanism. This method generates multiple training subsets via Bootstrap sampling and employs a multi-decision tree structure for ensemble prediction, thereby enhancing model stability and generalization capabilities while reducing the risk of overfitting. During model construction, key feature variables are introduced and combined with correlation analysis to optimize the input data, thereby improving the model's ability to identify critical information [1].

Within this framework, a predictive model was constructed using Olympic medal data as an example, and comparative analyses were conducted using methods such as multi-layer perceptrons, multiple linear regression, and support vector regression. Experimental results indicate that this model achieves superior predictive performance in complex multivariate data prediction tasks, providing a valuable modeling approach for predictive problems in similar data scenarios.

2. Data Preprocessing Method Based on Feature Selection and Correlation Analysis

2.1. Feature Selection

To improve the effectiveness and reliability of the medal prediction model, it is essential to identify informative input variables from the available dataset. Therefore, feature selection is conducted to determine the variables that significantly influence Olympic medal outcomes [2].

The dataset contains multiple attributes describing athletes and competitions, including competition year, country of origin, athlete ranking score, medal counts (gold, silver, and bronze), total medals, host country information, and the number of competition events. These attributes provide important information for modeling medal distributions.

From the perspective of athletic performance, age is a key factor affecting competitive outcomes. Athlete performance typically increases during the early stages of a career, reaches a peak period, and gradually declines afterward. Based on this observation, athletes participating in two Olympic Games are regarded as elite competitors with strong medal-winning

potential. Athletes participating in three or more Olympic Games are considered to have relatively declining physical performance, whereas athletes participating in only one Olympic Games are regarded as potential competitors with future development potential.

The country of origin may also influence athlete performance due to environmental familiarity and audience support. When a country hosts the Olympic Games, its athletes may benefit from a host advantage, which can positively affect medal outcomes. Furthermore, the number of competition events directly influences medal distribution, as more events increase the opportunities for medal acquisition.

Gender is also considered as an important feature variable. Physiological characteristics, training environments, and participation structures differ across populations, which may lead to different performance patterns across sports events. In addition, historical medal counts are incorporated as an important feature since previous achievements often reflect the long-term competitiveness of a country in international sporting competitions.

2.2. Correlation Analysis

After identifying candidate features, correlation analysis is conducted to evaluate the relationships among the selected variables and to verify the rationality of the feature selection process. The correlation coefficients between variables are calculated and visualized using a heatmap.

Figure 1 illustrates the correlation coefficient matrix of the selected features. The heatmap clearly reveals the strength of dependencies among the variables and provides insights into their influence on medal prediction. Features with stronger correlations generally indicate higher statistical relevance and stronger predictive potential.

The correlation analysis confirms that the selected variables exhibit meaningful relationships consistent with the characteristics of Olympic competition data. These findings provide a reliable foundation for subsequent machine learning model construction.

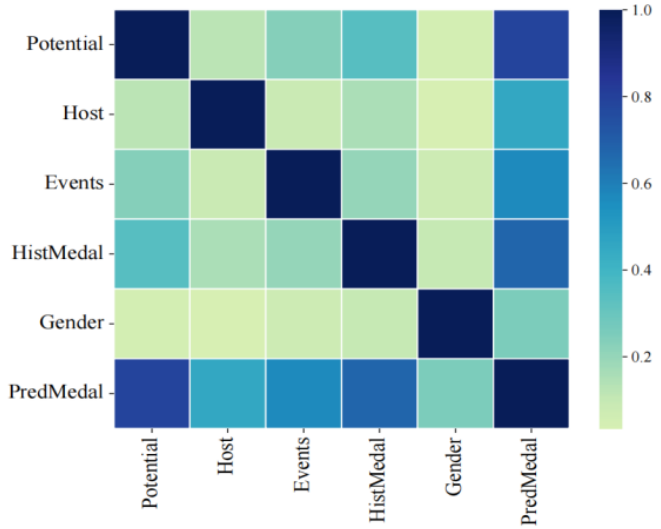

Figure 1. Feature Correlation Heatmap

3. Random Forest-Based Medal Prediction Model Construction

3.1. Random Forest Principle

Random Forest is an ensemble learning algorithm proposed by Breiman that combines bootstrap aggregation and random feature selection to construct multiple decision trees. By aggregating predictions from multiple trees, the model effectively captures nonlinear relationships within complex datasets and improves prediction accuracy.

In the Random Forest framework, each decision tree is trained using a bootstrap sample drawn from the original dataset. At each node of the tree, a random subset of features is selected for splitting. This mechanism introduces diversity among trees and significantly reduces the risk of overfitting, thereby enhancing the generalization capability of the model.

The number of decision trees and the number of candidate features selected at each split are two important parameters that strongly influence model performance. Proper parameter selection helps achieve a balance between model complexity and predictive accuracy [3].

3.2. Model Construction

(1) Data Preparation

To estimate the total number of medals obtained by each country, the numbers of gold, silver, and bronze medals are first predicted separately and then aggregated to obtain the total medal count.

The input data used in the Random Forest model include historical medal statistics, host country indicators, event information, and other relevant variables. After feature extraction and preprocessing, the final input variables used in the model are Potential, Host, Events, HistMedal, and Gender.

Bootstrap sampling is applied to generate multiple training subsets. The dataset is divided into training, validation, and test sets with proportions of 70%, 15%, and 15%, respectively. This process is repeated k times to generate k sub-training datasets denoted as $S_1, S_2, \ldots, S_k$.

(2) Parameter Selection

Two key parameters must be determined when constructing the Random Forest model: the number of candidate features selected at each node and the number of decision trees in the forest.

The number of candidate features selected at each split should not exceed the total number of input variables. After determining the candidate feature size, the number of decision trees is adjusted to ensure stable model performance while avoiding overfitting or underfitting.

(3) Feature Importance Evaluation

To evaluate the importance of each input feature, the weighted variance reduction criterion is adopted within the Random Forest framework.

For a regression tree with node N, the variance of node N is defined as

$$\sigma^2 = \frac{1}{C}\sum_{k\in N}(Y_k - \bar{Y})^2 \tag{1}$$

where C represents the sample size and Y_k denotes the output value.

The impurity reduction after splitting node N using feature I can be expressed as

$$\sigma^2(I,N) = \sigma^2 - \left(\frac{C_1}{C}\sigma_1^2 + \frac{C_2}{C}\sigma_2^2\right) \tag{2}$$

where C_1 and C_2 represent the sample sizes of the two child nodes, and σ_1^2 and σ_2^2 denote their corresponding variances.

The importance of each feature is determined by the magnitude of impurity reduction it produces during tree construction. Features generating larger impurity reductions contribute more significantly to the predictive performance of the model.

3.3. Result Analysis

Model parameters are determined through iterative experimentation. Different combinations of node sizes and numbers of decision trees are evaluated to identify the configuration that maximizes prediction accuracy.

Figure 2 illustrates the relationship between model parameters and the explained variance. The results show that when the node size is set to 4, the explained variance is higher than when the node size is set to 2 or 3. Therefore, the optimal node size is determined to be 4.

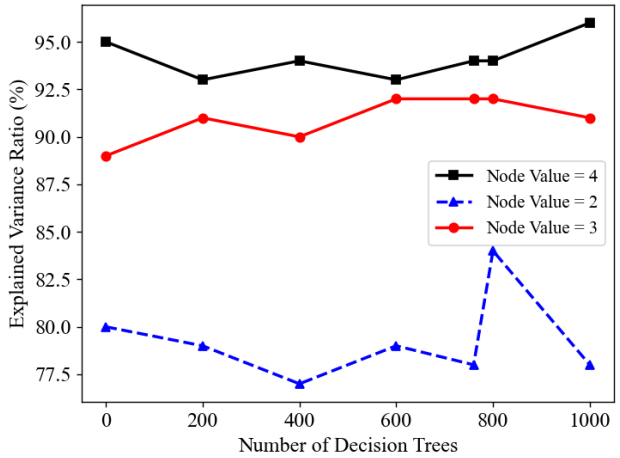

Figure 2. Parameter Sensitivity of Random Forest

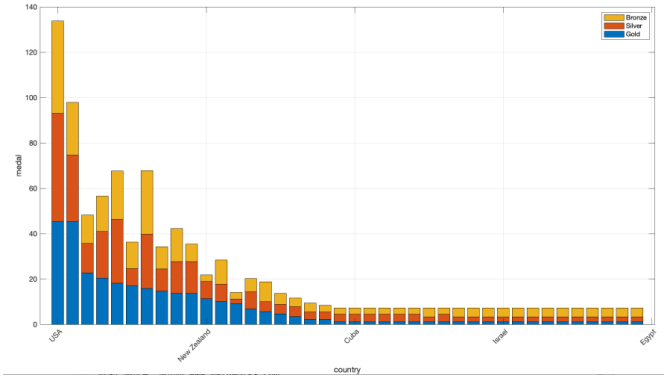

Figure 3. Random Forest Medal Prediction Results

When the number of decision trees is set to 700, 800, and 900, the explained variance values remain relatively stable. Considering computational efficiency and model interpretability, the number of trees is ultimately set to 700.

Using the optimized parameters, the Random Forest model is applied to the processed dataset to generate medal predictions. Figure 3 illustrates the predicted medal distribution results.

The results indicate that the Random Forest model provides stable and accurate predictions of medal outcomes, demonstrating its effectiveness in modeling complex relationships within Olympic competition data.

4. Construction of Multiple Machine Learning Prediction Models

4.1. Multilayer Perceptron Model

To further evaluate the performance of different machine learning approaches, a Multilayer Perceptron (MLP) model is constructed for Olympic medal prediction.

The sigmoid activation function is adopted to facilitate gradient computation and improve the convergence behavior of the neural network. The input variables of the MLP model are consistent with those used in the Random Forest model, including Potential, Host, Events, HistMedal, and Gender [4].

The prediction error of the neural network is defined as

$$Loss = (y - \hat{y})^2 \tag{3}$$

where y represents the true value and $\hat{y}$ denotes the predicted value.

The optimization objective of the model is to minimize the total loss across all training samples:

$$\min \sum_{i=1}^{N} Loss_i \tag{4}$$

Based on the neural network structure, the output of the hidden layer can be expressed as

$$Loss = \left(y - \sum_{j=1}^{m} w_{kj} \sigma \left(\sum_{i=1}^{n} v_{ji} x_i + b_j \right) \right)^2 \tag{5}$$

The parameters that require optimization include the weight matrices v and w, as well as the bias term b. Gradient backpropagation is used to update the parameters.

The gradient of the loss function with respect to the parameters is

$$\frac{\partial Loss}{\partial w^T} (y - \hat{y}) \cdot v^T \cdot \sigma' \cdot x \tag{6}$$

To improve computational efficiency, stochastic gradient descent (SGD) is adopted. Let θ represent the parameters to be optimized. The iterative update rule is

$$\theta_{k+1} = \theta_k - \eta_k d_k, \text{where } d_k = \frac{\partial Loss}{\partial \theta} = \nabla(\theta) \tag{7}$$

The iterative process continues until the loss function converges to a stable minimum, thereby completing the training of the MLP prediction model.

4.2. Multiple Linear Regression Model

To further analyze the relationships between explanatory variables and Olympic medal outcomes, a Multiple Linear Regression (MLR) model is established. This model describes the linear relationship between medal counts and several explanatory variables obtained from the feature analysis stage [5].

Let $Y^{\text{gold}}_{\text{medal},i,t}$, $Y^{\text{silver}}_{\text{medal},i,t}$, and $Y^{\text{bronze}}_{\text{medal},i,t}$ denote the numbers of gold, silver, and bronze medals obtained by country i at Olympic Games t. The regression models are defined as

$$\begin{aligned} &Y^{\text{gold}}_{\text{medal},i,t} \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + \beta_3 X_{it3} + \beta_4 X_{it4} + \varepsilon_i \\ &Y^{\text{silver}}_{\text{medal},i,t} \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + \beta_3 X_{it3} + \beta_4 X_{it4} + \varepsilon_i \\ &Y^{\text{bronze}}_{\text{medal},i,t} \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + \beta_3 X_{it3} + \beta_4 X_{it4} + \varepsilon_i \end{aligned} \tag{8}$$

The total medal count can therefore be expressed as

$$Y^{\text{total}}_{\text{medal},i,t} Y^{\text{gold}}_{\text{medal},i,t} + Y^{\text{silver}}_{\text{medal},i,t} + Y^{\text{bronze}}_{\text{medal},i,t} \tag{9}$$

where β_0 represents the intercept term, $\beta_1, \beta_2, \beta_3, \beta_4$ denote regression coefficients, and ε_i is the random error term.

The regression coefficients quantify the strength and direction of the relationships between explanatory variables and medal outcomes. The model therefore provides an interpretable baseline for Olympic medal prediction.

4.3. Support Vector Regression Model

Support Vector Regression (SVR) is employed to capture nonlinear relationships between input variables and medal outcomes. The SVR model constructs an optimal regression function by minimizing model complexity while maintaining prediction accuracy [6].

The optimization objective of the SVR model can be formulated as

$$min \quad _{w,b} \frac{1}{2}|w|^2, \text{ s.t. } |y_i - (wx_i + b)| \leq \varepsilon \qquad (10)$$

where w represents the weight vector, b denotes the bias term, and ε defines the allowable prediction error margin.

To account for samples located outside the ε-insensitive region, slack variables are introduced, resulting in the following optimization objective:

$$\min_{w,b} \frac{1}{2}|w|^2 + \text{loss} \qquad (11)$$

To model nonlinear relationships, a kernel function is used to map the input space into a higher-dimensional feature space. The regression function can therefore be expressed as

$$f(x)\omega^T\phi(x) + b\sum_{i=1}^{m}\alpha_i y_i \phi(x_i)^T\phi(x) + b\sum_{i=1}^{m}\alpha_i y_i K(x, x_i) + b \qquad (12)$$

where $\phi(x)$ denotes the feature mapping function and $K(x, x_i)$ represents the kernel function.

5. Performance Evaluation of Prediction Models Based on Multiple Metrics

Four machine learning models are constructed to predict Olympic medal distributions: Random Forest, Multilayer Perceptron, Multiple Linear Regression, and Support Vector Regression.

During the model training process, the dataset is divided into training and testing subsets. Model performance is evaluated using several standard regression metrics, including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination R^2.

Figure 4 illustrates the performance comparison among the four models. The Random Forest model consistently achieves lower prediction errors and higher R^2 values than the other models, indicating superior predictive performance.

Based on the Random Forest prediction results, the United States is expected to obtain approximately 135 medals in the next Olympic Games, including 44 gold medals, 48 silver medals, and 43 bronze medals. These findings demonstrate the effectiveness of ensemble learning methods in modeling complex Olympic medal distributions.

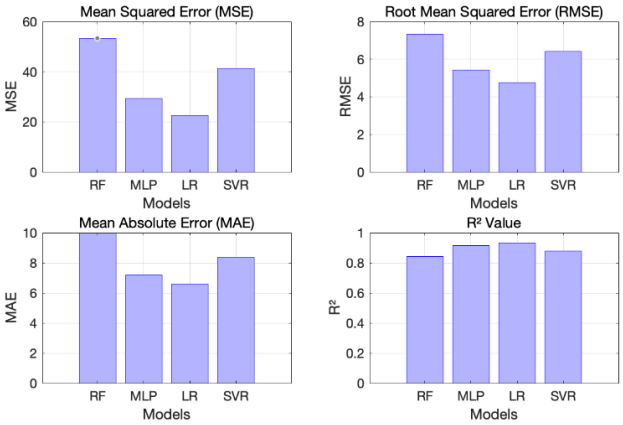

Figure 4. Performance Comparison of Prediction Models

6. Regression-Based Modeling of Event Contribution

The distribution of Olympic medals is closely related to the number and types of events in which a country participates. To quantify the contribution of different sports events to medal outcomes, a regression-based event contribution analysis is conducted.

Let $P^{\text{gold}}_{\text{medal},c,k}$, $P^{\text{silver}}_{\text{medal},c,k}$, and $P^{\text{bronze}}_{\text{medal},c,k}$ denote the numbers of gold, silver, and bronze medals obtained by country c in event k. The total medal counts can be written as

$$Y_{\text{gold},c}\sum_{k=1}^{m}P^{\text{gold}}_{\text{medal},c,k}, Y_{\text{silver},c}\sum_{k=1}^{m}P^{\text{silver}}_{\text{medal},c,k}, Y_{\text{bronze},c}\sum_{k=1}^{m}P^{\text{bronze}}_{\text{medal},c,k} \qquad (13)$$

Let N_k denote the number of events in category k. The event weight is defined as

$$w_k \frac{N_k}{\sum_{k=1}^{m}N_k} \qquad (14)$$

Using these weights, the total medal count for country c can be modeled as

$$Y_{\text{medals},c}\beta_0 + \sum_{k=1}^{m}\beta_k P_{\text{medal},c,k}w_k \qquad (15)$$

where β_0 represents the intercept term, β_k denotes the contribution coefficient of event k, $P_{\text{medal},c,k}$ represents the medal count obtained in event k, and w_k denotes the weight of the corresponding event.

Figure 5 presents the regression analysis results identifying key sports events for different countries. The results highlight the events that contribute most significantly to national medal performance and reveal event specialization patterns among participating countries.

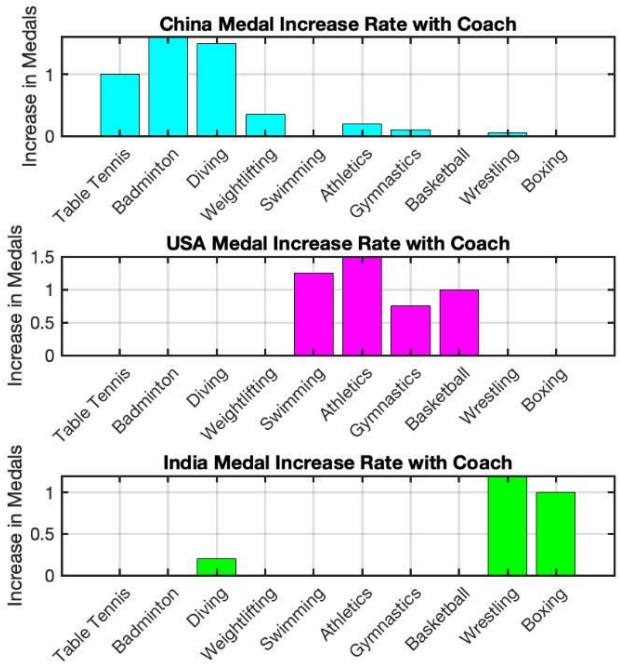

Figure 5. Contribution of Olympic Events to Medal Distribution

7. Conclusions

This paper focuses on the problem of predicting the number of Olympic medals and constructs a prediction model based on the Random Forest algorithm, which is then compared with Multi-Layer Perceptrons, Multiple Linear Regression, and Support Vector Regression models. Through feature selection and correlation analysis, key features are extracted from variables such as athlete potential, host country factors, number of events, historical medal records, and gender, and a data-driven prediction model is constructed based on these features. The model utilizes Bootstrap sampling and multi-decision tree ensemble mechanisms to enhance its ability to capture complex data relationships. Through parameter experiments, optimal structural settings were determined; for instance, the model performed stably when the number of nodes was set to 4 and the number of decision trees was approximately 700. Experimental results indicate that the Random Forest model outperforms other models in metrics such as mean squared error, root mean square error, and coefficient of determination, and is capable of accurately capturing the distribution characteristics of Olympic medals. For example, the model predicts that the United States will win approximately 135 medals at the next Olympic Games, including 44 gold, 48 silver, and 43 bronze medals. Future research could further incorporate additional data features and optimize the algorithm structure to enhance the model's predictive performance.

References

[1] Li Kai, Li Li, Yu Hang, et al. A Modeling Method for Atmospheric Precipitable Water Based on a Random Forest and Multi-Layer Perceptron Ensemble Algorithm [J]. Geodesy and Geodynamics, 2025, 45(05): 518–525. DOI: 10.14075/j.jgg.2024.05.231.

[2] Chen Xueying, Luo Chuan, Li Tianrui, et al. A Federated Evolutionary Feature Selection Algorithm Based on Rough Hypercubes [J]. Computer Research and Development, 2025, 62(11): 2710-2724.

[3] Zhao Kang, Li Xiaofan, Xue Jianru. Design of a Predictive Controller for Heavy-Duty Trucks Based on Random Forest Learning of Residuals [J]. Acta Automatica Sinica, 2025, 51(11): 2427–2440. DOI:10.16383/j.aas.c250207.

[4] Wang Xuemin, Xu Jingpei, He Yun. Stress and Temperature Prediction of Aircraft Engine Compressor Disks Based on Multi-Layer Perceptrons [J]. Journal of Aerodynamics, 2024, 39(04): 210–219. DOI: 10.13224/j.cnki.jasp.20220297.

[5] Feng Daguang, Feng Sizhe, Li Zhaoxing, et al. Comparison of Temperature Prediction Methods for Solar Greenhouses Based on Multivariate Linear Regression Models [J]. Journal of Shenyang Normal University (Natural Science Edition), 2025, 43(01): 75-81.

[6] Liu Xiaoyong, Zeng Chengbin, Liu Yun, et al. Construction of an Upper Boundary Model Based on Least-Squares Support Vector Regression [J]. Journal of South China University of Technology (Natural Science Edition), 2024, 52(12): 139-150.