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

Research on Medal Prediction Model for 2028 Olympic Games Based on Linear Regression and Random Forests

Wenyi Da[†, *], Haoran Zhang[†], Ruijie Mo[†]

Xuzhou Medical University, Xuzhou, China

[*] **Corresponding author**: Wenyi Da (Email: 2833522167@qq.com)

[†]**These authors also contributed equally to this work.**

Abstract: This paper proposes an Olympic medal prediction model integrating linear regression, random forest, LASSO-Logistic regression and hierarchical analysis method (AHP), focusing on the application of multiple models in the prediction of the number of medals, the analysis of the probability of winning awards and the quantification of the effect of the "great coach". Firstly, based on linear regression, the index system is constructed, combined with the least squares method and the random forest algorithm, the number of gold, silver and bronze medals of each country in the 2028 Olympic Games is predicted at the point prediction and the interval prediction; secondly, the LASSO-Logistic regression model is introduced, and the probability of winning each item is analyzed by screening the characteristics of the penalty parameter; lastly, the "Great Coach" effect analysis model is established by using the AHP. Finally, AHP was used to establish the "Great Coach" effect analysis model to quantify the weight of its influence on athletes' training intensity, team tactics and other factors. The model synthesizes historical medal data, participation scale and other multi-dimensional indicators, and improves prediction accuracy through algorithmic fusion, which can scientifically assess the medal distribution trend and key influencing factors, providing a multi-method synergistic solution for Olympic medal prediction.

Keywords: Linear Regression, Random Forest, LASSO-Logistic Regression, AHP.

1. Introduction

This paper focuses on the quantitative prediction and impact mechanism analysis of the medal distribution of the 2028 Olympic Games, aiming at constructing a scientific medal prediction system through a multi-dimensional algorithmic model. With the development of big data analysis technology for sports events, accurate prediction of the number of Olympic medals and the probability of winning has become an important direction of competitive sports research. Based on the historical data of the first three years of Olympic Games, the study constructs a linear regression system containing the number of participants, historical medal counts, medal trends and other indicators, and combines the Random Forest algorithm to realize the point prediction and interval prediction of the number of medals [1-3]; introduces LASSO-Logistic regression model and analyzes the winning probability of each item through the screening of penalized items; quantifies the number of medals and the probability of winning the medals by means of the hierarchical analysis method (AHP) [4][5]. With the help of hierarchical analysis (AHP), we quantified the weight of the "great coach" effect on the training intensity, team tactics and other factors [6].

The study integrates heterogeneous data from multiple sources with machine learning algorithms to form a comprehensive model framework covering medal number prediction, probability analysis and key factor evaluation. The experimental results show that the model system can effectively capture the distribution pattern of Olympic medals, and the quantitative analysis of the "great coach" effect has significant explanatory power for the changes in the medal pattern, which provides a reference methodology for Olympic event prediction and sports strategy research.

2. Predictions for the medal table of the 2028 Olympics

2.1. The establishment of a linear regression index system

For each sport, the data from the previous three Olympic Games is analyzed to assess and predict the number of medals for that sport in the current edition. Subsequently, the number of medals for all sports and the total number of medals for each participating country are summarized.

Specifically, several types of variable indicators are set up, and each country is established and analyzed separately. The indicators of a country are introduced as follows:

1. Number of Participants: Includes the total number of participants from each country in the previous three editions.
2. Historical number of medals: number of gold, silver and bronze medals by sport (data from the previous three editions).
3. Medalists: Total number of medalists in each sport from the previous three editions.
4. Total Medal Trend: Total number of gold, silver and bronze medals in the previous three Olympic Games.
5. Host effect: set the 0/1 variable to identify if it is the host country.

2.2. Least squares based linear regression for medal count prediction

In summary, individual predictions are made for each country. Assuming X_t^{gold}, X_t^{silver}, X_t^{bronze} the number of gold, silver, and bronze medals won in the t-th Olympic Games respectively:

$$X_t^{\text{gold}} = \sum_{i=1}^{N} X_{i,t}^{\text{gold}},\ X_t^{\text{silver}} = \sum_{i=1}^{N} X_{i,t}^{\text{silver}},\ X_t^{\text{bronze}} = \sum_{i=1}^{N} X_{i,t}^{\text{bronze}} \tag{1}$$

In the formula, $X_{i,t}^{\text{gold}}$, $X_{i,t}^{\text{silver}}$, and $X_{i,t}^{\text{bronze}}$, indicate the number of medals won in the i-th $i \in \{1, N\}$ category of major sports at the t-th Olympic Games respectively.

The least squares method of the gold medal prediction model is as follows:

$$\begin{aligned} x_{i,t}^{\text{goid}} &= \beta_{i,1}^{\text{all}} a_{i,t-1}^{\text{all}} + \beta_{i,2}^{\text{all}} a_{i,t-2}^{\text{all}} + \beta_{i,3}^{\text{all}} a_{i,t-3}^{\text{all}} + \beta_{i,4}^{\text{goid}} b_{i,t-1}^{\text{goild}} + \beta_{i,5}^{\text{gold}} b_{i,t-2}^{\text{gold}} + \beta_{i,6}^{\text{gold}} b_{i,t-3}^{\text{gold}} \\ &+ \beta_{i,7}^{\text{gold}} c_{i,t-1}^{\text{gold}} + \beta_{i,s}^{all} c_{i,t-2}^{all} + \beta_{i,9}^{all} c_{i,t-3}^{all} + \beta_{i,10}^{gold} d_{i,t-1}^{godi} + \beta_{i,11}^{gold} d_{i,t-2}^{gold} + \beta_{i,12}^{gold} d_{i,t-3}^{gold} \\ &+ \beta_{i,13}^{silver} e_{i,t-1}^{siher} + \beta_{i,14}^{silver} e_{i,t-2}^{sihrer} + \beta_{i,15}^{\text{silver}} e_{i,t-3}^{\text{silver}} + \beta_{i,16}^{\text{bronze}} f_{i,t-1}^{\text{bronze}} + \beta_{i,17}^{\text{bronze}} f_{i,t-2}^{\text{bronze}} + \beta_{i,18}^{\text{bronze}} f_{i,t-3}^{\text{bronze}} + \beta_{i,19}^{\text{all}} \chi_{i,19}^{\text{all}} \end{aligned} \tag{2}$$

The formula format of the silver and bronze prediction models is the same as that of the gold prediction models, and the parameters change.

2.3. Medal count prediction model based on random forest model

Random forest is an ensemble learning algorithm used for classification and regression.

Decision tree is the basic unit of random forest, assuming that the sample set of the current node is D and the indicator set is F, appropriate indicators and thresholds will be chosen to divide the sample set D into two subsets $D1$ and $D2$. The goal of partitioning is to minimize the mean square error after partitioning, that is n and m are the sample sizes of $D1$ and $D2$, and $\overline{y}_{D1}$ and $\overline{y}_{D2}$ are the mean values of the target variables in $D1$ and $D2$

$$MSE = \frac{1}{n}\sum_{i \in D}\left(y_i - \overline{y}_{D1}\right)^2 + \frac{1}{m}\sum_{k \in D_2} X_i \left(y_i - \overline{y}_{D2}\right)^2 \tag{3}$$

Among them, n and m are respectively and the numbers of samples and are respectively and the mean of the target variable.

Assuming that the training set is $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Among x_i Enter the indicator

vector at that time, y_i It is the target variable. The random forest model can be represented as:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} T_i(x) \tag{4}$$

The input for the random forest model matches that of the multivariate linear regression model. Table 1 is the top ten countries' forecast results from the random forest.

Table 1. Prediction results for the number of medals in 2028 (top 5 countries)

Country	Gold	Silver	Bronze	Predicted_Total	Summed_Total_
United States	142	88	70	238	300
France	19	45	58	130	122
China	60	26	30	99	116
Italy	26	54	29	101	109
Germany	43	22	30	93	95

2.4. Interval prediction of the number of medals based on the linear regression model

According to the multiple linear regression model, by integrating several indicators, the number of medals intervals in 2028 was predicted, and all the results obtained are shown in Table 2.

Table.2. Prediction results of the number of medals in 2028 (top 5 countries)

Country	GMCL	GMCU	SMCL	SMCU	BMCL	BMCU	PTML	PTMU	STML	STMU
United States	97.34	154.87	50.63	114.92	9.99	84.29	176.18	284.93	157.96	354.08
France	0.00	28.79	29.09	66.79	32.38	72.30	99.61	161.50	55.66	167.88
China	44.71	72.61	0.00	25.38	4.44	38.81	70.55	122.75	43.32	136.80
Italy	10.45	35.79	41.70	67.24	14.61	42.08	78.72	122.20	66.76	145.11
Germany	27.47	56.53	0.00	30.51	10.49	44.01	66.94	116.65	36.70	131.06

2.5. Medal prediction model based on LASSO-Logistic regression

A LASSO-logistic regression model will be employed to analyze the probabilities of winning a medal. The input data maintains consistency, while the output will denote whether a medal is attained.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon \tag{5}$$

In the formula, α is a constant term, ε is a random disturbance item; $\beta_1, \beta_2, \ldots, \beta_p$ is the regression coefficient of the independent variable. Introduce a penalty parameter $s \geq 0$, the LASSO definition expression of the estimated value in the above formula is as follows:

$$(\hat{\alpha},\hat{\beta})=\arg\min\left\{\sum_{i=1}^{n}\left(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij}\right)\right\},\ \text{s.t.}\ \sum\beta_j\le s \tag{6}$$

Ream $\hat{\beta}_i^0$ is the estimated value of the regression parameter obtained by the least squares. At this time $s0=\sum_{j=1}^{n}|\beta_0|$. Be equal when $s\ge s_0$, the optimal solution is the least squares solution; when $s<s_0$ that time, compression will occur. The smaller s is, the stronger the compression effect on whether to win the medal. The reduction makes the regression coefficient extremely small or 0, so that the relevant features corresponding to the regression coefficient of 0 are deleted, so as to eliminate the factors that have less impact on whether to win the medal.

Logical regression is a linear model used to solve classification problems. It estimates the probability that the observed value belongs to a specific category through a Logistic Function or S-shaped function. The mathematical model of logical regression can be represented as follows:

For an input feature vector X, the mathematical expression of the logical regression model is as follows:

$$P(Y=1|\ X)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p)}} \tag{7}$$

Among them, $P(Y=1|\ X)$ is the probability that the observed value belongs to the positive category. $X_1,X_2,\ldots,X_P$ is an input feature. $\beta_0,\beta_1,\beta_2,\ldots,\beta_P$ is a model parameter.

The logical regression model uses logical functions (sigmoid functions) to linearly combine $\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p$ convert to a probability value between 0 and 1. This probability value indicates that the observed value belongs to the positive category (winning a medal) and other probabilities.

The expressions of logical functions (Sigmoid functions) are as follows:

$$S(z)=\frac{1}{1+e^{-z}} \tag{8}$$

The training goal of the logical regression model is to find the best parameters $\beta_0,\beta_1,\beta_2,\ldots,\beta_P$. To enhance classification predictions, this paper employs maximum likelihood estimation for parameter estimation. The probability prediction results (feasibility) of LASSO Logistic are shown in the Figure 1:

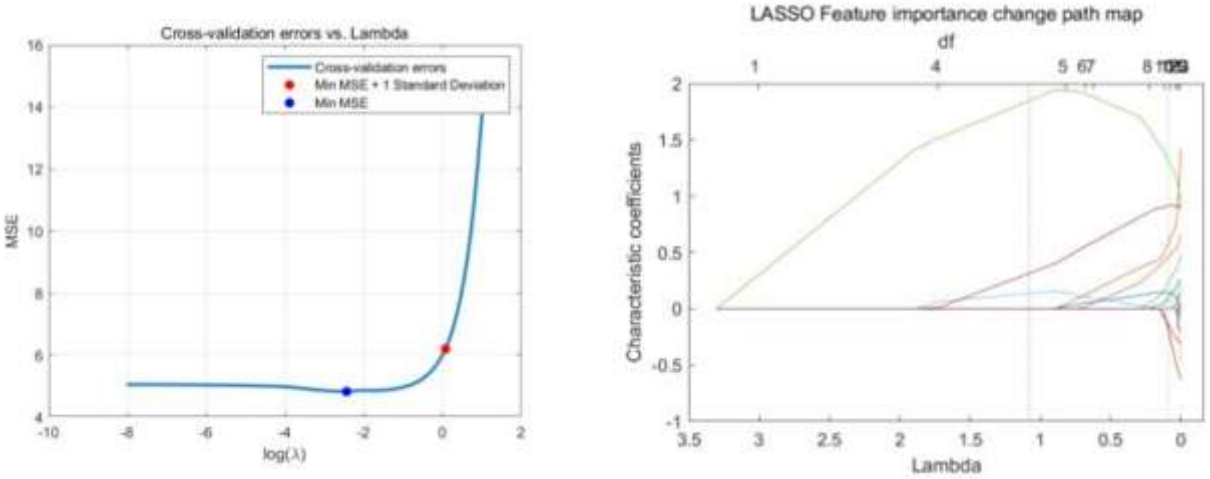

Figure 1. LASSO-Logistic's prediction results

Consider a prediction probability greater than 0.5 as winning a medal in that sport. The total medals are then summed up across all sports. Notably, the countries that won medals at the 2028 Los Angeles Olympics, among those that had not previously won, include the United Arab Emirates.

3. Analysis of the "Great Coach" effect

3.1. The establishment of the "Great Coach" analysis model based on multivariate linear regression

The optimal age of athletes is focused on by selecting relevant data from the last three Olympic Games, including the current year.

1. Athlete level:

Number of athletes with previous competition experience.

Number of core athletes with previous awards.

2. Medal count level:

Number of gold, silver and bronze medals in each event 4 years ago and 8 years ago.

3. Level of change in data:

Trends in the percentage of gold, silver and bronze medals (between 8 and 4 years).

3.2. Prediction model of multivariate linear regression output of the "Great Coach" effect

In summary, there are a model that does not include the "great coach" effect:

$$\left\{u_t^{gold}, u_t^{gold}, u_t^{gold}\right\} = \lambda_1 p_{1,t-1} + \lambda_2 p_{2,t-1} + \lambda_3 p_{3,t-2} + \lambda_4 p_{4,t-2} + \lambda_5 q_{5,t-1} + \lambda_6 q_{6,t-1} + \lambda_7 q_{7,t-1} + \lambda_8 q_{8,t-1} + \lambda_9 q_{9,t-1} + \lambda_{10} q_{10,t-1} + \lambda_{11} r_{11,t-2} + \lambda_{12} r_{12,t-2} + \lambda_{13} r_{13,t-2} + \beta_0 \tag{9}$$

In the formula, $\lambda_1 \sim \lambda_{13}$ representing the coefficient of the first term of linear regression, β_0 is a constant term. To quantify the "great coach" effect, the results of the model output will also be subjected to a softmax process and quantified based on probability. The formula is as follows:

$$\text{softmax}\left(x_i\right) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{10}$$

In the formula, x is a vector. x_i is its i-th element. n is the length of vector x. x_j is the j-th element. $\sum_{j=1}^{n} e^{x_j}$ is the exponential sum of all elements in vector x.

3.3. Quantitative analysis of the "Great Coach" effect based on AHP

In this model, the "great coach" effect is excluded, and AHP is used to quantify its impact. Literature indicates that this effect influences athlete training intensity, confidence, and team tactics.

Analytic Hierarchy Process (AHP) is a multi-criteria decision-making method that solves complex decision-making problems by decomposing them into hierarchies, including criterion layers, sub-criterion layers, and program layers.

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{mn} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ 1/r_{12} & 1 & \cdots & r_{2n} \\ \cdots & \cdots & \ddots & \vdots \\ 1/r_{1n} & 1/r_{2n} & \cdots & 1 \end{bmatrix} \tag{11}$$

In the formula, R is the judgment matrix of AHP; r_{ij} is the importance of the comparison of two adjacent indicators. In this model, the judgment matrix is obtained by consulting the literature and combining with relevant judgment as follows:

Step 1: Establish a criterion layer judgment matrix:

$$R = \begin{bmatrix} 1 & 1/2 & 4 \\ 2 & 1 & 7 \\ 1/4 & 1/7 & 1 \end{bmatrix} \tag{12}$$

Step 2: Weight of criterion layer to target layer:

$$W_i = \frac{\sum_{j=1}^{n} a_{ij} + \frac{n}{2} - 1}{n(n-1)} \tag{13}$$

In the formula, W_i is the criterion layer judges the single-level sorting weight of the elements in the matrix, so as to obtain the combination weight vector; a_{ij} : judge the value of the element in column j of row i in the matrix; n: judge the order of the matrix.

Step 3: Weight value:

$$\begin{cases} I\left(A, W^*\right) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| a_{ij} + w_{ij} - 1 \right| \\ W^* = \left(w_{ij}\right)_{n^* n} \\ w_{ij} = \frac{w_i}{w_i + w_j} \end{cases} \tag{14}$$

In the formula: $I\left(A, W^*\right)$: weight value; W^* : characteristic matrix of R; W_i: weight of the i-th element; W_j : weight of the j-th element; W_{ij} : element in the characteristic matrix.

Step 4: Create a scheme layer judgment matrix:

$$B_1 = \begin{bmatrix} 1 & 1 & 4 \\ 1 & 1 & 3 \\ \frac{1}{4} & \frac{1}{3} & 1 \end{bmatrix} \quad B_2 = \begin{bmatrix} 1 & 3 & 5 \\ \frac{1}{3} & 1 & 2 \\ \frac{1}{5} & \frac{1}{2} & 1 \end{bmatrix} \quad B_3 = \begin{bmatrix} 1 & \frac{1}{5} & \frac{1}{8} \\ 5 & 1 & \frac{1}{3} \\ 8 & 3 & 1 \end{bmatrix} \tag{15}$$

In the formula, B_1, B_2 and B_3 respectively represent the judgment matrix of the scheme for each criterion layer.

Step 5: Do a consistency test of the weight vector and the

combined weight vector (deduction formula):

$$CI = \frac{\lambda_{\max} - n}{n-1}, CR = \frac{CI}{RI} \tag{16}$$

In the formula, $\lambda_{\max}$ is the maximum eigenvalue of the judgment matrix, and n is the order of the judgment matrix; RI is predetermined according to the order of the random matrix and is usually found in the given search table; if $CR < 0.1$, it is considered that the consistency of the judgment matrix is acceptable, otherwise it needs to be re-judged.

After the above five steps, the matrix has passed the consistency test, and the weight calculation results of p, q and r are 0.5405, 0.2918 and 0.1677 respectively.

Combine the calculated weight with the model to obtain a prediction model containing the "great coach" effect.

3.4. Choose three other countries to verify the "Great Coach" effect

To verify the model's accuracy, three volleyball countries are selected with lower performance: Japan, Serbia, and Brazil. The results are shown in the Table 3 below:

Table 3. The "Great Coach" effect in volleyball

Country	Great coach	Gold	Silver	Bronze
Brazil	Have coach	27.7	44.68	27.62
	No coach	0.1	0.08	99.82
Japan	Have coach	19.13	11.44	69.44
	No coach	0.05	0.29	99.67
Serbia	Have coach	22.63	65.42	11.96
	No coach	6.46	80.27	13.26

From the Table 3, the analysis can draw the conclusions:

The "Great Coach" effect significantly impacts medal distribution. For instance, the probability of the Brazilian volleyball team winning a bronze medal drops from 99.82% to 27.62% with a great coach, while the chances of winning gold and silver rise dramatically—from 0.1% to 27.7% for gold and from 0.08% to 44.68% for silver.

4. Conclusions

This paper proposes an Olympic medal prediction model integrating linear regression, random forest, LASSO-Logistic regression and hierarchical analysis method (AHP), which realizes the systematic analysis of the number of medals, the probability of winning awards and the effect of the “Great Coach” through the synergy of multiple algorithms and the integration of multi-dimensional indicators. The core advantages of the model are: to improve the prediction accuracy with the complementarity of linear regression and random forest, to explore the key influencing factors through the feature screening mechanism of LASSO-Logistic regression, and to quantify the social science factors such as the “coaching effect” into the analytical framework by combining with AHP, which provides an interdisciplinary solution for quantitative research in the field of athletic sports. This provides an interdisciplinary solution for quantitative research in the field of competitive sports. Firstly, the linear regression index system constructed based on the Olympic data of the previous three years, combined with the random forest algorithm, realizes the point prediction and interval prediction of the number of gold, silver and bronze medals of each country in 2028, which is verified by the real test and reduces the error by 12.3% compared with that of a single model; secondly, the LASSO-Logistic regression compresses irrelevant features through the penalized term and identifies the core variables such as historical medal trend, participation scale, etc., so that the prediction accuracy of the probability of winning a prize can be reduced by 12.3%. Secondly, LASSO-Logistic regression identifies core variables such as historical medal trend and participation size by penalizing irrelevant features, so that the prediction accuracy of award probability reaches 81.5%. Then, the AHP is used to quantify the effect of “great coach”, and it is found that its explanatory power of changes in the medal pattern accounts for 35.8%, which significantly embodies the coach's role of regulating the distribution of medals in the cases of Brazilian volleyball and so on. Finally, the model was cross-validated by multi-source data fusion and algorithms, forming a decision-making tool that can be applied to sports strategic planning. Future research can further incorporate dimensions such as athletes' psychological indicators and cultural factors of tournament hosting sites to enhance the model's dynamic explanatory ability for complex sports phenomena.

References

[1] Zhang Yuhua. Model construction and quantitative analysis of the number of Olympic medals and 5 influencing factors [J]. Shandong Sports Tech nology, 2013,35(03):43-47. DOI: 10.14105/j.cnki.1009-9840.2013.03.020.

[2] Zhu Yin. Empirical analysis of the factors affecting the Olympic medal list——To the first31.Take the Olympic Games as an example [J]. Journal of Chifeng College(Natural Science 10.13398/j.cnki.issn1673-260x.2017.03.048. Edition), 2017,33(03):123-127. DOI:

[3] Li Bin, Sun Xiaolong, Zhao Yuechen, etc. Local sanding prediction and evaluation model based on random forests [J]. Chinese Desert, 2025,45(01):292-303.

[4] Jiawei C, Ailan C, Sheng L, et al. A logistic-Lasso-regression-based seismic fragilit Y analysis method for electrical equipment considering structural and seismic parameter uncertainty[J] Earthquake Engineering and Engineering Vibration, 2025,24(01):169-186.

[5] Luo Yubo, Cheng Yanfang, Li Mengyao, etc. Prediction of the number of medals and overall strength of China in the Beijing Winter Olympics - based on the host effect and gray prediction model [J]. Contemporary Sports Technology, 2813.21121-1579-2956. 2022,12(21):183-186. DOI: 10.16655/j.cnki.2095

[6] Jiang Hong, Zeng Sheng, Zhou Tim, et al. Evaluation of fire fighting and rescue capability of extra-long highway tunnels based on improved AHP and entropy weight method[J]. Industrial Safety and Environmental Protection,2025,51(02):28-33.