Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 4, No. 3, 2025

Neural Network Modeling and Multiple Linear Regression Modeling in Data Trend Prediction

Jingbo Ji[1,*], Ruiyi Wang[2], Youxu Liu[1]

[1]School of Information and Communication Engineering, Communication University of China, Beijing, China

[2]School of Economics and Management, Communication University of China, Beijing, China

[*] **Corresponding author**: Jingbo Ji (Email: 635809601@qq.com)

Abstract: This paper is based on neural network model and multiple linear regression model around data trend prediction. Firstly, feature engineering is carried out to provide rich data dimensions for the follow-up by extracting various features such as base features, contribution rate features, athlete features, etc. Second, the feed-forward neural network-based regression model was selected for medal data prediction after testing various models, and entropy weighting and hierarchical analysis were also used to construct models for specific data prediction; at the same time, multivariate linear regression models were used to evaluate the performance of the models in conjunction with K-fold cross validation. Finally, these models not only have high goodness-of-fit in data prediction, but also have high accuracy and credibility in prediction results, which is a significant advantage in data prediction.

Keywords: Neural Network Modeling, Linear Regression, Hierarchical Analysis, Monte Carlo Simulation.

1. Introduction

In this paper, with the help of neural network model[1] and multiple linear regression model[2], the prediction of the number of medals of each country and the related influencing factors are explored in depth. First, in order to accurately predict the number of medals of each country, the study constructed a regression model based on feed-forward neural network, which has excellent performance in predicting the number of medals. Secondly, for the prediction of medal breakthroughs of countries that have never won medals, entropy weighting[3] and hierarchical analysis[4] are used to establish an evaluation model. Through the comprehensive consideration of multi-faceted factors, the comprehensive score of each country is calculated to filter out the countries that may win the first medal. Finally, when studying coaching factors, a multiple linear regression model was utilized to quantify the effect of coaching effects on the number of medals by introducing new characteristic variables and coding and creating interaction terms, which were cross-validated by K-fold[5]. These algorithmic models deeply analyze Olympic medal data from multiple dimensions and provide a new perspective for Olympic medal research[6].

2. Medal data trend forecast analysis

2.1. Feature engineering

(1) Basic characteristics

The number of athletes from each country, whether a country is the host (Is_Host), the number of events (Program_Count), and the year are basic features that can influence medal predictions. The number of events in each Olympic Games is also included as a feature in the prediction model.

(2) Medal contribution of each program to each country

Since each country has its specialties, by exploring the relationship between the number of events and the number of medals, it is possible to determine which events are more important for each country to increase the number of medals or to maintain a high number of medals. This importance will be expressed as the medal contribution rate " $Medal\_Rate$" and the gold medal contribution rate "$Gold\_Rate$".

$$Medal_{Rate\,c,t} = \frac{Medal_{c,t}}{Total_t} \quad (1)$$

$$Gold_{Rate\,c,t} = \frac{Gold_{c,t}}{Total_t} \quad (2)$$

An event's contribution rate cannot be determined by a single data point and must be analyzed with historical data. To calculate the total contribution rate, we used the exponential smoothing method, which emphasizes recent data while diminishing the impact of older data. The formula for the smoothing index of the total contribution rate for a country c, a program s is:

$$TotalCR_{s,c} = \sum_{t=0}^{n-1}\left(\alpha \cdot (1-\alpha)^t \cdot Medal_{s,c,y-4t}\right) \quad (3)$$

(3) Athlete Characteristics by Country

In this paper, we generate the individual athlete weights "$Weight\_Athletes$" for predicting the number of medals by combining the year and event weights, where A_{ij} is the number of athletes in each program for country i, W_{event} is the weight for each program, and W_{year} is the year weight.

$$Weight_{Athletes_i} = \sum_j A_{ij} \cdot W_{event} \cdot W_{year} \quad (4)$$

2.2. Model Establishment

To predict the number of gold medals and the total number of medals of each country in the future Olympic Games, except for the slight difference in the pre-processing, we tested the training by multiple models, we selected the same features (LSTM and SARIMA have some differences in the selection of the features with other models, and they chose the same number of historical medals), and selected almost the same parameters for the model training, and finally, according to the MSE and R^2 values to get the visualized graphs of the corresponding different model prediction effects as in Figure 1.

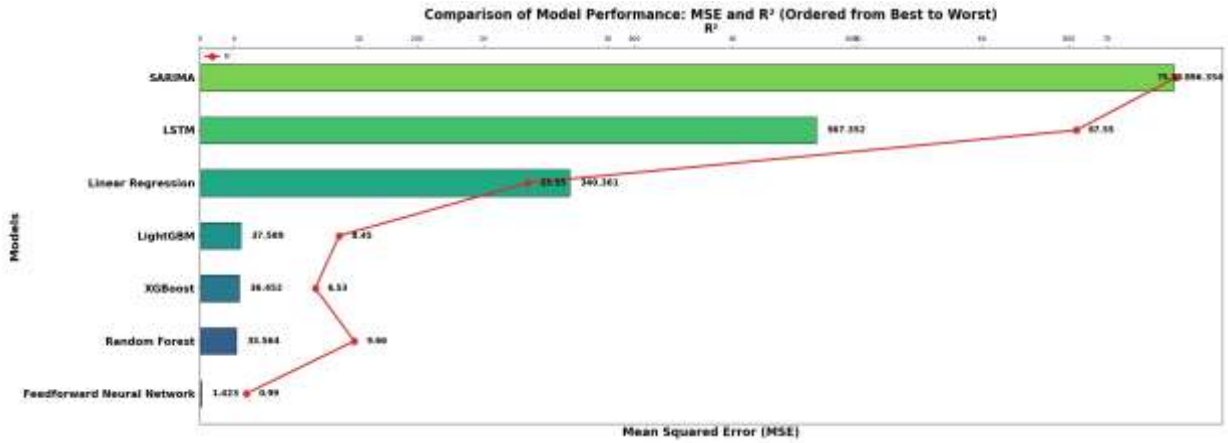

Figure 1. Comparison chart of model performance

Based on the selection of multiple models, we finally chose the regression model based on the Feedforward Neural Network (MFN) for prediction. We selected the Sequential model in Keras for training. The input of the model is the feature-engineered and preprocessed data, and the outputs are the predicted number of gold medals and the total number of medals for each country in the 2028 Olympics, respectively.

Our neural network architecture includes input, hidden, and output layers.

Input Layer: the input dimension is the number of feature variables d.

Hidden Layers: The model contains 3 hidden layers with 256, 128, and 64 neurons in each layer and uses the ReLU activation function:

$$\alpha_j = \max(0, z_j), j = 1.2 \ldots \ldots \quad (5)$$

Where z_j is the input weighted sum of the j th neuron:

$$z_j = \sum_{i=1}^{d} w_{ij} x_j + b_j \quad (6)$$

Where w_{ij} denotes the weights and b_j denotes the bias terms.

Output Layers: The output layer is a single neuron used to predict the target variables, i.e., the total number of medals and the number of gold medals. For the input features $x = [x_1, x_2, \ldots \ldots x_d]$,σ is the ReLU activation function $W^{(k)}$, and $b^{(k)}$ denote the weight matrix and bias vector of the kth layer, respectively.

The prediction formula for the entire neural network can be expressed as:

$$y_{predict} = W^{(3)} \cdot \sigma(W^{(2)} \cdot \sigma(W^{(1)} \cdot x + b^{(1)}) + b^{(2)} + b^{(3)} \quad (7)$$

The model uses the mean square error (MSE) as a loss function to measure the deviation between the predicted and true values.

For model training, we set the learning rate to 0.0001, the number of training rounds for both models for the total number of medals and gold medals to 100, and the Batch Size to 32. We will use the trained neural network models to predict the number of medals and gold medals for each country.

2.3. Analysis and Results

Finally, we get the predicted results of the total number of medals and gold medals for each country in the 2028 Los Angeles Olympics based on the above model, as shown in the table. For example, for the United States, the model predicts that it will continue to be ranked No. 1 in the coming Games, with a predicted total medal count of 131 with a prediction interval of [105,157] and a predicted gold medal count of 46 with a prediction interval of [35,56]. Projections are based on the USA's dominant programs, athlete profile, and the potential impact of the host effect.

(1) 2028 Los Angeles Olympics Medal Projections

Finally, we get the predicted results of the total number of medals and gold medals for each country in the 2028 Los Angeles Olympics based on the above model, as shown in the Table 1.

Table 1. The prediction of medal count table

Country	Gold_pred	Gold_internal	Total_pred	Total_interval
USA	45.655216	[35.00, 56.31	131.12738	[105.00, 157.26]
CHN	39.144978	[28.50, 49.79]	96.961555	[75.00, 118.92]
JPN	24.98492	[14.33, 35.64]	52.583954	[30.00, 75.17]
AUS	22.238344	[11.58, 32.89]	56.36582	[34.00, 78.73]
FRA	22.227583	[11.57, 32.89]	68.78807	[46.00, 91.58]
GBR	17.825064	[7.17, 28.48]	66.61704	[44.00, 89.23]
ITA	15.364998	[4.71, 26.02]	50.658813	[28.00, 73.32]
NZL	10.152424	[0.00, 20.81]	23.086222	[0.00, 46.17]

The model prediction effect was also evaluated along with the prediction results, and the evaluation results are shown in the Table 2. The R^2 of both models reached 0.99, the MSE of the model predicting the number of gold medals was only 0.164, and the MSE of the model predicting the total number of medals was only 1.23, the model fitting effect is very good, and the credibility of the prediction results in 2028 is high.

Table 2. Model Effect Table

	MSE	R^2
Gold Model	0.1646425	0.99744
Total Model	1.23044	0.99725

(2) Changes in award performance by country

Figure 2 shows bar charts of the progress and regression data for each country. By comparing the medal counts at the 2024 Paris Olympics, changes in the number of medals and gold medals for each country can be observed. Projections show that Japan, Canada, Italy, Romania, and France are expected to see increases in total medals by 10-12. The United States, Spain, and Finland are projected to see the largest improvements in gold medals, with increases of 4-5. The United States, which added five new sports and is hosting the 2028 Los Angeles Olympics, could gain more medals due to the host advantage.

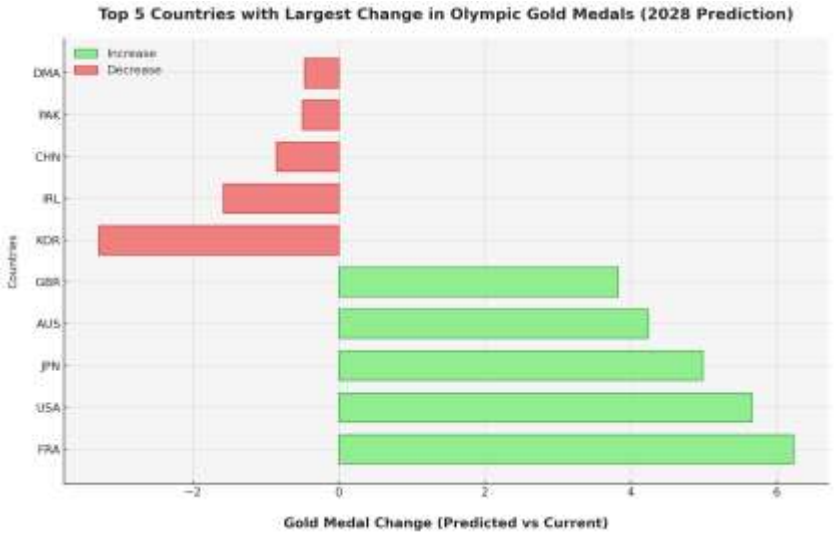

Figure 2. Bar chart of progress and regression data for various countries

However, some countries are expected to experience significant regression. Hong Kong, China, Israel, and Ecuador may see a reduction of 1-2 total medals, while China, South Korea, and New Zealand are likely to lose 1-4 gold medals, with China's decline of 4 golds possibly linked to changes in sports programs or individual athlete performance.

2.4. Getting the first medal in history National Prediction

To predict whether countries that have never won an Olympic medal will win their first medal at the 2028 Los Angeles Olympics, we developed an evaluation model using entropy weighting and hierarchical analysis. This model filters out countries that have not won medals in past Olympics and identifies ten countries most likely to win their first medal in 2028 based on their evaluation scores.

2.4.1. Model Preparation

We have extracted from the data provided the countries that have never won a medal in their history, as well as the athletes from those countries and the programs they participated in. The dataset was then merged and consolidated to include the name of the program, the number of programs participated in, the name of the country, the personal information of the athlete, and whether they won a medal. To reflect the importance of athletes' performance in recent years, year weights were introduced. Where $Year_{max}$ and $Year_{min}$ are the earliest and latest years in the data, respectively.

$$W_{year} = \frac{(Year - Year_{min} + 1)}{(Year_{max} - Year_{min} + 1)} \tag{8}$$

Also, in order to quantify the importance of the entries, we used the entropy weighting method to calculate the weights of each project W_{event} The mathematical definition of the project weights is as follows, where n_{ij} is the number of entries made by the country i on the project j and e_j is the entropy value of the project j.

$$p_{ij} = \frac{n_{ij}}{\sum_j n_{ij}} \tag{9}$$

$$e_j = -\frac{1}{\ln(m)} \sum_i p_{ij} \ln(p_{ij}) \tag{10}$$

$$W_{event} = \frac{1 - e_j}{\sum_j 1 - e_j} \tag{11}$$

Based on obtaining the year weights and program weights, the gender of athletes, participating programs, and year weights of each country are weighted and counted to obtain the combined program weights and combined athlete weights. For country i, the formula of the weighting index is as follows, where D_{ij} is the number of participating programs of country i and A_{ij} is the number of participating athletes in each program of country i.

$$Weight_{Disciplines_i} = \sum_j D_{ij} \cdot W_{year} \cdot W_{event} \tag{12}$$

$$Weight_{Athletes_i} = \sum_j A_{ij} \cdot W_{year} \cdot W_{event} \tag{13}$$

To maximize the comparability of the indicators, we have normalized the program composite weights $Weight\_Disciplines$ and the athlete composite weights $Weight\_Athletes$ so that the data can be on the same order of magnitude.

2.4.2. Multi-Feature Model Establishment

Subsequently, we constructed the judgment matrix and calculated the maximum eigenvalues and eigenvectors to get the weights of each attribute using the Analytic Hierarchy Process (AHP), which is a multi-attribute decision-making method that breaks down a complex problem into different levels and indicators by constructing a hierarchical structure and compares them two-by-two to derive the weights of each indicator and a composite score. In this model, we comprehensively consider factors such as the number of participants, the distribution of participation programs, gender characteristics historical participation, etc., and finally

calculate and rank the scores of each country. By comparing the relative importance of different attributes two by two, we define the judgment matrix A as:

$$A = \begin{bmatrix} 1 & 3 & 2 & 4 \\ 1/3 & 1 & 1/2 & 2 \\ 1/2 & 2 & 1 & 3 \\ 1/4 & 1/2 & 1/3 & 1 \end{bmatrix} \quad (14)$$

The maximum eigenvalue and the corresponding eigenvector of the judgment matrix A is then calculated by eigenvalue decomposition, and the weights of each attribute are obtained after normalization w_i, and the composite scores of each country are calculated based on the weighted sum of the attribute weights and the indicator values $Score_i$.

$$Score_i = \sum_j w_j \cdot X_{ij} \quad (15)$$

2.4.3. Analysis and Results

We have ranked the countries based on their combined scores and have come up with the top 8 countries and their scores to win their first-ever medals at the 2028 Olympics in Los Angeles. The exact results are shown in Table 3.

Table 3. The Comprehensive scores of countries table

NOC	Score	Probability
ANG	2.246085	9.67%
HON	2.03669	8.30%
MLI	0.756739	3.89%
GUI	0.629264	3.05%
SAM	0.522646	2.35%
CRC	0.368567	1.94%
BAR	0.272426	1.61%
MLT	0.244066	1.52%

We then evaluated the performance of the countries that achieved their first-ever medal in the last two decades at the current Olympics and also substituted it into our model. The model was tested based on the t-distribution of hypotheses, and after applying small perturbations to the judgment matrix several times, the 95% confidence interval for the highest scores was calculated to be [1.133514,3.536126], and the interval for the countries that had won medals was [3.373515,5.055129], and based on the range of values in the two intervals, we could get the highest probability of winning a medal of 9.67%. The probability is 9.67%. this is illustrated in Figure 3 below. Therefore, we can see that the probability that a country will make a breakthrough in the next Olympic Games is not high.

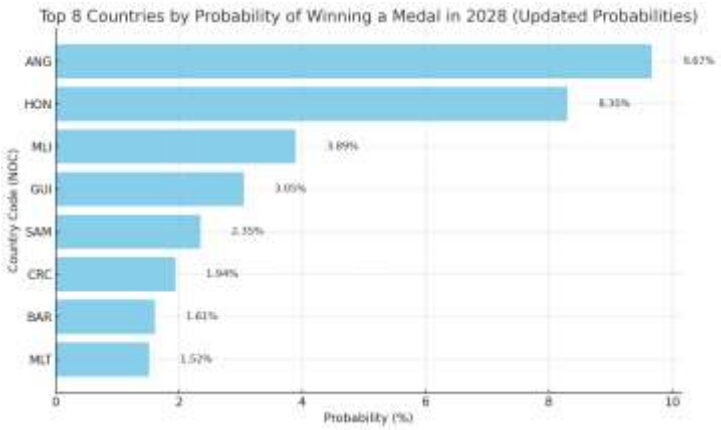

Figure 3. Probability chart of countries that may win their first medal

3. Analysis of coaching factors

3.1. Model Preparation

We add a new feature variable "Great Coach", which is 1 if there is a coach who meets the conditions of Great Coach in this program in this Olympics, and 0 if the opposite is true.

One-hot coding was performed on the categorical variables (NOC and Event), and a new interaction term was created to reveal the effect of the interaction of the characteristics "years of continuous coaching" and "presence of a great coach" on the number of medals.

$$Interaction = Years\ of\ Continue\ Coaching \times Great\ Coach \quad (16)$$

Then all the features that need to be put into the model are organized.

3.2. Model Establishment

We chose to use a multiple linear regression model to measure the magnitude of the "great coach effect" on the number of medals and gold medals. The formula for the linear regression model is as follows:

$$y_{predict} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (17)$$

Where $y_{predict}$ is the predicted target variable (total medals and gold medals), x_i is the feature variable, β_0 is the intercept term, β_i corresponds to the regression coefficients of the feature variables, which indicates the degree of influence of each feature on the target variable, and ε is the error term.

To evaluate the model performance, we used K-fold cross-validation. K-fold cross-validation divides the data into multiple subsets, selects one subset at a time as the validation set and the others as the training set, and repeats it for K times, and then evaluates the model performance by

summarizing the results. We tried different values of K ($[3, 5, 7, 10]$), and by comparing the average mean square error (MSE) under each value of K, we finally chose $K = 7$ as the optimal value. After fitting the linear regression model, the regression coefficients obtained for each feature allow us to analyze the degree of influence of each feature on the target variable. For example, $\beta_{greatcoach}$ denotes the average effect of the presence of a great coach on the number of medals, and $\beta_{interaction}$ denotes the effect of the interaction between the number of years of continuous coaching and the presence of a great coach on the number of medals.

3.3. Analysis and Results

By inputting the features into a linear regression model, we measure the effect of great coaching. The results from the test set indicate that the presence of great coaches significantly impacts both the total number of medals and the number of gold medals for certain countries and sports.

When great coaching is applied, Great Britain's (GBR) medal count in Cycling increases from 5.08 to 6.97, a rise of 1.89 medals. China's (CHN) Gymnastics medal count grows from 7.81 to 10.62, an increase of approximately 2.81 medals. The increase in gold medals follows a similar trend to the total medal count. However, in the United States (USA) Swimming event, the increase is modest, rising by only 1.66 medals.

4. Conclusions

In this study, the number of Olympic medals was comprehensively predicted and deeply analyzed by constructing various algorithmic models such as neural network model and multiple linear regression model. The regression model based on feed-forward neural network has excellent performance in predicting the number of medals of each country, with the R^2 of the model as high as 0.99, and the MSEs of predicting the number of gold medals and the total number of medals are only 0.164 and 1.23, respectively, which accurately predicts the number of medals of the United States, China and other countries as well as their rankings trends. Then, the model constructed by using entropy weighting and hierarchical analysis effectively screens out the countries most likely to win medals for the first time, providing a reference for the diversification of the Olympic medal pattern. The multiple linear regression model, on the other hand, quantified the coaching factor and found that its influence on the number of medals of different countries and events varied significantly. Overall, this study provides valuable methods and ideas for Olympic medal count prediction and sports event research.

References

[1] Shao Jiapeng. Research on time series data prediction based on deep neural network[D]. Jilin Institute of Chemical Technology, 2024.DOI: 10.27911/d.cnki.ghjgx.2024.000092.

[2] Zhang, W.R. Research on multiple regression big data prediction method based on Hadoop[D]. Dalian Jiaotong University,2016.

[3] Luo Fan, Jiang Yu. An attribute approximation algorithm based on information entropy weighting[J]. Computer Application Research,2024,41(04): 1047-1051.DOI: 10.19734/j.issn.1001-3695.2023.07.0366.

[4] Huang J, Yuan J, Cui Hastiness Long, et al. Research on the methodology of equipment data assessment model based on the combination of hierarchical analysis and fuzzy evaluation methods[J]. Network Security and Data Governance,2024,43(11):43-49+55.DOI: 10.19358/j.issn.2097-1788. 2024.11.008.

[5] Luo Aodan, Huang Zhensheng. K-fold cross validation criterion model averaging method and its application[J]. Journal of Hefei Normal College,2024,42(03):40-43.

[6] Wang F. Prediction of medal performance in 2020 Olympic Games based on neural network[J]. Statistics and Decision Making,2019,35(05): 89-91. DOI: 10.13546/j.cnki.tjyjc.2019.05.019.