World Journal of Management Science
ISSN: 2959-9636
DOI: https://doi.org/10.61784/wms3058

PREDICTION STUDY OF 2028 OLYMPIC MEDAL TABLE BASED ON WEIGHTED FUSION MODELING

QianFeng Jin*, RuiXin Yao
School of Chemistry and Chemical Engineering, North University of China, Taiyuan 030051, Shanxi, China.
Corresponding Author: QianFeng Jin, Email: 17200616170@163.com

Abstract: The purpose of this paper is to construct a weighted fusion model to predict the 2028 Olympic medal table with greater accuracy. Firstly, the data is cleansed and a relevant evaluation system is established using the K-Means clustering method. Then, a weighted fusion model integrated with a regression model and a time series model is adopted to predict the medal table of the 2028 Olympic Games. The results are as follows: the United United States is predicted to increase both the number of gold medals and the total number of medals, due to the home field advantage; China's number of gold medals may decrease, due to the abolition of the dominant events; and Japan's number of medals is expected to decrease, due to the loss of home field advantage. The study demonstrates the efficacy of the model in predicting the medal table, thereby providing a reference for countries to formulate their participation strategies and optimise the allocation of resources.
Keywords: Weighted fusion model; K-Means clustering; Olympic Games; Medal prediction

1 INTRODUCTION

The Olympic medal table is of significant global interest, as it serves as a barometer for the sporting prowess of nations worldwide. It offers insight into the effectiveness of national sports policies, the calibre of athlete training programmes, and the allocation of national resources towards sports. The number of gold medals attained is frequently regarded as a metric for evaluating a nation's sporting achievements. Given the growing importance countries attribute to sports events and the influence of a number of factors, the Olympic medal table has become more difficult to predict, and accurate prediction of the Olympic medal standings is therefore important.

In previous studies, scholars have utilised various methodologies to predict the medals. Based on the interpretable machine perspective, Shi Huimin et al. employed a random forest model to assess the predictability of medals in different sports, thereby demonstrating the feasibility of Olympic medal prediction[1]. Tian Hui et al. selected the host country of the 14th-23rd Winter Olympics as the research object, systematically studying the host country's dominant characteristics, such as the number of gold medals The ranking in the medal table, and the number of medals in different sports sub-items. The advantageous characteristics of the host country were then studied from the perspective of national competitive sports development policies and funding inputs. The dominant effect of the host country of the Winter Olympics was analysed, and logistic regression models were established to predict the number of medals of our athletes in the 202 2 Beijing Winter Olympics will be the best in the history of Winter Olympics participation[2]. Wang Fang used the non-linear method of neural network to fit and predict the per capita GDP data, and based on the prediction model proposed by Bernard and Busse, we made a prediction of the medals of the 2020 Tokyo Olympics, and then made a prediction for the 2020 Tokyo Olympics. 2020 Tokyo Olympic Games medals were predicted[3]. However, these models are not without their limitations. Neural networks require a substantial amount of data and are susceptible to quality issues. They also have a high computational demand and are less robust to outliers. PSO and multiple linear regression analysis models become more complex and may introduce greater uncertainty, affecting the stability and reliability of the model. Logistic regression Logistic regression models are sensitive to linearly divisible data and have a weak ability to deal with feature interactions. Time series analysis is suitable for simple, stable and cyclical data. In order to more fully and efficiently predict the Olympiad scores, it is necessary to comprehensively consider the phasic factors and use a combination of multiple analysis methods.

The objective of this paper is to accurately predict the medal list of the 2028 Olympic Games by combining a weighted fusion model integrated with a regression model and a time series model. This will take into account historical medal data, athletes' situations, and Olympic events, as well as analysing the related factors in depth. The aim is to provide references for countries to formulate their participation strategies and optimise the allocation of resources.

2 MATERIALS AND RESEARCH METHODS

2.1 Data Preprocessing

Data from the official International Olympic Committee website used in this article.

The initial phase of the study involves detecting missing values and visualising the missing values in the dataset using heatmaps. In the event that the dataset contains missing values, the heatmap will visibly highlight the missing areas, thus helping to identify problematic parts of the data. If duplicate records are found, appropriate measures such as deletion or merging will be taken to ensure the accuracy and reliability of the data.

2.2 Research Methods

In this paper, three models are proposed for predicting the number of gold, silver and bronze medals won by countries with well-developed sports systems and countries with developing sports, respectively. For time series variables, such as the number of medals in previous years, it is customary to employ time series models. For continuous and sub-type variables, including quantitative ratings of athletes' abilities and the number of gold, silver and bronze medals, regression models are utilised to predict the number of medals in the current Olympics. Consequently, a weighted fusion model combining the regression model and the time series model was employed in constructing the prediction model for the number of medals in future Olympic Games, as illustrated in Figure 1 Flow chart.

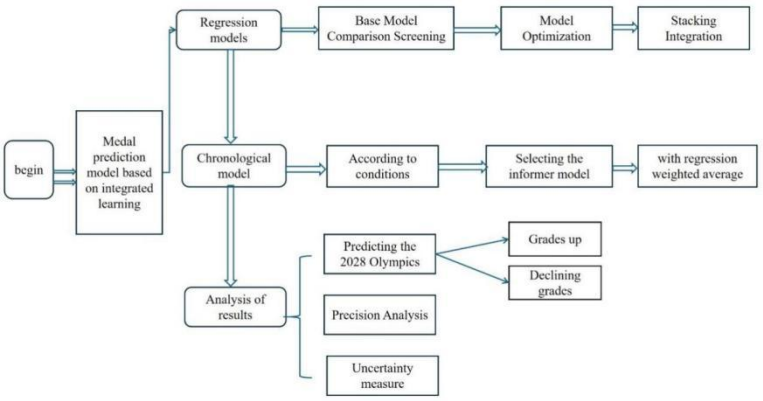

Figure 1 Flow Chart

In the context of selecting a nonlinear model, the tree model has been observed to exhibit an overfitting problem on the dataset due to the tree model itself. To address this challenge, this paper proposes a stacking method for integrating multiple tree models. The proposed model involves the utilisation of several tree models, with a meta-model being implemented to merge their predictions. Examples of such models include the GBDT model[4], XG-Boost model[5], Cat-Boost model [6]and Extra Trees model[7]. This integration approach effectively solves the overfitting problem inherent in individual tree models, thus improving the generalisation and performance of the models on unseen data, as illustrated in Table 1.

Table 1 Plot of Optimal Parameters for Each Base Model

Mould	GBDT	ExtraTrees	XGBoost	Cat-boost
optimal parameter 1	learning rate=0.1	profundity=14	learning rate=0.05	profundity=6
optimal parameter 2	profundity=10	Node ratio=0.7	profundity=10	learning rate=0.1
optimal parameter 3	Number of decision trees=200	Number of decision trees=300	Number of decision trees=200	Number of iterations=200

3 MODEL CONSTRUCTION

Initially, the number of medals won by each country at each Olympic Games was enumerated. Thereafter, these data were visually represented and analysed, as demonstrated in Figure 2.

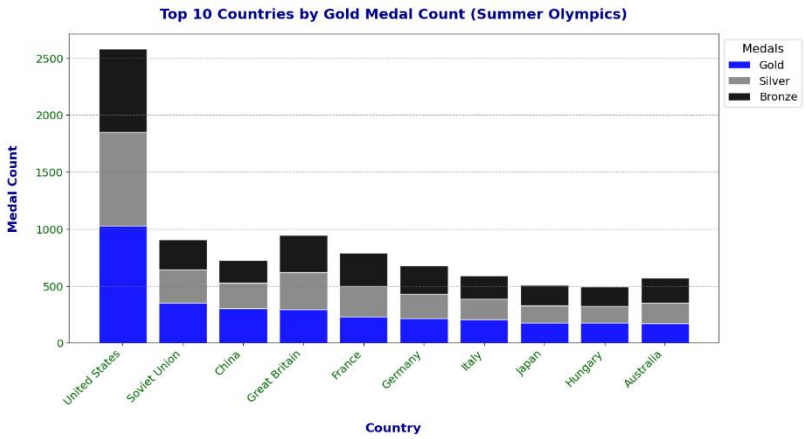

Figure 2 Top 10 Countries in Terms of Gold Medals

As demonstrated in the accompanying visualisation chart, there exists a considerable disparity in the level of sports development between nations, a discrepancy that has the potential to exert an adverse influence on the subsequent prediction model if not effectively addressed. Data discrepancies have the capacity to engender model prediction bias, resulting in imbalanced data comparisons between sports powerhouses and developing countries. Furthermore, such discrepancies can lead to the distortion of training data, an occurrence that has the capacity to compromise the accuracy of the model. Additionally, the presence of oscillatory behaviour in the model can impede its ability to converge stably during the training process. This, in turn, gives rise to a flattening of the objective function, changes in the direction of the gradient, and unstable results, which in turn adversely affects prediction accuracy and generalizability.

In order to address these issues and to enhance the predictive validity of the subsequent model, this paper proposes a categorisation of countries according to their level of sports development. Specifically, the paper proposes a two-category classification system, namely 'mature sports system countries' and 'emerging sports countries'. Mature sports system countries refer to those countries that already have a perfect sports system, and have a high level of sports competition and stable medal output, such as the United States, China and so on. These countries have accumulated substantial resources and advantages over time in the domains of training, infrastructure development, and financial investment in sports programs. In contrast, emerging sports countries are those that are in the process of establishing or progressively enhancing their sports systems. Although their sports resources and infrastructure may not be optimal, these countries have demonstrated potential in certain sports programs and are anticipated to enhance their sports standards over time.

In order to realise this classification, this paper adopts the K-Means clustering algorithm. The number of gold, silver and bronze medals won by each country in different Olympic years is input as the features in the data set. The clustering algorithm then divides these countries into different groups. By employing this method, this paper can automatically identify which countries have strong sports competitiveness and which countries are gradually developing their sports potential. The clustering results are shown in Figure 3.

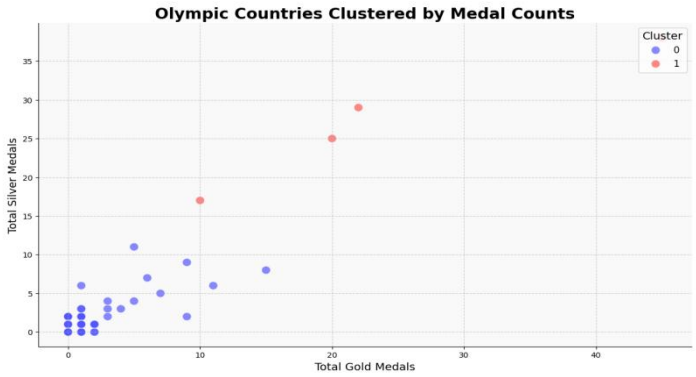

Figure 3 Graph of Clustering Results

The objective of this paper is to use clustering analysis to identify two categories of countries: those with a mature sports system and those with an emerging sports system. The results of this categorisation will facilitate the subsequent construction of a more accurate prediction model. The model will focus on the strengths and experiences of the former category in multiple sports, while the model will pay more attention to the potential and development trends of the latter in specific sports. The development of distinct prediction models for these two categories is expected to mitigate the impact of potential differences in data, thereby enhancing the accuracy and reliability of the model. This study will also establish an evaluation system to quantify the abilities of athletes.

3.1 Establishment of Evaluation Indicators

3.1.1 Comprehensive national competitiveness

(1) Obtaining the athletes of the target country who are competing in a particular sport.
(2) Calculation of athlete scores
(3) We begin by clarifying the rules for allocating points, as shown in Table 2.

Table 2 Points Allocation Table

Awards	Gold	Silver	Bronze	None
Score	10	6	3	1

In this paper, time-decreasing weights are assigned to the athletes' historical performances, and decreasing weight factors are set for each Olympics. The kth Olympics is weighted as w_k, $w_k = 1 - 0.2(k-1)$, $k = \{1,2,3,4,5\}$and S_{t_k,A_i} is the athletes' scores in the kth Olympics. The following formula is used to calculate the athletes' total weighted scores P_{A_i}:

$$P_{A_i} = \sum_{i=1}^{5} S_{t_k,A_i} \cdot w_k \# \quad (1)$$

(4) Country's total score in a given minor event: assuming that all the country's athletes (NOC) participating in minor event E are A_i, the country's total score $T_{NOC,E}$ in this minor event is the sum of the scores of all the participating athletes.

$$T_{NOC,E} = \sum_{i=1}^{n} P_{A_i} \# \quad (2)$$

(5) Normalisation of overall scores: In order to quantify the competitiveness of each national team in the sub-event, we normalised the scores in each sub-event.

$$N_{NOC,E} = \frac{T_{NOC,E}}{\max T_{NOC',E}} \# \quad (3)$$

(6) The normalised score $N_{NOC,E}$ takes values in the range [0,1], with higher normalised scores indicating greater competitiveness for that country in that subprogramme.

(7) Calculation of the country's combined competitiveness in all small projects

$$C_{NOC} = \sum_{i=1}^{m} N_{NOC,E} \# \quad (4)$$

3.1.2 Country highlights competitiveness

A country's outstanding competitiveness is an important indicator for assessing its excellence and ability to win gold, silver and bronze medals in small events. By analysing the historical results of the athletes competing in each minor event, the competitiveness of each country for gold, silver and bronze medals can be calculated and provide data support for subsequent forecasting and analysis, and the steps for constructing the indicator are as follows:

(1) Access to all athletes competing in a specific minor sport in a target year.

(2) Calculate each athlete's score using the Historical Performance Score formula.

$$P_{A_i} = \sum_{i=1}^{5} S_{t_k,A_i} \cdot w_k \# \quad (5)$$

(3) Identify the top three scoring athletes in a minor event: After calculating the scores of all participating athletes, rank the athletes according to the size of their scores, and write down the results of the ranking as:

$$P_{A(1)} \geq P_{A(2)} \geq P_{A(3)} \geq \cdots P_{A(n)} \# \quad (6)$$

Calculation of Gold, Silver and Bronze Medal Indicators For each sub-item E, the countries that will receive gold, silver and bronze medals will be determined by ranking their scores in the following order；

$$\begin{cases} G_{NOC,E} = \begin{cases} 1, & if\ NOC = NOC_{(1)} \\ 0, & otherwise \end{cases} \\ S_{NOC,E} = \begin{cases} 1, & if\ NOC = NOC_{(2)} \\ 0, & otherwise \end{cases} \# \\ B_{NOC,E} = \begin{cases} 1, & if\ NOC = NOC_{(3)} \\ 0, & otherwise \end{cases} \end{cases} \quad (7)$$

(4) Aggregate gold, silver and bronze medal metrics for each sub-event: For a given country (NOC), sum the gold, silver and bronze medal metrics for all sub-events to obtain the total gold, silver and bronze medal metrics for that country.

$$\begin{cases} G_{total,NOC} = \sum_{i=1}^{m} G_{NOC,E_i} \\ S_{total,\ NOC} = \sum_{i=1}^{m} S_{NOC,E_i} \# \\ B_{total,NOC} = \sum_{i=1}^{m} B_{NOC,E_i} \end{cases} \quad (8)$$

(8) Evaluating the outstanding competitiveness of countries: The $G_{total,NOC}$, $S_{total,\ NOC}$, $B_{total,NOC}$ calculated above can be used to systematically evaluate the outstanding competitiveness of countries. These indicators reflect the country's performance at the level of elite athletes and also provide a quantitative basis for the subsequent prediction of the Olympic gold, silver and bronze medal lists.

3.2 Fusion of Regression and Time Series Models

Regression models and time series models[8] have very different architectures, but both have their own advantages in solving specific problems. Although they give close results, an integrated model with better performance can be obtained by model fusion. For the fusion, this paper uses a simple weighted average method and uses sklearn for auto-tuning the coefficients to obtain the optimal fused model. The final integrated model combines the strengths of both the tree model and the linear regression model and can therefore achieve a higher performance score than the individual models. Although the score of the integrated tree model is higher than that of the linear regression model, the structure of the optimal output allows the two models to complement each other's strengths, further improving the overall performance. Figure 4 illustrates the fusion architecture of the integrated model:

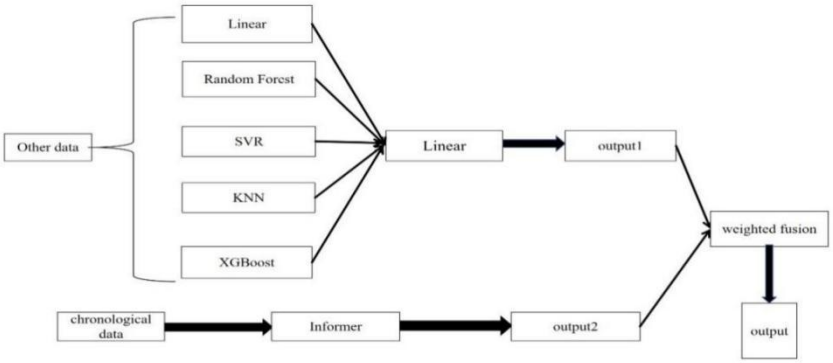

Figure 4 Weighted Fusion Model Structure

After the weighted fusion of the models, this paper validates their performance and finds that their performance on the dataset will indeed be stronger than any single model, and can effectively mitigate the overfitting phenomenon, and more accurate predictions are obtained on the validation set, and the performance graphs of their fusion models are shown in Table 3 :

Table 3 Fusion Model Performance Graph

	MSE	MAE
fusion model	3.875	0.9680

3.3 Calculate the Uncertainty

In the stacking integration model, in this paper we can calculate the prediction uncertainty of the model by analysing the prediction results of each base learner. Specifically, in this paper, we can obtain the prediction values of each base learner for the test set and calculate the variance of these predictions. With the variance, in this paper we can derive the standard deviation and then calculate the confidence interval for each prediction. In this paper, we choose the 95% confidence interval: $Y \pm 1{,}96\delta$, and draw a visualisation to show the stability and instability of the prediction values, as shown in Figure 5 :

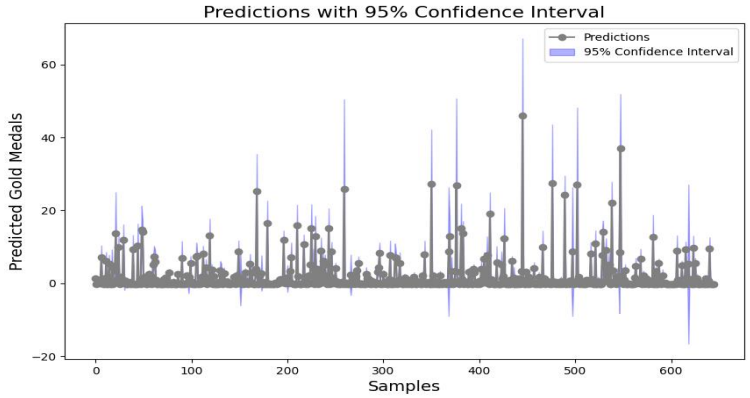

Figure 5 Visualization of Predicted Values

The model's predictions for some of the samples have large uncertainties, as evidenced by large standard deviations and even an excessively wide range of confidence intervals for some of the samples, with negative values and unusually large upper bounds. This may indicate that the model is unstable with certain data, possibly due to noise, overfitting or problems with the data itself.

To improve the stability of the model, consider analysing the characteristics of the high uncertainty samples, adding regularisation to prevent overfitting, and tuning the hyperparameters of the model. In addition, using more robust models and increasing the granularity of data processing will help to reduce prediction uncertainty and increase model confidence.

3.4 Predicted Results of the 2028 Olympics Medal Table

By analysing information on the 2028 Summer Olympics in Los Angeles, the Los Angeles Organising Committee has decided to drop traditional sports such as weightlifting and boxing and to add five new major sports: baseball and softball, racquet tennis, cricket, squash and flag football. Based on these changes, and assuming that the list of

participants does not change significantly from 2024, the predictive model in this paper gives only a large bonus to the performance of the United States, with no significant change in the number of medals for other countries. Incorporating this information into the previously constructed medal prediction model, the following conclusions can be drawn - the United States, as the host, will benefit from home advantage and its performance at the 2028 Summer Olympics in Los Angeles will be significantly improved, with the number of gold medals and total medals expected to increase significantly. In contrast, the performances of China and Japan are likely to decline. While China is a strong competitor in many events, the elimination of traditionally dominant Chinese sports such as weightlifting and boxing in 2028 could lead to a decline in its gold medal tally, while Japan, which shone at the 2024 Games, is expected to see a decline in its medal tally as a participant in the 2028 Olympic Games due to the loss of its home advantage, as shown in the results in Table 4 .

Table 4 Olympic Medal Prediction Table

NOC	Gold medal	Total medals
United States	51	151
China	31	86
Japan	19	45
Australia	18	52
France	17	62
Netherlands	16	38
Great Britain	14	64
South Korea	11	35

According to this forecast, the United States is expected to improve its performance at the Olympics due to its strength and home advantage, while China and Japan are likely to see their medal tally decline. Such a change highlights the importance of home advantage and the far-reaching impact of changes to the programme of events on the performance of different countries. In addition, the addition of new events may provide new opportunities for other countries, particularly the United States, to capitalise on their local culture and spectator support to excel in these new events.

4 CONCLUSION

In this paper, the participating countries are categorized by cluster analysis, and then the weighted fusion model is successfully constructed to predict the medal table of the 2028 Olympic Games by constructing feature engineering, establishing the relevant evaluation index system, and integrating time series and regression models. There are many factors that affect the performance of the Olympic Games, in addition to the pattern that can be found from historical data, other factors such as comprehensive national power and host effect should be comprehensively considered, and intelligent methods such as neural networks should be used to make more perfect predictions, and more factors will be incorporated in the future.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic medals be predicted? --An interpretable machine learning perspective. Journal of Shanghai University of Physical Education and Sports, 2024, 48(04): 26-36.
[2] Tian Hui, He Yiman, Wang Min, et al. Medal Prediction and Participation Strategy of Chinese Athletes in the Beijing 2022 Olympic Winter Games: Based on the analysis of the Olympic home advantage effect. Sports Science, 2021, 41(02): 3-13+22.
[3] Wang F. Prediction of medal performance in 2020 Olympic Games based on neural network. Statistics and decision making, 2019, 35(05): 89-91.
[4] Hanbo Zhang, Ji Yang, Wenlong Jing, et al. Research on downscaling method of precipitation data by multiple characterisation factors combined with GBDT. Chinese Environmental Science, 2023, 43(04): 1867-1882.
[5] MU Yanzi, ZHAO Zhifei, WU Wei. Research on sea surface small target detection method based on cost-sensitive learning DBN-XGBoost [J/OL]. Systems Engineering and Electronics, 2025, 1-11. http://kns.cnki.net/kcms/detail/11.2422.tn.20250415.1514.003.html.
[6] XING Zhikai, LIU Yongbao, WANG Qiang, et al. Application of Cat Boost Algorithm in Shipboard Bearing Fault Diagnosis. Ship Science and Technology, 2022, 44 (23): 117-122.
[7] E Okoro E, Obomanu T, E Sanni S, et al. Application of artificial intelligence in predicting the dynamics of bottom hole pressure for under-balanced drilling: Extra tree compared with feed forward neural network model. Petroleum, 2022, (02): 227-236.
[8] Li Muying. A time series analysis model based on multi-source data. Information Technology, 2025, (01): 112-118+125.