World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3049

MEDAL PREDICTION BASED ON REGRESSION MODELS

XinLei Wang[1*], ZiHan Gao[2], ZiYe Chen[3]
[1]*Overseas Chinese College, Capital University of Economics and Business, Beijing 100070, China.*
[2]*Accounting School, Capital University of Economics and Business, Beijing 100070, China.*
[3]*School of Artificial Intelligence, Capital University of Economics and Business, Beijing 100070, China.*
Corresponding Author: XinLei Wang, Email: 18516835518@163.com

Abstract: This research focuses on the prediction of the 2028 Los Angeles Olympic medal tables, constructing a predictive model based on multiple linear regression. Through systematic quantitative analysis of individual athlete performance data and national-level sports development factors, a comprehensive medal count prediction system has been constructed. This study thoroughly considers key aspects of athlete performance, including medal-winning records, number of participations, and recent competitive performance. At the same time, crucial variables at the national level, such as home-field advantage and historical performance, are incorporated. By integrating technical means such as time-decay functions, error term settings, and normalization processing, the accuracy and stability of the prediction model have been significantly enhanced, systematically addressing the 2028 Los Angeles Olympic medal count. By combining rigorous statistical validation with complex predictive modeling, the study demonstrates the significant advantages of the constructed model in terms of predictive effectiveness and analytical comprehensiveness, revealing its universal applicability across various sports scenarios. Finally, the research integrates its findings to provide decision-making references for national Olympic committees, facilitating strategic planning for sports events and optimal allocation of resources.
Keywords: Medal prediction; Multiple linear regression; Resource allocation; Home-field advantage

1 INTRODUCTION

The Olympic Games, as the world's top-level sports event, the medal table is not only a key indicator for measuring the sports strength of various countries but also a significant symbol of national honor and national pride [1]. The competition for the medal table has always been a crucial part of the Olympics, with governments and sports organizations investing substantial resources to achieve excellent results.

Zhou Xiaobo et al. analyzed the correlation between population quality, political institutions, and Olympic performance [2]. Wen Jing et al. employed multiple methods including literature analysis and mathematical statistics to predict the number of medals won by the Chinese team in the Winter Olympics [3]. Shi Huimin et al. employed a random forest model to evaluate the predictability of medals in different sports, thereby demonstrating the feasibility of Olympic medal prediction [4]. Yuan Junjie conducted medal prediction based on big data models using historical award-winning data from past events [5]. However, these prediction methods often require massive data and rely heavily on historical medal statistics, failing to fully account for key factors such as changes in Olympic event settings, host-country advantages, and the "great coach" effect. As a result, their prediction accuracy is significantly affected.

This study aims to establish a more comprehensive and accurate Olympic medal table prediction model that comprehensively considers multiple factors, including historical medal data, Olympic event settings, host-country advantages, and the "great coach" effect. By deeply analyzing the impact mechanisms of these factors on the medal table, this research will not only improve the accuracy and reliability of medal table predictions but also provide a scientific basis for national Olympic committees to formulate preparation strategies and optimize resource allocation.

2 MEDAL PREDICTION BASED ON REGRESSION MODELS

2.1 Model Establishment

First, by reading the research literature of Bian X et al [6-8]. This study proposes three basic hypotheses: stable performance of athletes, stable national strategies, and home-field advantage. Meanwhile, this paper preprocesses the relevant data, unifies the code for team names, and addresses missing values and outliers.

In this study, a "comparative score" is derived based on various indicators of countries and athletes. The indicator scores are allocated 50% to countries and 50% to athletes. The percentage of medal distribution for each country is calculated. First, the percentage of medal distribution for each country was calculated. Taking into account the total number of medals from 2016 to 2024, the total number of gold medals and overall medals that each country is likely to obtain in 2028 was predicted based on this “ comparative score ”.

$$Y_i = \frac{(W_i \cdot \sum_{j=1}^{n} X_j + W_2 \cdot Z_i)}{\sum_{i=1}(W_i \cdot \sum_{j=1}^{n} X_{ij} + W_2 \cdot Z_i)} \times 100\% \times N \quad (1)$$

where W_1 and W_2 denote the weights of the total score of the country and the score of the athletes, respectively. Let $\sum_{j=1}^{n} X_j$ be the total score of all athletes in the i-th country. Let Z_i be the score at the level of the i-th country. Let $\sum_{i=1}^{n} X_{ij}$ be the sum of the comprehensive weighted scores of all countries. Let Y_i be the medal-distribution percentage of the i-th country.

2.2 Athlete Performance

To accurately evaluate an athlete based on the available information, this study has established a quantifiable metric, "Athlete Performance". The performance of athletes is primarily gauged by their medal-winning record in recent Olympic Games. The quantities of gold, silver, and bronze medals can directly mirror an athlete's competitive level. Meanwhile, a time-decay coefficient is introduced. As time elapses in the context of successive Olympics, the results of more recent competitions carry higher reference value. If an athlete participates in multiple events, additional points can be awarded, indicating a stronger comprehensive ability and a higher probability of winning medals. Considering the positive impacts of home-field advantage and environmental adaptability on an athlete's performance, if an athlete competes in the Olympics on behalf of the United States, extra points can be added.

$$X_{ij} = \sum_{i=1}^{4} T_i \cdot M_i \quad (2)$$

where X_{ij} represents the comprehensive score of the j-th athlete from the i-th country. T_i denotes the score obtained by this athlete for the i-th characteristic value, and M_i is the weight assignment for the i-th characteristic value.

The quantity and quality of medals are important indicators for measuring an athlete's comprehensive performance. However, a time-decay mechanism is required to more objectively evaluate their current competitive level. Therefore, the indicator is the medal score, and the formula is as follows:

$$T_1 = m_j \cdot \frac{1}{1 + 0.1 \times (P - p_i)} \quad (3)$$

where m_j represents the medal weight (Gold = 5, Silver = 3, Bronze = 1, No medal = 0), which reflects the relative value of Olympic medals, emphasizing the scarcity and significance of gold medals. P represents the year 2024, namely the most recently held Paris Olympics. p_i represents the year in which the medal was won. By $\frac{1}{1+0.1\times(P-p_i)}$ reducing the weight of medals won in earlier years, more influence is given to recent achievements.

$$T_3 = \sum_{i=1}^{n} n_i \quad (4)$$

where T_3 represents the total number of events an athlete has participated in within a single Olympic Games.

$$T_4 = \begin{cases} 2, & \textit{Win} \text{ a Gold Medal in 2024} \\ 1.5, & \textit{Win} \text{ a Silver Medal in 2024} \\ 1, & \textit{Win} \text{ a Bronze Medal in 2024} \\ 0, & \text{No Medal} \end{cases} \quad (5)$$

where T_4 represents the medal-winning situation of the athlete in the Paris Olympics. An athlete's performance in the most recent Olympic Games is generally more reflective of their current competitive level. Therefore, the importance of the performance in the most recent Olympics is higher, and the weight assigned to the performance in the 2024 Olympics is increased.

$$M_i = 0.6 \cdot Q_i + 0.4 \cdot L_i \quad (6)$$

where Q_i denote the correlation weight. The Pearson correlation coefficient between each indicator and the Medal Score is employed to measure the influence of the indicators on medal-winning performance. This approach ensures that the contribution of each indicator to the total score in the model is proportional to its actual influence, thereby avoiding the subjective assignment of weights.

$$Q_z = \frac{\left|correlation(T_x, T_z)\right|}{\sum_{y=1}^{4}\left|correlation(T_y, T_z)\right|} (x, z=1, 2, 3, 4) \quad (7)$$

The load values of each feature are calculated using standardized Principal Component Analysis (PCA), and then these load values are normalized to obtain weights, denoted as the PCA weights. PCA extracts the common characteristics

among various indicators, reducing redundant information. This process ensures the simplicity and robustness of the final model.

$$L_i = \frac{|l_i|}{\sum_j |l_i|} \tag{8}$$

where L_i denotes the loading of the i-th feature in the first principal component of PCA.

2.3 National Performance

The performance of a country is mainly measured by three key factors. First, is the number of times a country has hosted the Olympic Games. Countries that host the Olympics usually have strong economic strength, which enables them to invest more in competitive sports. And the more they invest in competitive sports, the higher the medal output tends to be. Second, the country's medal-winning record in previous Olympic Games. When considering this record, different weights are assigned to gold, silver, and bronze medals according to their medal rankings. This approach enables a better reflection of the value of each medal. Third, the home-field advantage. An exponential time-decay coefficient is introduced to more accurately represent the current competitive level.

Time-decay coefficient: This coefficient is designed to endow medals from more recent years with greater value.

$$f(t) = e^{-\lambda \cdot (p - p_i)} \tag{9}$$

Calculate medal score: For each type of medal (Gold, Silver, Bronze), scores are calculated according to the time-decay coefficient and the preset medal weights (5 for Gold, 3 for Silver, and 1 for Bronze).

$$\text{Medal Score} = \text{Medal count} \times f_{(t)} \times Weight \tag{10}$$

Host Scores: Each time a country hosts the Olympic Games, it receives a fixed bonus score.

Host advantage bonus:

$$\text{bonus (i)} = \begin{cases} 3, & if\ \text{Team}_i = \text{host country} \\ 0, & otherwise \end{cases} \tag{11}$$

Total Score: By using multiple linear regression for modeling, this study can derive the regression equation [9]. Calculate the total score for each country according to the scientific weights.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{12}$$

where X_i represents the indicator, β_i denotes the weight of the indicator, and ε stands for the error term.

2.4 Prediction of the Medal Table for the 2028 Los Angeles Olympics

Due to the fact that the original dataset does not provide the total number of Sports and Events in 2028, a multiple linear regression model based on a sliding time window is employed. By leveraging the number of Sports and Events in previous sessions provided in the dataset, the number of Sports and Events in 2028 is predicted. Subsequently, the range of the total number of medals is inferred.

$$m = \frac{\text{Weighted Average} - \text{Last Year Value}}{\text{Last Year} - \text{First Year}} \tag{13}$$

$$\text{Weight Average} = \frac{\sum_{i=1}^{n} e^{-\lambda \cdot t_i} \cdot x_i}{\sum_{i=1}^{n} e^{-\lambda \cdot t_i}} \tag{14}$$

$$w_i = e^{-\lambda \cdot t_i} \tag{15}$$

$$y = m \cdot \frac{2018 - Last\ \text{Year}}{Last\ \text{Year} - \text{First Year}} + x_{last\ \text{year}} \tag{16}$$

Given the significance and referential value of recent Olympic data, coupled with the year-on-year increase in Olympic events and the number of medals, this paper selects the sliding-time-window model for prediction.

Based on the above-mentioned calculations and analyses, the number of Olympic events generally exhibits an upward trend. However, the prediction indicates that the number of events in 2028 may experience a slight decrease. This is likely due to considerations of the current trends and certain limiting factors, such as budget constraints and venue availability.

3 ANALYSIS

The data used in this article is derived from https://www.olympics.com/.

3.1 Medal Prediction Based on Regression Models

3.1.1 Gold medal rankings

In the 2028 Olympic Games, the United States leads the gold-medal table with 43 gold medals, and China is expected to rank second with 40 gold medals. Countries including Japan, Australia, and France are also expected to obtain a certain quantity of gold medals, as demonstrated in Figure 1. In the predictive results, the rankings of countries based on the number of gold medals largely maintain consistency with those in 2024. There are fluctuations in the rankings of individual countries, but the overall pattern remains relatively stable.

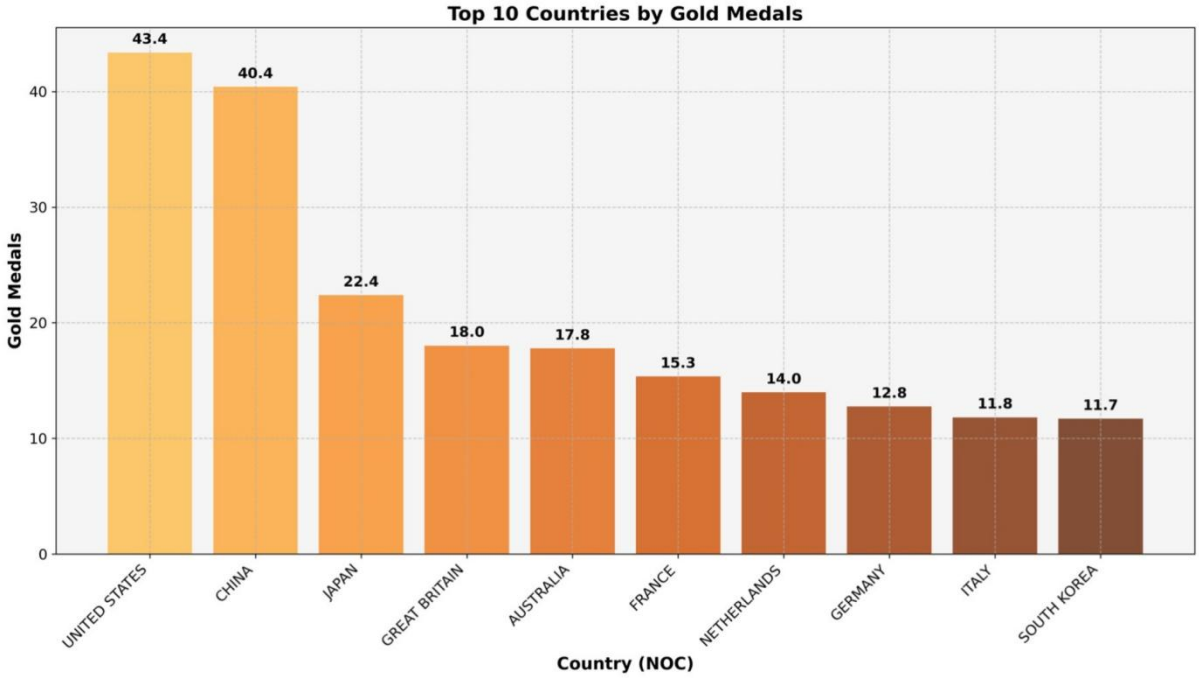

Figure 1 Gold Medal Rankings

3.1.2 Total medal rankings

In the 2028 Olympic Games, the United States leads the total-medal table with a total of 126 medals, followed by China with 91 medals. Countries such as the United Kingdom, Japan, France, and Australia also obtained a relatively large number of medals, as demonstrated in Figure 2. In the predicted scenario for the 2028 Olympic Games, although the rankings of countries based on the total number of medals exhibit certain changes compared to those in 2024, the overall trends remain similar.

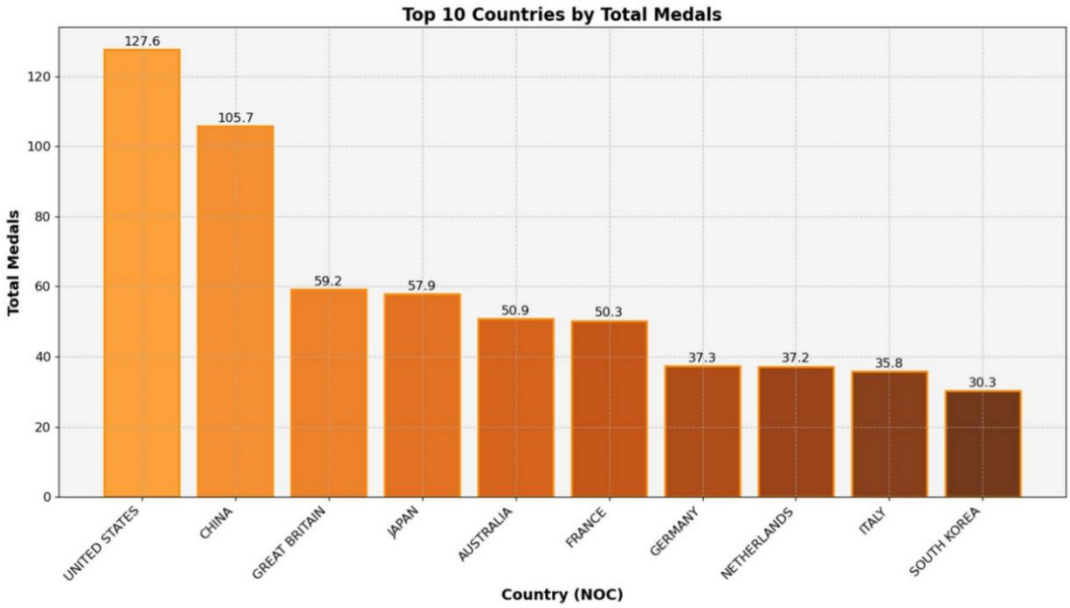

Figure 2 Total Medal Rankings

3.2 Sensitivity Analysis

In the process of quantitative modeling, two important factors are identified for the two indicators of countries and athletes respectively. One consists of the time-decay coefficient and athlete participation, while the other includes the intensity of sports events and international competition experience. These four parameters are being adjusted to simulate Olympic competitions at different times. Previously, to better present these results, the score changes of the top ten countries and the top ten athletes are now calculated. As shown in Figure 3, the model is quite sensitive to the setting of weights for each factor.
In practical applications, weights need to be carefully determined to ensure the accuracy of the prediction results.

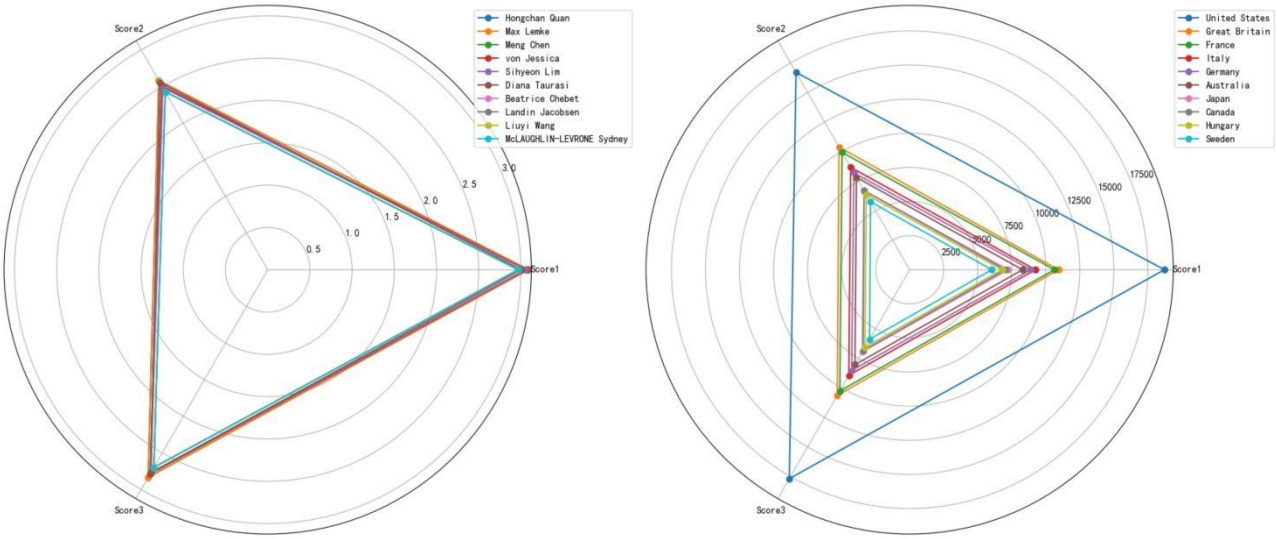

Figure 3 Sensitivity Analysis

4 CONCLUSION

This study presents a linear regression model for predicting the total medal standings and gold medal rankings of the 2028 Olympic Games. The model is developed through a comprehensive analysis of data from participating nations and athletes, employing a weighted fusion approach that incorporates time decay coefficients to account for temporal variations in performance metrics. Unlike conventional methodologies, this framework integrates multiple contextual factors beyond historical data alone, including host-nation advantage, the impact of elite coaching, and Olympic program configurations. These enhancements significantly improve the predictive accuracy and reliability of medal standings forecasts, offering actionable insights for national Olympic committees to optimize training strategies and resource allocation. The study concludes by advocating for further refinements of the model and its broader application across sporting contexts, thereby contributing to the theoretical and practical development of sports analytics.

COMPETING INTERESTS

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

REFERENCES

[1] Houlihan B, Zheng J. The Olympics and Elite Sport Policy: Where Will It All End? The International Journal of the History of Sport, 2013, 30(4): 338-355.
[2] Zhou Xiaobo, Zhou Liqun. Population Quality, Political Institutions, and Olympic Performance—Evidence from Four Olympic Games. South China Journal of Economics, 2016(08): 1-11.
[3] Wen Jing, Li Weiping, Lei Fumin. Predictive Study on Gold Medals and Medals Won by China at the Beijing Winter Olympics Using Multiple Methods//China Sports Science Society. School of Physical Education and Health, Hangzhou Normal University; Statistics Teaching and Research Section, Xi'an Physical Education University, 2022: 20-22.
[4] Shi Huimin, Zhang Dongying, Zhang Yonghui. Can Olympic Medals Be Predicted?—From the Perspective of Explainable Machine Learning. Journal of Shanghai University of Sport, 2024, 48(04): 26-36.
[5] Yuan Junjie. Preliminary Study on Gold Medal Prediction Model for the Olympic Games in the Big Data Era—Taking the Results of the World Athletics Championships as an Example. Bulletin of Sport Science & Technology, 2021, 29(06): 132-134.
[6] Aygün M, Savaş Y. Analysing Winter Olympic Medals Through Economic Variables: A Comprehensive Examination. Research in Sport Education and Sciences, 2024, 26(4): 197-209.

[7] Wilson D, Ramchandani G. A comparative analysis of home advantage in the Olympic and Paralympic Games 1988–2018. Journal of Global Sport Management, 2021, 6(2): 170-184.
[8] Luo Yubo, Cheng Yanfang, Li Mengyao, et al. Prediction of China's Medal Count and Overall Strength in the Beijing Winter Olympics—Based on the Host Effect and Gray Prediction Model. Contemporary Sports Technology, 2022, 12(21): 183-186.
[9] Etemadi S, Khashei M. Etemadi multiple linear regression. Measurement, 2021, 186: 110080.