Random Forest Regression-Based Olympic Medal Prediction Model

Xu Wang*, Tianyu Wang, Zhenyu Li

School of Physics, Jilin University, Jilin, 130012, China

*Corresponding author: 2894942595@qq.com

Abstract. As one of the most influential sporting events in the world, the Olympic medal table has always been an important benchmark for national sports development. To enhance the accuracy and interpretability of Olympic medal predictions, this study constructs a comprehensive Olympic medal prediction framework based on a random forest regression model. The model extracts rich features from multiple dimensions, including historical performance, host country advantage, event scale, and athlete strength, providing a comprehensive reflection of each country's Olympic participation strength. Analysis of the prediction results for the 2028 Los Angeles Summer Olympics indicates that the United States will continue to lead the medal table, with China and Japan also emerging as major medal-winning nations. Traditional powerhouses such as Great Britain, Australia, and France, as well as emerging strong competitors like Germany and the Netherlands, are also expected to achieve remarkable results in this Olympics. This prediction model provides scientific decision-making support for national sports management departments, helping optimize national sports strategies and resource allocation.

Keywords: Olympic medal, prediction model, random forest regression model, sports development.

1 Introduction

With the increased attention on Olympic Games, tracing back to first modern games in Greece in 1896, the Games more and more became a competition among nations compared to its initial spirit of individual competition [1-3]. In the latest summer Olympic Games in Paris, 2024, the United States and China had a particularly fierce competition on the number of golden medals. One way of wining more medals in Olympic is to hire more specific coach for program training. Provided the same base and talent for athletes, the development of coach has been identified as a key for the competition. Besides the technical support, good mental health assistance is also vital for improving athletes' performance [4].

In the existing research, there have been many studies that attempt to establish mathematical models to predict future Olympic medal tables. Early research was mostly based on statistical analysis and regression models, combined with the historical medal

A. Mallik et al. (eds.), *Proceedings of the 2025 6th International Conference on Management Science and Engineering Management (ICMSEM 2025)*, Atlantis Highlights in Economics, Business and Management 4,
https://doi.org/10.2991/978-94-6463-845-5_74

data of each country for prediction [5]. Scholars who have proposed statistical learning methods for medal prediction have considered the influence of historical medals and related factors on the number of medals [6]. In recent years, with the development of machine learning and big data technology, some studies have begun to adopt more complex algorithms, such as support vector machines, random forests, and neural networks, which can better handle the nonlinear relationships and complexity in the data [7]. Some scholars have used deep learning models to predict Olympic medal counts and proposed models that consider the composition of athletes in each country and the types of events, which have achieved relatively good predictive effects [8]. In addition to the traditional medal prediction, in recent years, some scholars have focused on the research of the "great coach effect", pointing out that the influence of coaches may have a significant impact on the performance of national teams, especially in high-level sports, where the strategic vision and training methods of coaches can greatly improve the competitive state and probability of winning medals for athletes [9].

Compared with previous models, which often rely on limited features or focus solely on medal statistics, our approach incorporates multi-dimensional variables (e.g., event participation rate, athlete performance trajectory, and host advantage) and leverages ensemble learning to better handle nonlinearity and feature interaction. This comprehensive and data-driven framework marks a significant advancement over traditional regression-based or rule-based predictions.

2 Methodology

2.1 Data Preprocessing

The data in this study primarily comes from the records of the International Olympic Committee (www.Olympics.com) during specific Olympic Games. The data includes complete national medal statistics for all Summer Olympics, a list of host countries for the Summer Olympics, all participants along with their sports, years, and results, as well as the number and total count of events categorized by sport/discipline. As with all datasets, there may be anomalies in the records. Therefore, we conducted a data quality analysis and preprocessing on the raw dataset, which mainly involved missing value analysis, outlier detection, and consistency analysis.

2.2 Feature Engineering

The goal of feature engineering is to transform raw data into more predictive features while maintaining data interpretability. Only by identifying more valuable features can we effectively enhance model performance. Therefore, based on our understanding of Olympic competition patterns and data analysis results, we have constructed a multidimensional feature system.

Historical Performance Features. Historical performance features play a crucial role in predicting Olympic success, as they capture the continuity and evolution of a nation's competitive strength. These features include both absolute and relative measures of past achievements. For medal counts, we analyze incremental changes and growth rates across gold, silver, and bronze categories, revealing how quickly a country's performance is improving or declining. Historical rankings provide context about a nation's standing relative to others. To assess consistency, we calculate a weighted moving average of total medals from the last three Games, giving more importance to recent performances. Performance stability is measured through standard deviation and coefficient of variation of medal counts. For sport-specific predictions, we apply linear regression to recent results within defined time windows and compute historical medal probabilities by event, highlighting countries' traditional strengths in particular disciplines. These comprehensive metrics collectively provide a nuanced understanding of a nation's athletic trajectory, enabling more accurate forecasting of future Olympic performance. The multi-dimensional approach accounts for both overall trends and sport-specific capabilities, offering a robust foundation for predictive modeling.

Host Country-Related Features. Since host countries often have a certain homefield advantage, we designed the following two features: A binary variable indicating whether a country is the host nation. Historical host advantage, measured as the difference between the host country's medal count in its hosting year and the average of the two adjacent Olympics, quantifying the magnitude of the homefield advantage. Host country-related features: Since host countries often have a certain homefield advantage, we designed the following two features: A binary variable indicating whether a country is the host nation. Historical host advantage, measured as the difference between the host country's medal count in its hosting year and the average of the two adjacent Olympics, quantifying the magnitude of the homefield advantage. This dual feature design can not only identify the identity of the host country, but also accurately measure the strength of the home advantage, thereby more accurately predicting its performance differences. In addition, this method has strong versatility and can be applied to the home advantage analysis of other international sports events such as the World Cup and the Asian Games with slight adjustments.

Event and Athlete Features. We constructed features from three dimensions: event scale, participation scale, and athlete strength. Event Scale: Includes the total number of events, total disciplines, and total sports in past Olympics. A greater number of total events increases the number of medals available. The increase in disciplines and sports means that some countries might unexpectedly win medals in newly added events. Participation Scale: Includes the number of specific events each country participates in and its event coverage rate. While it is generally believed that a country's medal count increases with the total number of events it competes in, if all participation is concentrated in highly competitive events, the improvement effect may be unpredictable. The introduction of event coverage rate helps identify whether a country fully utilizes medal-winning opportunities across different events, allowing for a more accurate prediction of its medal potential. Athlete Strength: Includes the number of participating athletes, the number of times an athlete has competed, and their past medal-winning records.

Generally, the more athletes a country sends, the higher its probability of winning medals.

2.3 Random Forest Regression Model

The random forest regression model is an ensemble learning method that uses regression trees as base learners, enhancing the model's predictive accuracy by introducing randomness. This randomness is mainly reflected in two aspects: feature selection and dataset sampling [10-12]. The former refers to the random selection of m features during node splitting, where m is typically one-third of the total number of features for regression problems. The latter involves random sampling of the dataset using the Bootstrap strategy, which performs resampling with replacement from the original dataset to create sub-training sets, so each regression tree is generated based on a different training dataset. Each tree produces a prediction, and the final prediction is determined through majority averaging (for regression problems).

In random forest regression, the prediction of each tree can be represented as:

$$\hat{y}_t = \sum_{i=1}^{N} w_i x_i + b \tag{1}$$

where $\hat{y}_t$ is the prediction of the t-th tree. w_i is the weight of the i-th feature. x_i is the i-th feature. b is the bias term. Among the predictions of all trees, the final prediction is obtained by averaging all tree predictions.

3 Results

In order to verify the performance of our model, we use RMSE to visualize the result since it is a common indicator to evaluate the prediction accuracy of regression models, as show in Fig. 1.

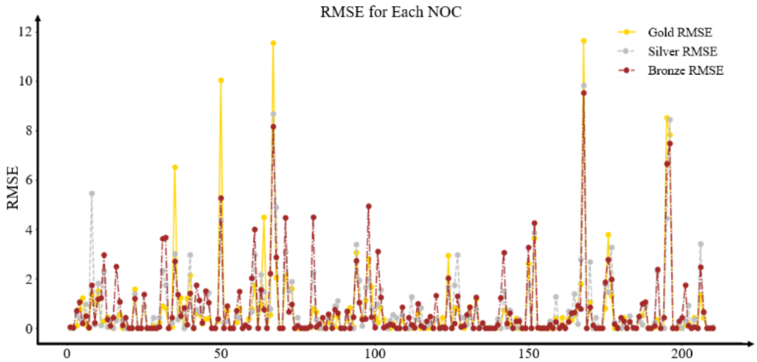

Fig. 1. Accuracy evaluation.

From Fig. 1, we can see that most of the points are concentrated in the regions with low RMSE value, indicating that the medal number prediction accuracy of most countries is relatively high. But there still exist some countries with high RMSE which means it might have a wide deviation in one medal category. According to the model, the top ten countries in the medal table of the 2028 Los Angeles Summer Olympics are finally predicted, as show in Fig. 2.

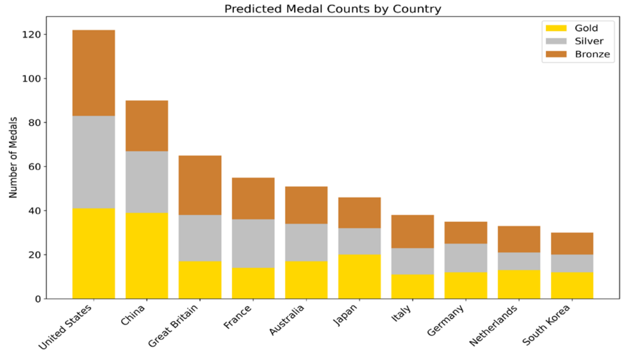

Fig. 2. Medal prediction.

From Fig. 2, it can be seen that the United States is expected to lead the medal table with a significant margin, totaling 122 medals. The U.S. is projected to secure 41 golds, 42 silvers, and 39 bronzes, maintaining its historical dominance in the Olympic Games. China follows closely with 90 medals, including 39 golds, 28 silvers, and 23 bronzes, which positions them second in the rankings. Japan, with a projected 46 medals, is likely to remain a strong contender, placing third. Great Britain, Australia, and France are also predicted to perform well, with 65, 51, and 55 medals respectively. Great Britain, in particular, is expected to achieve a balanced medal distribution across gold (17), silver (21), and bronze (27). The Netherlands, Germany, South Korea, and Italy are all expected to secure a respectable number of medals, placing them among the top ten in the medal tally. Overall, the U.S. is projected to dominate the 2028 Summer Olympics, but the competition remains strong, with several nations making significant contributions to the global medal tally.

4 Conclusion

This study successfully constructed an Olympic medal prediction model based on a random forest regression model. The prediction results indicate that the United States will continue its dominance in the Olympics, with an expected total of 122 medals, including 41 gold medals. China and Japan are also projected to be major medal-winning nations, ranking second and third, respectively. Additionally, traditional powerhouses like Great Britain, Australia, and France, as well as emerging strong competitors such as Germany and the Netherlands, are expected to achieve outstanding results at the 2028 Olympics. This data-driven and machine-learning-based medal prediction model provides scientific decision-making support and reference for national sports management departments.

This model can serve multiple functions in national sports management. For example, it can be used to identify underperforming disciplines with high medal potential, allocate training resources more efficiently, and inform long-term Olympic preparation strategies. Furthermore, policymakers can leverage prediction insights to assess the return on investment for elite athlete development programs or to guide strategic recruitment and coaching initiatives tailored to high-impact sports.

References

1. Nagpal, P., Gupta, K., Verma, Y., & Kirar, J. S. Paris Olympic (2024) Medal Tally Prediction Check for updates. Data Management, Analytics and Innovation: Proceedings of ICDMAI 2023, 662, 249 (2023).
2. Tchamkerten, A., Chaudron, P., Girard, N., Monnier, A., Pyne, D. B., & Hellard, P. Career factors related to winning Olympic medals in swimming. Plos one, 19(6), e0304444 (2024).
3. Weber, A. C., De Bosscher, V., Shibli, S., & Kempf, H. Strategic analysis of medal markets at the Winter Olympics: Introducing an index to analyse the market potential of sports disciplines. Team Performance Management: An International Journal, 25(3/4), 229-252 (2019).
4. Gupta, K. G., & Arora, A. Olympic Data Analysis: Uncovering New Insights into Athletic Performance and Competition. Grenze International Journal of Engineering & Technology (GIJET), 10 (2024).
5. Yao, Y., & Niu, X. Research on the champion physical fitness model of freestyle skiing aerials athletes in preparation for the Beijing Winter olympics. Scientific Reports, 14(1), 29107 (2024).
6. Zheng, J., Dickson, G., Oh, T., & De Bosscher, V. Competitive balance and medal distributions at the Summer Olympic Games 1992–2016: Overall and gender-specific analyses. Managing Sport and Leisure, 24(1-3), 45-58 (2019).
7. Yeh, C. C., Peng, H. T., & Lin, W. B. Achievement Prediction and Performance Assessment System for Nations in the Asian Games. Applied Sciences, 14(2), 789 (2024).
8. Zheng, J., Dickson, G., Oh, T., & De Bosscher, V. Competitive balance and medal distributions at the Summer Olympic Games 1992–2016: Overall and gender-specific analyses. Managing Sport and Leisure, 24(1-3), 45-58 (2019).
9. Sajadi, S. M., Baghaie, S., & Rezaei, R. Optimizing sports development: Identifying and prioritizing key indicators for professional and competitive sports. World Development, 180, 106651 (2024).
10. Xu, Z., Lin, Y., Cai, H., Zhang, W., Shi, J., & Situ, L. Risk assessment and categorization of terrorist attacks based on the Global Terrorism Database from 1970 to 2020. Humanities & Social Sciences Communications 11, 1103 (2024).
11. Xu Z. Coupling Coordination Development and Driving Factors of New Energy Vehicles and Ecological Environment in China. Wuhan University Journal of Natural Sciences 30(1), 79-90 (2025).
12. Xu, Z. Machine Learning-Based Quantitative Structure-Activity Relationship and ADMET Prediction Models for ERα Activity of Anti-Breast Cancer Drug Candidates. Wuhan University Journal of Natural Sciences 28(3), 257-270 (2023).