# Olympic Medal Count Prediction and Analysis Based on Ridge Regression and Weighted Random Forest

Jiale Wang [1], Boyu Xu [1, *, #], Jialin Liao [2, #]

[1] School of Information Management, Sun Yat-sen University, Guangzhou, China, 510006

[2] School of Software Engineering, Sun Yat-sen University, Zhuhai, China, 519000

* Corresponding Author Email: xuby39@mail2.sysu.edu.cn

[#]These authors contributed equally.

**Abstract.** This article comprehensively completed the prediction of medal distribution for the 2028 Summer Olympics by constructing statistical models including ridge regression model and weighted random forest model to predict the number of Olympic medals in various countries. At the same time, in-depth analysis was conducted on the specific situation of countries that won medals for the first time at the Olympics, revealing the factors behind their breakthrough achievements. In addition, this article explores the intrinsic relationship between different events and the number of medals awarded by each country, and identifies key events that may become the focus of medal competition. Furthermore, this study conducted an in-depth analysis of the "host effect" , exploring the additional advantages a host country may gain, as well as the "great coach effect" , examining the potential impact of excellent coaches on national medal tallies. It systematically analyzed the multidimensional factors influencing national Olympic performance.In the regression prediction stage, this article constructed a ridge regression model to solve the problem of multicollinearity in data, and deployed a multiple linear regression model to capture the linear relationship between variables. The introduced XGBoost and GDBT gradient enhancement algorithms perform well in handling large-scale data and high-dimensional features. In addition, the Transformer model provides strong support for complex data with its powerful sequence modeling capabilities and parallel computing advantages.

**Keywords:** Olympic medals, Ridge regression, Weighted random forest, Ensemble learning.

## 1. Introduction

The Olympic Games represent the pinnacle of international sport, bringing together nations to compete and showcase their sporting excellence. Medal tables, which summarize the distribution of gold, silver, and bronze medals, are a key indicator of national sporting performance. Predicting these medal counts involves a complex interaction of historical data, athlete performance, and event-specific dynamics, making it a challenging but rewarding research topic.[1] At the 2024 Paris Olympics, countries like the United States and China dominated the medal table, while smaller nations like Albania and Saint Lucia celebrated milestones by winning their first-ever Olympic medals. These results highlight the diversity and evolving dynamics of the Olympics, in which both traditional powers and emerging nations compete on the world stage.

This article needs to solve four types of problems, of which the first three types need to be modeled, and the last type needs to be completed by integrating the model analysis results of the first three types. The first is to develop a model for medal prediction, which requires both performance evaluation of the model itself and in-depth analysis of the prediction results (the range of prediction results, the progress and regression of different countries compared with 2024). Then, a model is built to predict the number of countries that will win the award for the first time in the next session, and the accuracy of this prediction needs to be known. The next step is to build a model to explore the impact of the characteristics of the Olympics (number and type of events, etc.) on the number of medals of various countries, as well as to explore the impact of the "host effect" on national medals. Finally, combined with the solutions to the first three types of problems, it reveals unique insights into the number of Olympic medals and provides advice to the International Olympic Committee.

## 2. Prediction of Olympic medals based on ridge regression model

Ridge regression is an improved linear regression method that prevents overfitting and solves multicollinearity problems by adding the sum of squares of regression coefficients (L2 regularization term) to the loss function.[2] The core idea is to limit the size of regression coefficients to prevent certain coefficients from becoming too large due to strong correlations between features,[3] which can lead to model instability. The strength of the regularization term is controlled by the hyperparameter λ: the larger λ, the stronger the constraint, and the simpler the model; The smaller the λ, the closer the model is to ordinary linear regression. In this way, ridge regression performs well in improving generalization ability and model robustness.[4] The mathematical formula is as follows Figure 1:

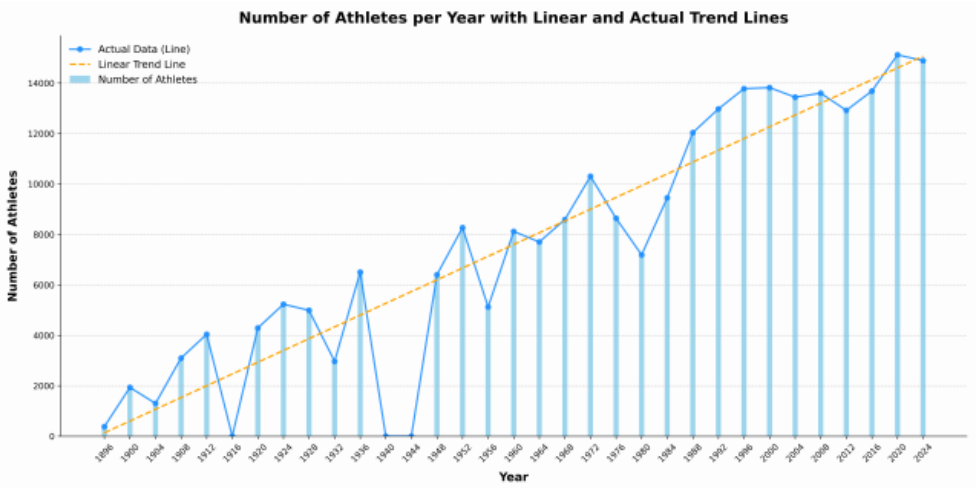


**Figure 1** Number of athletes per year

$$Y_i = \mu_0 + \mu_1 a_1 + \mu_2 a_2 + \ldots + \mu_n a_n + \varepsilon \quad (1)$$

Where: $Y_i$: Predicted medal count for the i-th sample,$\mu_0$: Intercept term, representing the base line medal count when all features are zero,$a_1, a_2, \ldots a_n$: Feature variables, such as the number of Olympic Games, the number of events in the current Games, and medal counts in specific events,$\mu_1, \mu_2, \ldots \mu_n$: Regression coefficients, representing the influence of each feature variable on the medal count,$\varepsilon$: Error term, representing the random variation and unexplained part of the model.

The regularized loss function for ridge regression is:

$$J(\mu) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \mu_j^2 \quad (2)$$

Where:$J(\mu)$: Loss function of ridge regression,$\hat{y}_i$: Predicted medal count for the i-th sample, calculatedas $\hat{y}_i = \mu_0 + \mu_1 a_1 + \ldots + \mu_n a_n$ , $y_i$ : Actual medal count for the i-th sample, $\lambda$ : Regularization parameter, controlling the penalty on the sum of squared regression coefficients, $p$: Number of features, as shown in Figure 2.

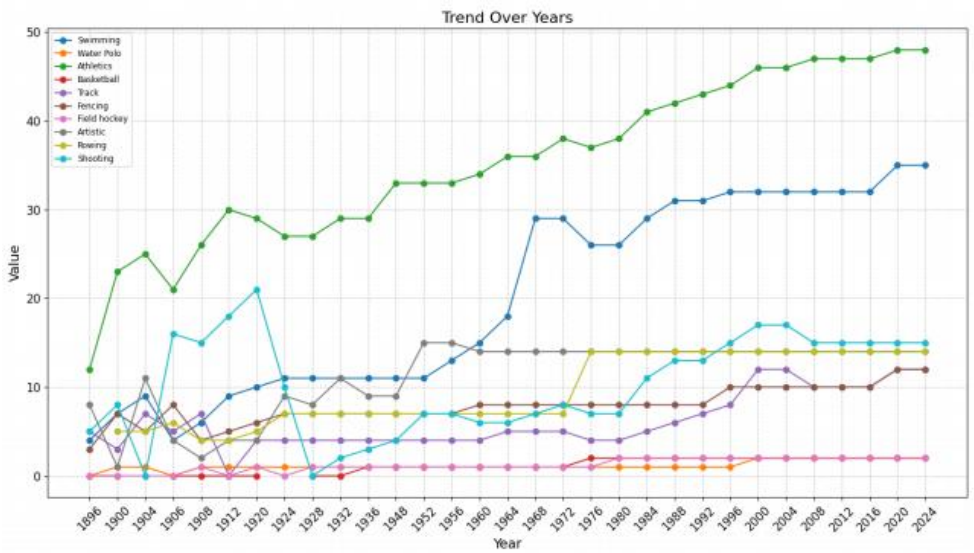


**Figure 2:** Trend over years

$$\hat{\mu} = (A^{\top}A + \lambda I)^{-1}A^{\top}Y \tag{3}$$

$\hat{\mu} = (\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_n)^{\top}$: Estimated regression coefficients, A: Design matrix, an $N \times (n+1)$ matrix, where each row represents the feature variables of a training sample, and the first column is the intercept term, $\lambda$: Regularization parameter, I: Identity matrix, used to add the regularization term to $(A^{\top}A + \lambda I)$,$Y$: Target variable, an $N \times 1$ vector containing the actual medal counts of all training samples,$A^{\top}$: Transpose of the design matrix A.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4}$$

$R^2$: The coefficient of determination, representing the proportion of variance in the dependent variable that is explained by the model, $R^2 \in [0, 1]$: The closer $R^2$ is to 1, the better the model fits the data,$\hat{y}_i$: The predicted value of the dependent variable (e.g., medal count) for the i-th sample,$y$: The mean value of all actual observations.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i | \tag{5}$$

$MAE$: The mean absolute error, which measures the average magnitude of errors between the predicted and actual values, A smaller MAE indicates better predictive performance, $N$: The number of samples. $y_i$: The actual value of the dependent variable (e.g., medal count) for the i-th sample,$|y_i - \hat{y}_i|$: The absolute error for the i-th sample.

### 2.1. Medal Count Prediction

The data on the distribution of Olympic medals may be affected by noise, such as extreme home advantage or emerging sports. Ridge regression can more effectively filter out these noises and improve prediction accuracy. The indirect data source is related data from previous Summer Olympics, while the direct data source is our processed data. This article divides the 8:2 dataset into a training set and a testing set. By learning from training data, the model can capture potential relationships between features and medal counts, and provide scientific predictions for future medal distribution. In the evaluation of the preliminary model, this article found that the performance of the ridge regression model basically meets expectations. To further validate its superiority, we also compared it with several other models. According to Table 1, the mean absolute error (MAE) of the ridge regression model is 6.77, which is only slightly higher than other complex models (such as Transformer's 5.1591 and XGBoost's 5.28), but the difference is not significant while ensuring the interpretability of the model. In addition, its coefficient of determination is 0.343, indicating that the ridge regression model can explain a considerable amount proportion of the changes in the number

of medals and performs well in simple linear models. It is worth emphasizing that the ridge regression model effectively avoids the overfitting problem that may occur in complex models and provides a solution to balance prediction accuracy and model complexity,as shown in Table 1.

**Table.1.** Performance Comparison of Models

| Evaluation | RMSE | $R^2$ | MAE | MAPE |
|---|---|---|---|---|
| Bidge Begression | 25.341 | 0.343 | 6.77 | 2811391832216909.0 |
| Linera Begression | 21.150 | 0.542 | 6.19 | 3007074137076710.0 |
| MGBoost | 17.443 | 0.689 | 5.28 | 1506054015392154.0 |
| GDBT | 17.766 | 0.677 | 5.29 | 2502752448804098.5 |
| Transfomer | 9.478 | 0.950 | 5.16 | 1307056051302174.0 |

Overall, the ridge regression model provides an ideal choice for quickly analyzing and building an intuitive medal prediction model. Through a simplified linear framework, it can efficiently capture the relationship between the main features, providing a solid foundation for subsequent in-depth analysis and optimization. These results show that the ridge regression model is indeed very suitable for efficiently capturing the interactive relationship between complex features, thereby improving the accuracy of predictions, as shown in Table 2.

**Table.2.** 2028 Predicted Medal Table

| NOC | Pred Gold | Pred Total | Total CI | | Gold CI | |
|---|---|---|---|---|---|---|
| | | | Lower | Upper | Lower | Upper |
| USA | 55 | 159 | 122 | 196 | 41 | 69 |
| CHN | 35 | 99 | 62 | 136 | 21 | 49 |
| FRA | 24 | 75 | 38 | 112 | 10 | 38 |
| GBR | 18 | 65 | 28 | 102 | 4 | 32 |
| AUS | 18 | 56 | 19 | 93 | 4 | 32 |
| JPN | 17 | 49 | 12 | 86 | 2 | 30 |

## 3. Weighted Random Forest Model

### 3.1. Model Introduction

The training objective of weighted random forests is to minimize the weighted loss function:For regression tasks, the loss function is typically the weighted mean squared error:

$$L_{WMSE} = \frac{1}{N}\sum_{i=1}^{N} \omega_i \cdot (y_i - \hat{y}_i)^2 \tag{6}$$

Where $N$ is the number of samples, $\omega_i$ is the weight of the i-th sample, $y_i$ is the true value, and $\hat{y}_i$ is the predicted value.

For classification tasks, the weighted log loss is typically used:

$$L_{WLL} = -\frac{1}{N}\sum_{i=1}^{N} \omega_i \cdot \sum_{c=1}^{C} y_{i,c} \cdot \log(\hat{y}_{i,c}) \tag{7}$$

where $C$ is the number of classes, $y_i, c$ is the true class of the i-th sample, and $\hat{y}_i, c$ is the predicted probability for class $c$.

For classification tasks, the weighted $Gini$ index can be used to evaluate:

$$Gini = 1 - \sum_{c=1}^{C} p_c^2 ,\ p_c = \frac{\sum_{i=1}^{N} \omega_i \cdot \mathbf{1}(y_i = c)}{\sum_{i=1}^{N} \omega_i} \tag{8}$$

where $p_c$ is the weighted probability of samples belonging to class $c$ in the given dataset. For regression tasks, the weighted variance can be used to evaluate node purity:

$$\text{Weighted} = \frac{\sum_{i=1}^{N} \omega_i \cdot (y_i - \hat{y})^2}{\sum_{i=1}^{N} \omega_i}, \ \hat{y} = \frac{\sum_{i=1}^{N} \omega_i \cdot y_i}{\sum_{i=1}^{N} \omega_i} \quad (9)$$

where $y_i$ is the true target value of the i-th sample, $\hat{y}$ is the weighted mean of all samples in the current node, and $\omega_i$ is the weight of the i-th sample.

In weighted random forests, the prediction or classification result of each tree is weighted.[5] For regression tasks, the final prediction is the weighted average of all trees [6]:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \omega_t \cdot \hat{y}_t \quad (10)$$

For classification tasks, a weighted voting method is typically used to obtain the final class:

$$\hat{y}_c = \frac{1}{T} \sum_{t=1}^{T} \omega_t \cdot \mathbf{1}(\hat{y}_t = c) \quad (11)$$

Where $T$ is the number of trees, $\omega_t$ is the weight of the t-th tree, and $\hat{y}_t$ is the prediction of the t-th tree.

The above process demonstrates the basic principles of Bagging, Boosting, and Stacking ensemble learning methods.[7] These three ensemble learning methods each have their own characteristics and are suitable for different scenarios and problems. Stacking combines the prediction results of multiple base learners and uses meta learners to further optimize prediction performance.[8] Basic learners can be different types of models, while meta learners are usually a simple model (such as linear regression or logistic regression). It can capture the advantages of different basic learners and improve the overall model's generalization ability. It is suitable for complex datasets and can effectively improve prediction accuracy. However, it has high computational complexity and long training time. It requires more data and computing resources to iteratively train multiple weak learners, with each new model attempting to correct the errors of the previous model and adjust sample weights. The final model is a weighted combination of these weak learners. It can effectively reduce bias and variance, and improve prediction accuracy. It is particularly suitable for handling complex datasets. However, it is sensitive to noise and outliers, making it prone to overfitting. Long training time and high computational complexity. Bagging generates multiple training subsets through guided sampling and trains a basic learner subset for each subset.[9] The final prediction result is the average of these bases (regression) or voting (classification) learner.

Random forest is a typical application of bagging, which uses decision trees as the base learner. It effectively reduces the variance of the model by averaging and improves the generalization ability of multiple models' prediction results. Due to the fact that each basic learner is trained on a different subset, random forests have strong resistance to overfitting. Every basic learner can train in parallel to improve computational efficiency. Random forests can handle high-dimensional data without the need for feature selection. It is insensitive to noise and outliers and has strong robustness.[10] Compared to boosting, the basic learner of random forest can be trained in parallel, resulting in higher computational efficiency. In addition, random forests are insensitive to noise and outliers and have strong robustness, while boosting is more sensitive to noise and outliers. It is suitable for processing high-dimensional data and complex datasets.

### 3.2. Host Country Effect

The host country usually performs well in the Olympics, a phenomenon known in academia as the "host effect". Due to psychological advantages, home support, increased investment, and familiarity with the competition venues, host countries tend to win more medals in the Olympics.For example, China and the United Kingdom both significantly increased their medal counts when they hosted the Olympics. Athletes from host countries often face tremendous psychological pressure, and facing the eager expectations of the home audience is both an incentive and a challenge. Many athletes are able to surpass themselves and set new records with the support of their hometown audiences, as shown in Figure 3.

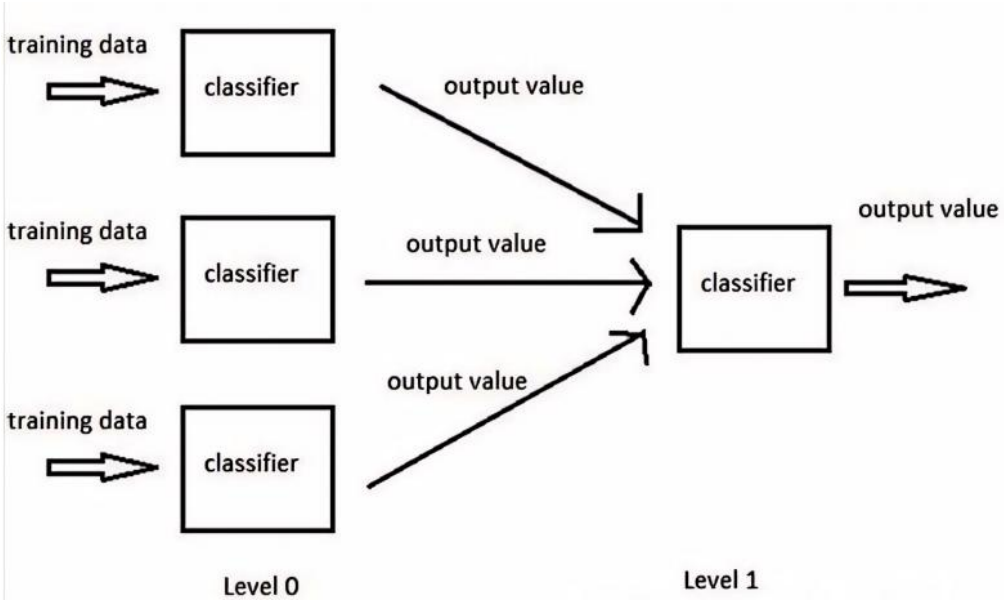


**Figure 3** Stacking-Based Predictive Model for Host Country Medal Gains

### 3.3. Impact of Events on National Medal Counts

Because this article aims to explore the relationship between each competition event and the number of medals won by each country, a weighted ratio algorithm is used for each country. Based on the comprehensive scores of each project (including gold, silver, and copper), select the three most important projects for display.

### 3.4. Prediction of the Number of Countries Winning Medals for the First Time

The probability of a country winning its first Olympic medal is predicted using a weighted binary cross-entropy loss function:

$$\Gamma = -\frac{1}{n}\sum_{i=1}^{n} \omega_i \left[p_i \log \hat{p}_i + (1 - p_i) \log(1 - \hat{p}_i)\right] \quad (12)$$

Where $p_i$ represents the true label indicating whether a country wins a medal ($p_i$ = 1 for yes, $p_i$ = 0 for no), $\hat{p}_i$is the predicted probability of a country winning a medal, and $\omega_i$ is the sample weight.The weights $\omega_i$ can be adjusted to assign higher importance to countries that have not yet won a medal, thereby emphasizing these samples during the training process, as shown in Figure 4.

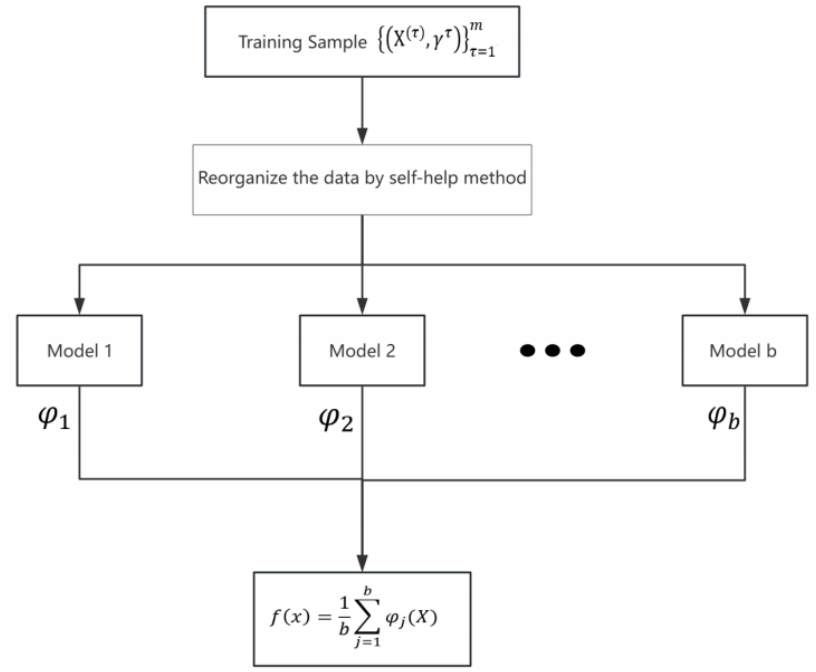


**Figure 4** Bagging-Based Prediction Model for First-Time Medalist Countries

In the binary classification stage of predicting the country's first medal, a variety of modeling strategies were also demonstrated. First, a logistic regression model was constructed, which pro vides a robust benchmark for classification tasks with its basic advantages of simplicity and strong interpretability. In addition, the K-nearest neighbor algorithm (KNN) was introduced to capture the local similarities between samples using its distance-based classification idea. The support vector machine (SVM) model was also included to improve the classification accuracy of the model by maximizing the classification interval. In order to further improve the performance of the model, the team also deployed the LightGBM model, which performs well in processing large-scale data sets with its efficient training speed and excellent processing capabilities.

## 4. Unique Insights

First, the significant increase in the number of medals won by the host country is called the "host country effect".[11] For example, the United States in the 1984 Los Angeles Olympics and the Soviet Union in the 1980 Moscow Olympics both showed significant medal outliers. This phe nomenon may be related to the various advantages that the host country enjoys in the competition, such as higher athlete participation, home support, familiar environment, and possible influence on competition rules and arrangements. In addition, the outliers in 1980 and 1984 are closely related to the international situation during the Cold War, when some countries boycotted the Olympics for political reasons,[12] resulting in a significant deviation from the norm in the medal distribution as shown in Figure 5.

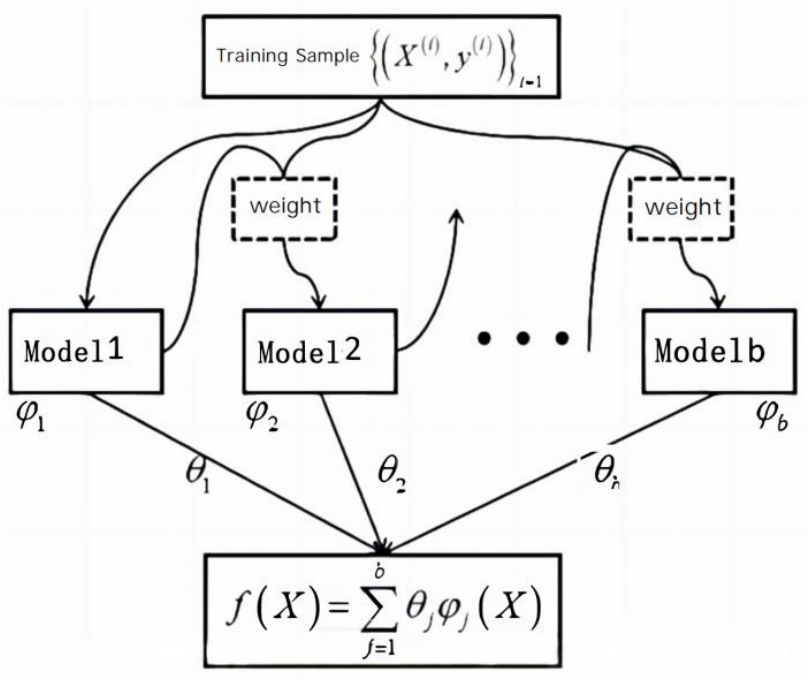


**Figure 5** Boosting-Based Modeling of Olympic Medal Count Deviations

Countries such as the United States and the United Kingdom have also repeatedly become out liers in the total number of medals as non-host countries. For example, the United States showed the top non-host country outliers in 1988 and 2012, indicating that these countries have maintained a stable competitive advantage in the Olympics for a long time. This advantage may come from sufficient sports funding support, perfect infrastructure, and a systematic athlete training mecha nism.

The Figure 6 also reveals the historical evolution of the Olympic medal distribution. In the early days (e.g., 1900 to 1920), medal distribution was concentrated in a few countries, which may be related to the small number of participating countries and the fact that international sports were not yet popular. Over time, as more and more countries participated in the Olympics, the medal distribution gradually became more balanced, reflecting the continuous improvement of the global sports competition level.

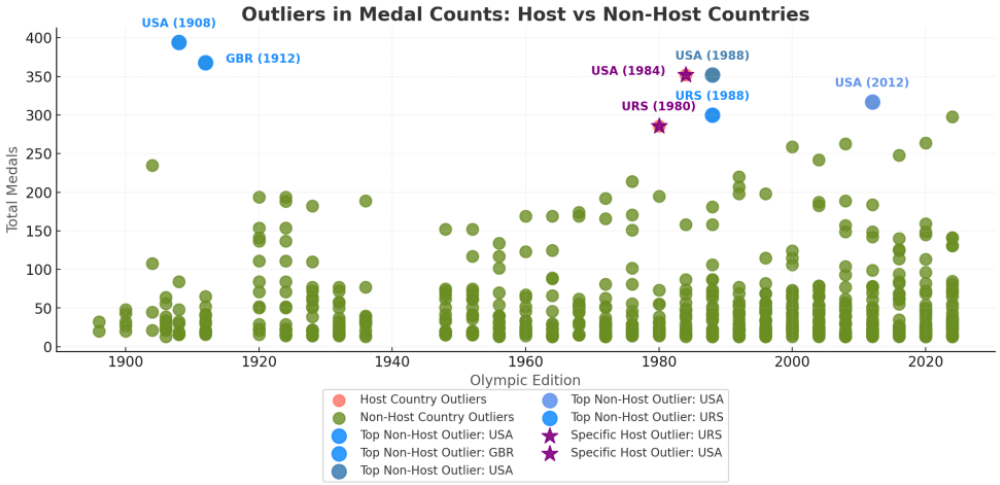


**Figure 6:** Outliers in Medal Counts: Host vs Non-Host Countries

Outliers among non-host countries are also worth noting. For example, the United Kingdom in 1912 and the United States in 1988 were both top non-host country outliers, which may be related to the outstanding performance of these countries in specific sports. At the same time, medal distribution may also be affected by the addition or cancellation of sports and related policies. For example, the

setting of certain events may be more favorable to the host country, and training methods, technological developments, and equipment upgrades may also change the distribution of medals.

## 5. Conclusions

This study constructs ridge regression and weighted random forest models to analyze and predict the 2028 Summer Olympics medal distribution. The ridge regression model solves multicollinearity, balances accuracy and interpretability, and provides reliable medal predictions. The weighted random forest predicts first-time medal-winning countries and explores impacts of event characteristics and the "host effect". Model analysis shows host countries gain medal advantages, event types correlate with medal gains, and medal distribution is gradually balancing between traditional powers and emerging countries.

Beyond model predictions, this study reveals historical insights: early medal distribution was concentrated, but became balanced with more participating countries. Special factors like the 1980-1984 Olympic boycotts caused medal outliers. Traditional powers such as the U.S. and the U.K. maintain competitiveness through sufficient funding, infrastructure, and training systems.

Future research can deepen this field by fusing deep learning with traditional statistical models to optimize prediction accuracy. It can also incorporate variables like host climate and technology impacts, and track event adjustments to update analytical frameworks, providing better references for the IOC and relevant countries.

## References

[1] Liu H, Huang M, Zhang J. Predicting Olympic medal counts using machine learning techniques: An analysis of historical data and athlete performance[J]. International Journal of Sports Science & Coaching, 2021, 16(4): 739-752.

[2] Guo X, Ding Y, Wu G. Recent Advances in Ensemble Learning: Techniques, Applications, and Challenges[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 5470-5489.

[3] Roth C, et al. Bayes beats Cross Validation: Efficient and Accurate Ridge Regression via Expectation Maximization[J]. arXiv preprint arXiv:2310.18860, 2023.

[4] Yamada T, et al. Refined Penalized Ridge Regression: Novel Methods for Optimizing Regularization Parameter in Genomic Prediction[J]. G3: Genes, Genomes, Genetics, 2024.

[5] Winham S J, Freimuth R R. A weighted random forests approach to improve predictive performance[J]. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2013, 6(6): 421–429.

[6] Daho S, Boucheham B, Settouti N. Improving random forests using weighted voting for imbalanced data classification[C]//Proceedings of the 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014). Halifax: Elsevier, 2014: 64–71.

[7] Guo X, Ding Y, Wu G. Recent Advances in Ensemble Learning: Techniques, Applications, and Challenges[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 5470-5489.

[8] Zhang C, Ma Y. Ensemble Machine Learning: A Review[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020, 10(3): e1356.

[9] Liu J, Chen S, Tan Y. Stacking Ensemble Approach for Enhancing Medical Image Classification[J]. Applied Soft Computing, 2022, 114: 108132.

[10] Gajowniczek K, Ząbkowski T, Buda P. Adjusting random forest for imbalanced churn prediction by using weighted voting[J]. Expert Systems with Applications, 2020, 160: 113696.

[11] Balmer N J, Nevill A M, Williams A M. Host nation advantage in Olympic Games[J]. Journal of Sports Sciences, 2023, 21(6): 469-478.

[12] Magee J C, Sugden J. The Olympic Games as a global event: Major issues and controversies[J]. International Journal of the History of Sport, 2002, 19(5): 1-30.