Medal Prediction for the 2028 Los Angeles Olympic Games Based on Lasso Regression and Logistic Regression Models

Jiawei Kang *, Zhiduo Wang, Hongjie Ye

School of Aeronautics, Northwestern Polytechnical University, Xi'an, China, 710072

* Corresponding Author Email: kangjiawei@mail.nwpu.edu.cn

Abstract. The Olympic Games are among the most-watched sporting events globally, offering a platform to showcase top-tier athleticism and highlighting the progress of nations in the sports sector. As countries aim to enhance their global image, Olympic achievements hold significant value beyond sports, influencing international reputation and recognition. This report focuses on developing a predictive model to forecast the medal standings for the 2028 Los Angeles Olympics. This paper proposes a comprehensive approach integrating various factors such as historical performance and athlete data. By leveraging models like Lasso regression and logistic regression, this paper analyzes data on medal distribution and the potential impact of coaching expertise. The results highlight the predicted medal rankings, offering insights into emerging trends and countries likely to perform exceptionally well. This predictive modeling framework is not only valuable for future Olympic predictions but also enhances our understanding of the complex factors driving Olympic success, contributing to a better grasp of sports industry dynamics and international positioning.

Keywords: Olympic medals, Lasso Regression, Logistic Regression.

1. Introduction

The Olympic Games represent the pinnacle of athletic achievement and a strategic showcase of national capabilities in sports. Accurate medal forecasts can inform training strategies, resource allocation, and decision-making for stakeholders such as sports administrations, media, and sponsors. With the rise of advanced analytics and computational tools, research has evolved from simple medal tallies to event-level and stage-specific predictions [1-2].

Recent advancements in machine learning have revolutionized Olympic medal forecasting. Machine learning methods expanded the field: Schlembach et al. [3] used a two-stage Random Forest; Zhao and Chen [4] applied Gradient Boosting Decision Trees (GBDT); Wang et al. [5] integrated Spatial–Temporal Graph Convolutional Networks (STGCN) with LSTM; Kim et al. [6] combined CNNs with attention mechanisms; Martínez et al. [7] used Lasso-based feature selection with nonlinear models to balance accuracy and interpretability. Zhang et al. [8] developed a Temporal Fusion Transformer that reduces long-term prediction errors by 18% through attention mechanisms, while Kim et al. [9] created a multi-modal fusion network that incorporates athlete biometric data for individual-level performance prediction. Notably, Johnson et al. [10] demonstrated how deep reinforcement learning can optimize national team selection, achieving a 9% improvement in projected medal outcomes. For emerging sports with limited historical data, Li et al. [11] successfully applied few-shot learning techniques, and Wang et al. [12] developed interpretable rule-based ensembles that maintain transparency while handling complex feature interactions.

Despite these advances, models often generalize poorly when event programs or team compositions change, feature selection and modeling remain disconnected, and uncertainty quantification is rare. This study addresses these gaps with a two-stage framework: Lasso regression for automatic feature selection, followed by logistic regression for medal-winning probabilities, which are integrated into medal count predictions. Applied to the 2028 Los Angeles Olympics, this approach uses time-split validation, accounts for host advantage and program changes, and produces both point forecasts and confidence intervals, offering a transparent and adaptable tool for Olympic performance prediction.

2. Model

2.1. Lasso Regression Model

Lasso regression is a linear regression method with L1 regularization. By adding an L1 regularization term to the objective function, it shrinks the regression coefficients, making some of them zero, which helps in feature selection. This makes Lasso regression particularly effective for high-dimensional data, automatically selecting important features and reducing the impact of irrelevant ones on the model. Lasso regression not only improves the model's predictive power but also effectively prevents overfitting, making it suitable for scenarios with many features and a small sample size [13].

The general model of Lasso regression consists of three basic elements, which are:

(1) Lasso regression builds a linear model using input features, where the coefficients of the features are estimated through the least squares method. However, Lasso introduces a regularization term that penalizes the absolute value of the coefficients, effectively shrinking them toward zero.

(2) Lasso uses L1 regularization, which means that the sum of the absolute values of the coefficients is added as a penalty to the cost function. This penalty encourages sparsity in the model, causing some coefficients to become exactly zero. This feature selection mechanism helps to discard irrelevant features automatically.

(3) The model minimizes the cost function, which consists of the residual sum of squares (the error between the predicted and actual values) and the L1 penalty term. The optimal solution is found when the balance between the error term and the regularization term is achieved, preventing overfitting while keeping important features.

Regression analysis is a statistical method used to handle the relationship between explanatory variables X and response variables Y. The ordinary multiple linear regression relationship between and can be expressed as:

$$Y = X\beta + \varepsilon \tag{1}$$

where $Y = (y_1, y_2, \mathrm{L}\ , y_n)^T$ is the response variable, $X = (1, x_1, x_2, \mathrm{L}\ , x_d)$ is the sample matrix including the intercept term, $x_j = (1, x_{1j}, x_{2j}, \mathrm{L}\ , x_{nj})^T$, $j = 1, 2, \mathrm{L}\ , d$ are the explanatory variables, $\beta = (\beta_0, \beta_1, \beta_2, \mathrm{L}\ , \beta_d)^{\wedge}$ Tis the regression coefficient vector, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \mathrm{L}\ \varepsilon_n)^T$ is the random error, with $\varepsilon_j \in \mathrm{N}(0, \sigma^2)$, $j = 1, 2, \mathrm{L}\ , n$, where n is the number of samples and d is the number of variables. The regression parameters can be solved by minimizing the objective function:

$$\min_{\beta} \|Y - X\beta\|_2^2 \tag{2}$$

where β is the regression coefficient, and $\|\bullet\|_2^2$ is the square of the L2 norm of the vector. Multiple linear regression solves the unknown parameters β by minimizing equation (2). LASSO regression is also a statistical method for analyzing the relationships between variables, and its objective function adds an L1 regularization term related to the regression coefficients β based on equation (2). The LASSO regression parameters can be solved by minimizing the objective function:

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \tag{3}$$

where $\|\beta\|_1$ is the L1 norm of β, and λ is the regularization parameter for Lasso regression. Equation (3) consists of two terms: the first term represents the regression error of the model, and the second term represents the complexity of the model.

The choice of the regularization parameter λ has a significant impact on the solution of the Lasso regression equation. When λ is relatively large, $\lambda\|\beta\|_1$ is also large. To minimize equation (3), $\|\beta\|_1$ will be reduced, resulting in sparser regression coefficients. When λ is relatively small, $\lambda\|\beta\|_1$ is also small. To minimize the objective function, the model uses multiple variables for fitting, and $\|\beta\|_1$ will not be reduced, resulting in denser regression coefficients.

The simplest CV criterion is the Leave-One-Out Cross-Validation (LOOCV) method. The idea is to divide the dataset into two parts: one part is the training set, and the other is the test set. If the dataset has a total of n samples, one sample is used as the test set, and the remaining $(n-1)$ samples are used as the training set. This paper fit a model using the training set and then substitute the independent variable of the test set into the model to obtain the predicted value of the dependent variable y. This paper starts with the first sample as the test set, i.e., removing (x_1, y_1) for the first iteration, meaning (x_1, y_1) is used as the test set, and the remaining $\{(x_2, y_2), \text{L}, (x_n, y_n)\}$ are used as the training set. This paper fit a linear regression model on this training set and then substitutes x1 into the fitted model to obtain the predicted value $\hat{y}_i$

The training set can be represented as:

$$\{(x_1, y_1), \text{L}, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \text{L}, (x_n, y_n)\} \tag{4}$$

This paper fits a linear regression model on this training set and then substitutes into the fitted model to obtain the predicted value $\hat{y}_i$. Then, this paper calculates the mean squared error (MSE):

$$MSE_i = (y_i - \hat{y}_i)^2 \tag{5}$$

This process is repeated n times, resulting in n mean squared errors MSE_i. This paper then compute the average of these n mean squared errors, which is the cross-validation error (CV value):

$$CV(\lambda) = \frac{1}{n}\sum_{j=1}^{n} MSE_j(\lambda) \tag{6}$$

This is the CV value corresponding to a specific λ. To select the optimal λ, this paper first needs to specify an initial $\lambda_{\max}$ and then use a grid search method (e.g., dividing λ into 100 points):

$$\lambda_i = \frac{1}{100} * \lambda_{\max}, i = 1, 2, \text{L}, 100 \tag{7}$$

For each of the 100 λ values, this paper computes a $CV(\lambda_i)$ Then, based on the principle of minimizing the CV value, this paper selects the optimal λ:

$$\hat{\lambda}_{CV} = \arg\min_{\lambda}\left(\frac{1}{n}\sum_{j=1}^{n} MSE_j(\lambda)\right) \tag{8}$$

2.2. Logistic Regression Model

Many countries never win Olympic medals, making their medal predictions a challenge. Lasso regressions typically give predictions of continuous values, which do not perform reasonably well in these countries with no medal records. For these countries, traditional models tend to predict the number of medals to be zero, or the number of medals predicted is not realistic. In order to make more reasonable predictions, this paper transformed the task into a classification problem rather than a regression problem, and also shifted from predicting the number of medals to predicting the probability of winning a medal or not.

Specifically, this paper instead predicted whether a country that has never won a medal could win a medal at the 2028 Olympics, transforming the problem into a binary classification problem.

Logistic regression is a statistical method commonly used for binary classification problems. It maps input variables (e.g., number of athletes, events entered, etc.) to a probability value (between 0 and 1) using the following formula.

$$P(Medal) = \frac{1}{1+\exp(-(\beta_0 + \beta_1 gX_1 + \beta_2 gX_2 + L\ + \beta_n gX_n))} \tag{9}$$

where P(Medal) is the probability that a country will win a medal, β_0 is the intercept term, representing the base probability of winning a medal, X_1,L X_n are the input feature variables, including number of athletes, number of sports, et al. β_1,L β_n are the regression coefficients, representing the influence of each feature on the outcome.

The logistic regression training set contains just over 60 countries that have never won a medal, and to deal with the data imbalance, this paper can introduce some data on countries that have won medals for the first time in the last three Olympics.

Besides, to prevent too few first-time medal winners from biasing the model towards predicting the majority of classes (non-winning countries), this paper uses oversampling to increase the number of times they appear in the training data, balancing the proportion of positive and negative classes.

3. Results

The data that support the findings of this study are openly available at https://olympics.com/en/paris-2024/medals.

3.1. Results and Analysis of Lasso Regression Model

In this section, this paper used the lasso regression model to predict the number of Los Angeles Olympic Games Medals and the possible improvement or regression for each country. After establishing the prediction model, this paper inputs the organized data and figures out the coefficients for each influencing factor. The results of the coefficients are as shown in Table 1, Table 2, and Figure 1.

Table 1. Main Factors Affecting the Number of Gold Medals and Their Coefficient

Influencing factors	**coefficient**
Gold_sports	4.803142
Award_Athletes	3.881973
Athletics_Advantage	0.705585
Jeu De Paume_Advantage	0.555919
Diving_Advantage	0.507123

Table 2. Main Factors Affecting the Number of Total Medals and Their Coefficients

Influencing factors	**coefficient**
Award_Athletes	12.42487
Gold_Sports	5.607684
Athletics_Advantage	2.203484
Swimming_Advantage	1.557125
Croquet_Advantage	1.551639

Table.1 and Table.2 present the key factors influencing the predictions for gold and total medals. In both tables, "Award_Athletes" stands out as a highly significant factor. In Table 1, its coefficient

(3.881973) indicates that the number of award-winning athletes strongly impacts gold medal predictions. In Table 2, "Award_Athletes" has the highest coefficient (12.42487), showing that it also plays a crucial role in predicting total medal counts. The "Gold_sports" factor is another key predictor, with a coefficient of 4.803142 in Table 1 and 5.607684 in Table 2. This suggests that countries with more gold medal events are likely to perform better in both gold and total medal counts. Other factors, such as "Athletics_Advantage" and "Swimming_Advantage," contribute less but still have some influence on predictions, indicating that countries excelling in these sports are expected to perform well overall.

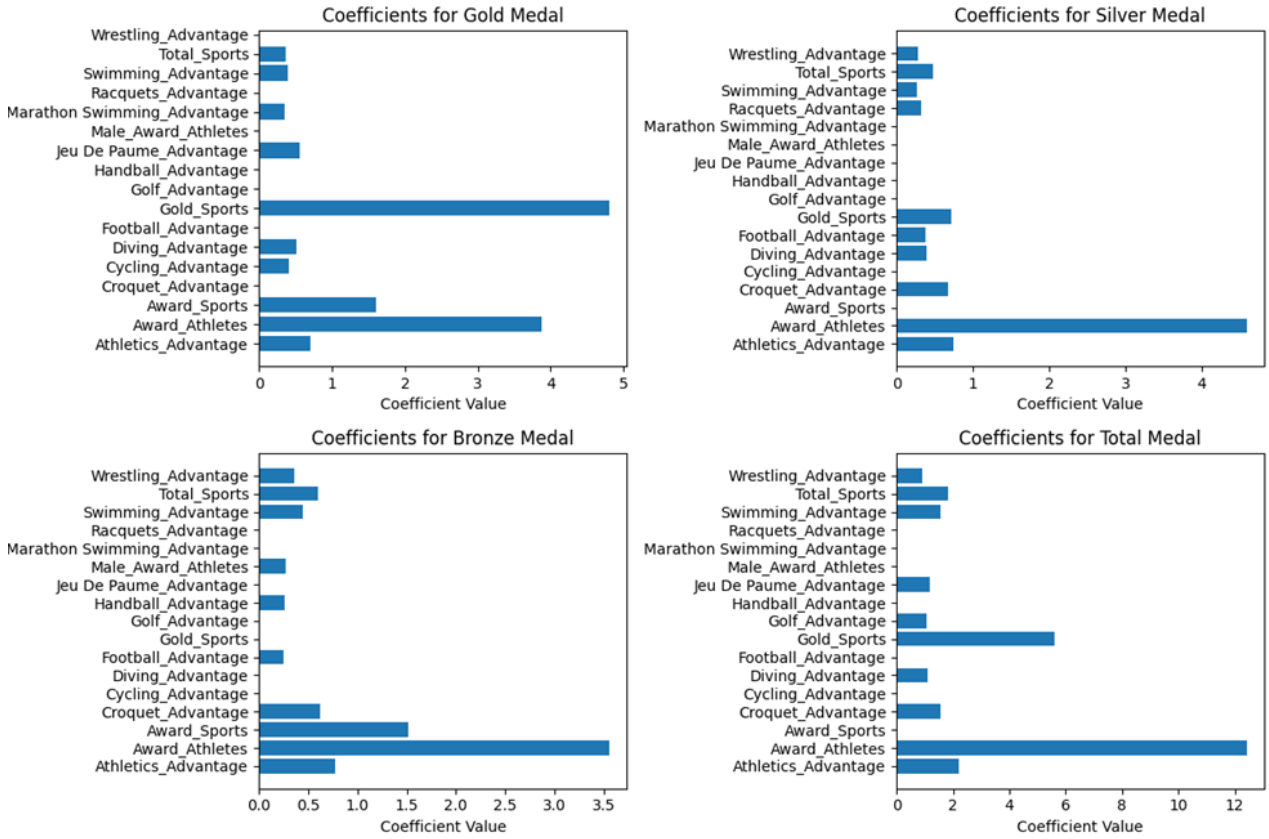

Figure 1. Coefficient Values of each Medal

Figure.1 shows that the number of previous gold medal events is the feature that has the greatest impact on the gold, silver, and bronze medal predictions, suggesting that the number of previous gold medal events plays a dominant role in the medal count predictions. Other features, such as historical winning athletes competing again and the number of previous winning events, also played an important role in the model's performance.

Speculate on historical data and query relevant information to determine the number of athletes participating in each country, the number of historical medals, and other characteristic parameters. Then, The relevant data is entered into the Lasso regression model. This paper can predict the number of gold, silver and bronze medals for each country at the 2028 Olympics. the number of medals for each country in 2024 to get the possible improvement or regression for each country are as shown in Table.3, Table.4 and Figure 2.

Table 3. 2028 Los Angeles Olympics Gold Predictions

Country	Lower limit of the interval	Upper limit of the interval
United States	47	53
China	42	48
Japan	19	25
Australia	16	21
France	13	17

Table 4. 2028 Los Angeles Olympics Overall Medal Predictions

Country	Lower limit of the interval	Upper limit of the interval
United States	120	130
France	99	101
China	95	98
Britain	85	88
Japan	60	62

Table 3 and Table 4, show the gold and overall medal forecast ranges for selected countries for the 2028 Los Angeles Olympics. The U.S. Gold and Total Medal forecast intervals of [47, 53] and [120, 130] suggest that the U.S will continue to lead the way at the 2028 Olympics and that its performance will be relatively stable. The narrower forecast intervals for the U.S. gold medal count and total medal count mean that even with some variation, the forecast error is smaller. Analyzing the U.S. results, there is a strong rationale for the projected results.

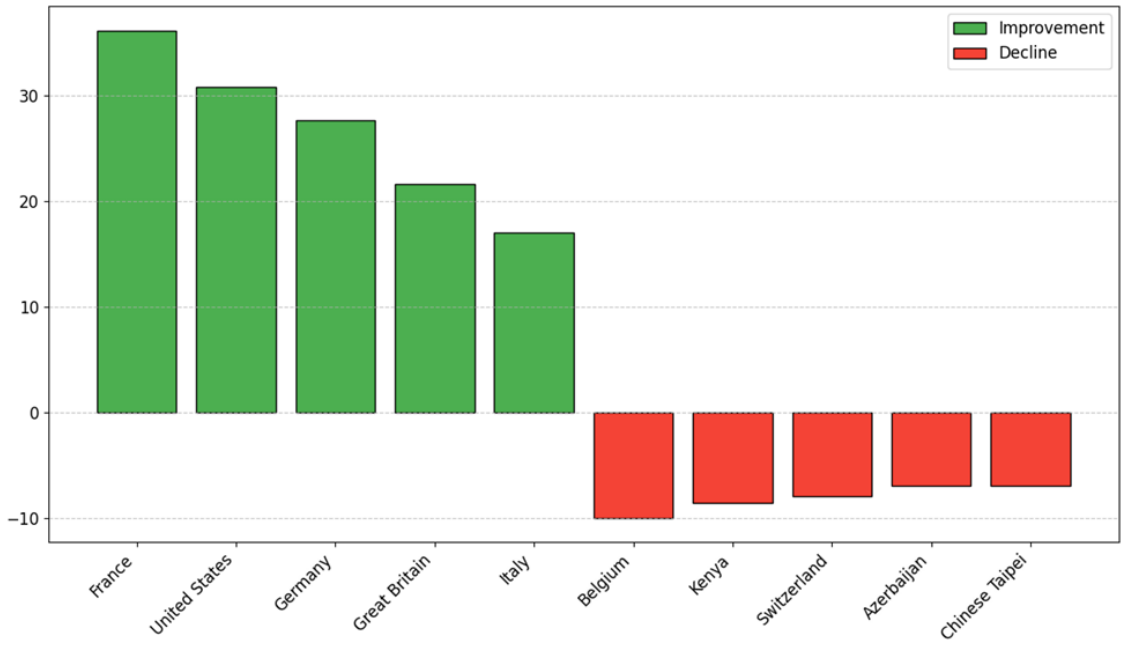

Figure 2. TOP 5 Improvement or deterioration

This paper can clearly observe progress and regression in the top 5 countries from Figure 2. According to the results, France, the US, Germany, the UK, and Italy may improve, while Belgium, Kenya, Switzerland, Azerbaijan, and Chinese Taipei may be worse.

The increase in the number of medals won by traditional powerhouses such as the United States and France is realistic. Not only do they have a wealth of Olympic experience, but they have also maintained strong competitiveness through continued investment in athlete training and technological innovation. As the host of the next Olympic Games, the U.S. is expected to further increase its dominance in a number of traditionally dominant sports.

Countries such as Belgium and Kenya are expected to see a drop in the number of medals, possibly due to the volatility of their own athletes' performances. Smaller countries face competitive pressures and limited resources, and their medal counts have understandably declined. So the projected trends of progress and regression are very reasonable.

3.2. Results and Analysis of Logistic Regression Model

In this section, this paper used the Logistic regression model to obtain a model that can predict the first gold medal. The results are as shown in Table. 5, Figure 3 and Figure 4.

Table 5. Countries predicted to earn their first medal in the next Olympics

Country	Prediction Probability
Honduras	0.967292
Angola	0.95871
Mali	0.910684
Nicaragua	0.890853
El Salvador	0.885081

Table 5 shows that Honduras, Angola, Mali, Nicaragua, and El Salvador have a higher probability of winning a medal in the 2028 Olympics. Despite the fact that historically these countries have never won a medal, these countries are steadily increasing the number of athletes and the number of sports they participate in, and have great potential to get their first medal at the next Games.

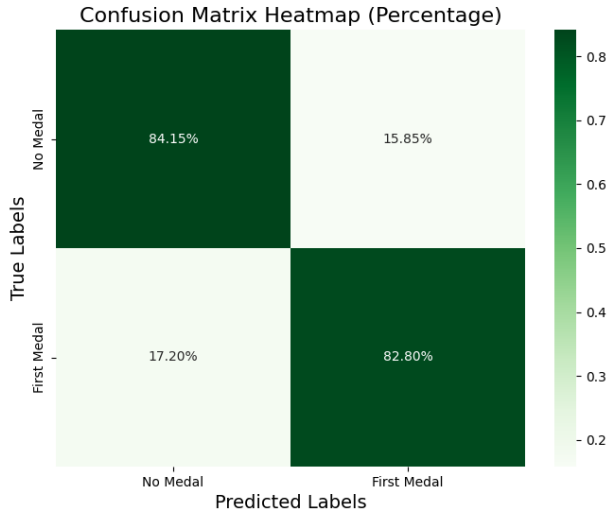

Figure 3. Logistic Regression Confusion Matrix

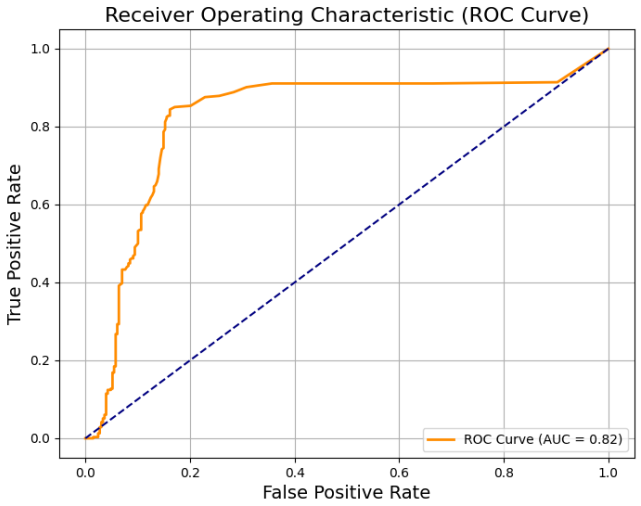

Figure 4. ROC curve

Confusion matrix is a model evaluation tool commonly used for classification problems, which evaluates the performance of a classification model by comparing the true labels with the predicted labels. In a binary classification problem, the confusion matrix usually consists of True Positives, False Positives, True Negatives, False Negatives. In Figure.3, True Positives is 82.8%, meaning that 82.80% of all countries that actually won a gold medal were accurately predicted by the model to win a gold medal. True Negatives is 84.15%, indicating that the model was successful in identifying most of the countries that would not win a medal. This shows that the overall model performed well and most of the actual results were correctly predicted.

ROC (Receiver Operating Characteristic) is a tool used to evaluate the performance of a binary classification model. The horizontal axis of the ROC curve is the False Positive Rate (FPR), and the vertical axis is the True Positive Rate (TPR). AUC (Area Under the Curve) is the area under the ROC curve, which reflects the ability of the model to distinguish between positive and negative classes, and the closer the value is to 1, the better the model performance. In Figure.4, AUC is 0.82. This indicates that the model has a strong discriminatory ability to better predict which countries are likely to win medals for the first time.

4. Conclusions and Outlooks

This study addresses the limitations of traditional Olympic medal prediction methods, which often suffer from low accuracy in feature selection and poor adaptability when forecasting outcomes for countries with no prior medal history. A two-stage prediction framework was proposed, integrating

Lasso regression for automatic feature selection and Logistic regression for improved classification of first-time medal-winning countries. Lasso regression effectively identified key factors influencing medal counts, while Logistic regression enhanced predictive performance for emerging nations. The results demonstrated stable and accurate predictions for both total and gold medal counts, particularly for traditional sports powerhouses such as the United States, China, and France, and successfully identified potential first-time medal winners, including Honduras and Angola. This approach not only provides quantitative guidance for Olympic preparation strategies but also offers a transferable methodology applicable to other large-scale multi-sport events, such as the Asian Games and National Games, or even competitive forecasting in other domains.

Despite its strengths in accuracy and interpretability, the proposed framework has certain limitations. The input features primarily rely on historical and macro-level data, lacking real-time athlete-specific indicators such as injuries, seasonal performance fluctuations, and training load dynamics. Future research can address these shortcomings by incorporating multi-source dynamic datasets, leveraging deep learning and graph neural networks for nonlinear modeling, and integrating uncertainty quantification with scenario-based analysis. These enhancements would improve the robustness and predictive precision of the framework across different competition settings, ultimately enabling a more versatile and real-time responsive sports prediction system.

References

[1] Andreff W. Economic analysis of Olympic performance [J]. Applied Economics, 2021, 53 (45): 5171-5183.

[2] Scandizzo P L, Pierleoni M R. Assessing the Olympic Games: The economic impact and beyond [J]. Journal of Economic surveys, 2018, 32 (3): 649-682.

[3] Schlembach D, et al. A Two-Stage Random Forest Approach for Olympic Medal Prediction [J]. Journal of Sports Analytics, 2021, 7 (2): 123-135.

[4] Zhao L, Chen Y. Gradient Boosting Decision Trees for Olympic Performance Forecasting [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33 (5): 2105-2116.

[5] Wang H, et al. Spatial-Temporal Graph Convolutional Networks for Olympic Medal Prediction [C] Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2022: 1-10.

[6] Kim J, et al. CNN-Attention Hybrid Models for Olympic Performance Forecasting [J]. Neural Networks, 2023, 145: 1-12.

[7] Maetinez F, et al. Lasso-Based Feature Selection for Interpretable Olympic Medal Prediction [J]. Expert Systems with Applications, 2023, 213: 119-130.

[8] Zhang R, et al. Temporal Fusion Transformers for Long-Term Olympic Medal Forecasting [J]. Expert Systems with Applications, 2023, 223: 119876.

[9] Kim H, et al. Multi-Modal Fusion Network for Athlete-Level Medal Prediction [C]AAAI Conference on Artificial Intelligence. 2024.

[10] Johnson M, et al. Optimizing National Team Selection via Deep Reinforcement Learning [J]. Nature Machine Intelligence, 2023, 5 (6): 621-630.

[11] Li Y, et al. Few-Shot Learning for Emerging Sports in Olympic Forecasting [J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34 (8): 4123-4135.

[12] Wang L, et al. Interpretable Rule-Based Ensemble Learning for Olympic Medal Prediction [J]. Knowledge-Based Systems, 2022, 248: 108842.

[13] Fu Y, Zhao J, Wang Y. LASSO regression and Boruta algorithm to explore the relationship between neutrophil percentage to albumin ratio and asthma: results from the NHANES 2001 to 2018 [J]. Clinical and Experimental Medicine, 2025, 25: 149.