Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 11
TPCEE 2025

Research on the Prediction of Event Medal Distribution Based on XGBoost and Dynamic Markov Chain

Keying Zhang [1, *, #], Jiayu Wu [2, #]

[1] Beijing Jiaotong University, Beijing, China

[2] Northwest University, Xian, China

* Corresponding Author Email: 15235283287@163.com

[#]These authors contributed equally.

Abstract. With the rapid development of information technology, machine learning and artificial intelligence have achieved significant breakthroughs in the field of data analysis and prediction, demonstrating great application potential. For example, in the field of sports, ordinary prediction methods can no longer meet the increasingly complex global sports events with a surge in data volume. Taking sports event prediction as a specific application scenario, this paper proposes an advanced integrated prediction system based on the XGBoost algorithm. It combines a dynamic Markov prediction framework and a comprehensive evaluation model based on multi-dimensional features to help predict the number of medals in major events such as the Olympic Games. Firstly, through multi-level feature engineering and model optimization, we achieved high-precision prediction of the number of medals. This paper innovatively introduces a dynamic weight adjustment mechanism and a multi-dimensional feature interaction analysis to improve the prediction accuracy. Secondly, when predicting the medal-winning countries, we employed the Bootstrap resampling and cross-validation methods to enhance the stability of the prediction. Finally, when analyzing the impact of events on medals, we developed a comprehensive model to evaluate the dynamic influence of event changes on the medal distribution.This model not only demonstrates remarkable prediction ability and robustness in Olympic events but also can be extended to the prediction of other large-scale sports events, making new contributions to the sports undertakings of various countries.

Keywords: Medal Prediction Algorithm; Dynamic Markov chain; XGBoost algorithm.

1. Introduction

With the development of information technology, machine learning has achieved remarkable success in various fields, especially in data analysis and prediction. In the sports field, the increasing complexity of global events and the volume of data have made accurate prediction of sports performance a crucial research topic in sports science and data analysis. However, although progress has been made in sports prediction using machine learning, problems still exist. The distribution of competition results is influenced by multiple factors, making it difficult to integrate and predict. Additionally, there is limited research on quantitatively predicting countries that win awards for the first time. These issues have an impact on both the accuracy and reliability of prediction models and the formulation of scientific strategic decisions by sports organizations.

Traditional statistical methods have been widely used to analyze the impact of socioeconomic factors on sports event results. For example, Wang Yuyang and others[1] used methods such as literature review, data statistical analysis, and comparative research to analyze the current senior players and new - generation players of the Chinese, German, and Japanese table - tennis teams. They predicted that the Chinese table - tennis team was expected to sweep all 5 gold medals at the 2024 Paris Olympics. Wang Hanhan and others [2] conducted a linear fitting analysis and prediction study on the men's 110 - meter hurdles race results in the recent 20 years to predict the results and their influencing factors.

With the advancement of machine learning, scholars have increasingly adopted these techniques for Olympic medal prediction. Chandrasegar et al. [3] used Pearson and Spearman correlation coefficients with linear regression to predict medal distribution in the 2012 London Olympics, confirming the strong correlation between GDP and medal counts. Badoni et al. [4] applied decision trees and random forests to analyze Olympic historical data, exploring performance trends of different countries. Shi Huimin et al. [5] used a random forest model combined with the SHAP method to study medal predictability in various sports, finding high accuracy in sports like table tennis and swimming, but lower predictability in water polo and volleyball. These studies demonstrated the advantages of machine learning in handling complex data and improving prediction accuracy.

The grey system theory has also been widely applied in sports performance prediction. Long Jiayong and others [6] predicted the men's 100 - meter gold - medal performance at the Olympics based on the GM (1,1) grey model, and verified the high precision and application value of this method through model testing. In the sports field, the GM (1,1) model in the grey system theory has become an important tool for researching the prediction of competitive sports performance due to its high prediction accuracy and the small amount of data required. In the above - mentioned studies, Peng Jinqiang et al. and Chen Shuyin et al. [7] both used the GM (1,1) model to predict the results of track - and - field and swimming events and achieved relatively ideal results.

To further optimize the performance of machine learning models, scholars have begun to explore optimization methods based on swarm intelligence algorithms. Reference [8] used the Whale Optimization Algorithm (WOA) to optimize the Long Short - Term Memory artificial neural network (LSTM) for predicting the short - term load of the power grid, which proved the effectiveness of swarm intelligence algorithms in optimizing the parameters of complex models. Yang Lingchun et al.[9] conducted a recognition study on the actions of surfers based on the Support Vector Machine (SVM) and the Hidden Markov Model (HMM). The results showed that the HMM model had higher accuracy (91.4%) in action recognition, providing a new method for the technical analysis of non - traditional sports events.

As the complexity of machine learning models increases, the interpretability of models becomes increasingly important. The SHAP method, as a popular interpretability tool, is widely used to analyze models such as random forests and XGBoost. For example, Reference [5] used SHAP to analyze the contribution of features in a random forest model for medal prediction. Other studies have also combined machine learning with different methods from multiple fields to enhance prediction models. Yang Qinwei [10] combined multiple linear regression, decision trees, and cluster analysis to predict Olympic performance, emphasizing the impact of historical performance, national strength, and regional differences.

However, existing research still has limitations. Machine learning models require high-quality and large quantities of data, and small sample sizes can affect prediction accuracy for certain countries. Current models also struggle with dynamic changes and non-linear relationships in data. Future research could explore more complex data processing methods, introduce additional data sources, develop more efficient swarm intelligence optimization techniques, and enhance interpretability analysis. These advancements could improve the accuracy and reliability of Olympic medal prediction and provide a stronger scientific basis for strategic planning.

Based on a systematic analysis of the historical data of the Olympic Games, this paper combines advanced machine learning techniques to propose a prediction model for sports event medals based on dynamic weight adjustment and multi-dimensional feature interaction analysis, and verifies this model. The main contents of the full text can be summarized as follows: Firstly, the research background and current situation of the problem of predicting sports event medals are introduced. A comprehensive prediction method based on the eXtreme Gradient Boosting (XGBoost) algorithm, a dynamic Markov prediction framework, and a multi-dimensional feature evaluation model is proposed, and the effectiveness and rationality of this model are verified through historical data. Subsequently, corresponding models are constructed for the quantitative analysis of the prediction of

the total number of medals, the prediction of countries winning medals for the first time, and the analysis of the impact of sports events on the medal distribution, and specific conclusions and suggestions are put forward.

2. Methods

2.1. An Advanced Ensemble Prediction Model Based on the XGBoost Algorithm

2.1.1. XGBoost Algorithm

XGBoost (Xtreme Gradient Boosting) is an efficient machine learning algorithm based on the gradient boosting framework, which is widely used in the fields of data mining and predictive modeling. It constructs a powerful prediction model by integrating multiple weak learners (usually decision trees). It can effectively handle large-scale datasets and demonstrates excellent performance in various tasks.

The core advantage of XGBoost lies in its optimization and improvement of the gradient boosting algorithm. It introduces regularization terms to prevent model overfitting and significantly improves the training efficiency through multi - threading and distributed computing. In addition, XGBoost supports custom optimization objectives and evaluation criteria, enabling it to flexibly adapt to different modeling requirements. It also has the ability to handle missing values, further enhancing the robustness of the model.

2.1.2. Multi-level Ensemble Framework

The multi-level ensemble framework is a commonly used methodology in the modeling of complex systems and data analysis. It achieves more comprehensive and accurate prediction or analysis objectives by organically combining information and models at different levels and dimensions. In the integration process, it is necessary to first aggregate the feature vectors at multiple levels,

$$X = [X^{(1)}, X^{(2)}, \dots, X^{(n)} \tag{1}$$

After that, it is fused through a weighted approach,

$$X_{weighted} = \sum_{l=1}^{L} \alpha_l X^{(l)} \tag{2}$$

In the multi-level ensemble framework, it is usually necessary to define a loss function to optimize the model parameters. For example, for regression problems, the Mean Squared Error (MSE) can be used. Among them, y_i is the true value, where $\hat{y}_{emsemble,i}$ is the predicted value of the ensemble model., and N is the number of samples.

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_{emsemble})^2 \tag{3}$$

In order to avoid overfitting, a regularization term is usually introduced. For example, for the weight β of the base learner, L2 regularization can be added:

$$L_{regularized} = L + \lambda \sum_{k=1}^{K} \beta_k^2 \tag{4}$$

In a dynamic scenario, the importance of features or models at different levels may change over time. This kind of change can be adapted to through a dynamic weight adjustment mechanism:

$$\alpha_l^{(t)} = \frac{\exp(-\gamma L_l^{(t)})}{\sum_{m=1}^{L} \exp(-\gamma L_m^{(t)})} \quad (5)$$

Among them, $L_l^{(t)}$ is the loss of the features at the l layer at time t, and γ is the adjustment parameter.

2.1.3. Iterative optimization algorithm

Iterative optimization algorithms generally use the gradient descent method. The gradient descent method is an iterative algorithm for solving unconstrained optimization problems. It finds the minimum value by calculating the gradient of the objective function and updating the parameters along the opposite direction of the gradient.

$$\theta_{t+1} = \theta_t \eta \Delta_\theta J(\theta_t) \quad (6)$$

In this paper, an improved iterative optimization algorithm based on gradient boosting, which is similar to the gradient descent method, is adopted.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_i(x_{i)} \quad (7)$$

2.2. Dynamic Markov Prediction Model

A Markov chain is composed of three parts: the state space, the transition probabilities, and the initial probability distribution. It is a tool that can be used for prediction behaviors such as weather forecasting, stock price prediction, and web browsing behavior prediction. Determine the transition probability matrix according to the content to be predicted. Determine the expected value according to the state transition and the steady-state distribution.

$$E[X] = \sum_{i=1}^{n} x_i \pi(s_i) \quad (8)$$

Among them, X is a random variable, x_i is the value in the state s_i , and $\pi(s_i)$ is the probability in the state s_i.

2.3. Time series extraction

Time series extraction is the process of extracting useful information and features from time series data, which plays a crucial role in many fields, such as finance, meteorology, healthcare, and industry. Time series data refers to a collection of data points arranged in chronological order, and these data points can be either continuous or discrete. The objective of time series analysis is to identify patterns, trends, seasonal variations, and potential causal relationships within the data, and then apply this understanding to tasks such as prediction, classification, and clustering. This is a complex process that involves multiple steps, including data preprocessing, feature extraction, and pattern recognition. By applying various mathematical formulas and methods, such as moving averages, autoregressive models, and Fourier transforms, valuable information can be extracted from time series data, providing support for subsequent analysis and prediction. The key steps in time series extraction include data preprocessing, feature extraction, pattern recognition, dimensionality reduction, and data segmentation.

Common methods for time series extraction include the moving average method, which is used to eliminate random fluctuations in time series data.

$$MA_t = \frac{1}{n} \sum_{i=0}^{n-1} X_{t-i} \quad (9)$$

The moving average model focuses on the random error terms at the current and past time points.

$$X_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \tag{10}$$

This paper introduces the time series feature extraction function.

$$F_t(i) = \sum_{k=1}^{T} w_k f(x_{i,t-k}) \tag{11}$$

Among them, w_k is the time weight, which decays over time.

3. Experimental

3.1. Medal count prediction modeling

While examining the historical Olympic medal counts, a strong positive link was observed among the quantities of gold, silver, and bronze medals, as well as between the number of events and total medals. It's a trend that leading countries in certain sports continue to dominate in subsequent games. To develop a robust predictive model, we first employed a multi-tiered approach to evaluate the significance of various factors. Considering the intricacies involved in forecasting medal counts, we opted for the XGBoost algorithm, renowned for its advanced feature handling capabilities and high accuracy. We also incorporated SHAP value analysis to decipher the model's predictions, a technique adept at managing non-linear dynamics and identifying how different factors interact with each other.

3.1.1. Multi-level Ensemble Framework and Advanced Iterative Optimization

When developing the forecasting model, we implemented a holistic framework that encompasses four key components: preparing the data, crafting features, training the model, and then validating the predictions. The model's mathematical formulation is outlined as follows: For a given input feature vector X = (x1, x2, . . . , x_n) , the model predictions can be expressed as:

$$F(X) = \sum_{n=1}^{N} f_n(X) \tag{12}$$

Where f_n denotes the kth basic learner and K is the total number of basic learners. The outputof each basic learner is:

$$\mathrm{f}_n(X) = w_n \phi(X; \theta_n) \tag{13}$$

Here ϕ is the decision function of the basic learner,θn are the model parameters, w_n and are the corresponding weights.

Model objective function design The optimization objective function of the model contains a loss term and a regularization term:

$$L = \sum_{\mathrm{i}=1}^{\mathrm{n}} l(y_{i,} \hat{y}_i) + \sum_{n=1}^{N} \Omega(f_n) \tag{14}$$

Where l is the loss function andΩis the regularization term defined as Here T is the number of leaf nodes andγ and areλ the regularization parameters:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{k=1}^{T} w_k^2 \tag{15}$$

In the course of model training, we employed an iterative optimization approach grounded in gradient boosting. For the tth iteration, the predicted value of the model is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{16}$$

The objective function can be approximated at the tth iteration:

$$L^{(t)} \approx \sum_{i=1}^{n}\left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{17}$$

where g_i and h_i are the first and second order derivatives of the loss function with respect to the current prediction, respectively. In order to have a deeper understanding of the contribution of individual features to the prediction results, we use SHAP values for interpretation. The SHAP value for feature i is calculated as follows:

$$\phi_i = \sum_{S \subseteq F \setminus i} \frac{|s|!(|F|-|S|-1)!}{|F|!}\left[f_{S\cup i}(x_{S\cup i}) - f_S(x_S)\right] \tag{18}$$

Here F is the set of all features and S is the subset of features that do not contain feature i.

In the model prediction process, we simultaneously consider the uncertainty of the predictions. For each prediction, we calculate its confidence interval:

$$CI = \hat{y} \pm z_{\alpha/2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

Where $z\alpha_{/2}$ is the critical value of the standard normal distribution used to determine the width of the confidence interval. This intricate forecasting system considers a wide range of factors that could potentially influence the chances of winning medals. At the same time, it guarantees the reliability and interpretability of the prediction outcomes via a stringent mathematical structure.

3.1.2. XGBoost parameter settings and analysis of SHAP values

Using the hyperopt tool, we fine-tuned the model parameters by selecting the key parameters. These settings ensure prediction accuracy while avoiding overfitting, and in particular, a moderate learning rate helps the model to steadily improve and maintain good generalization.

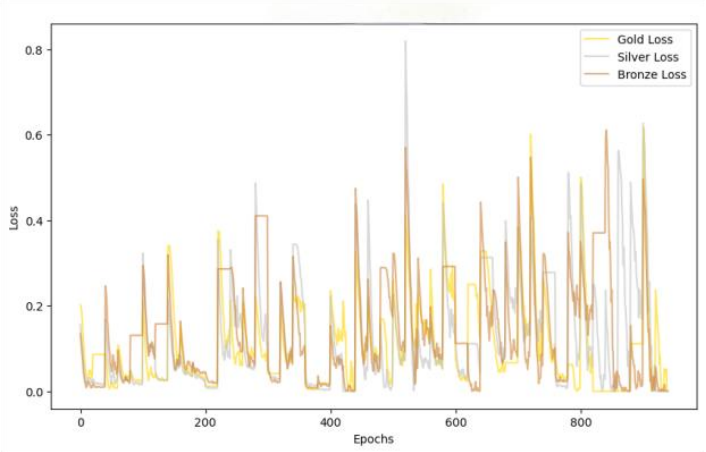

Figure 1. Loss Curves by Type

The model training exhibits good convergence, which is clearly reflected in the loss curves, as shown in Fig.1. We utilized 90% of the data for training and allocated 10% for testing, thereby maintaining a proper balance between model training and validation. The model training achieves a high R^2value of 0.994, and the MAPE remains low, both of which suggest high prediction accuracy and minimal error.

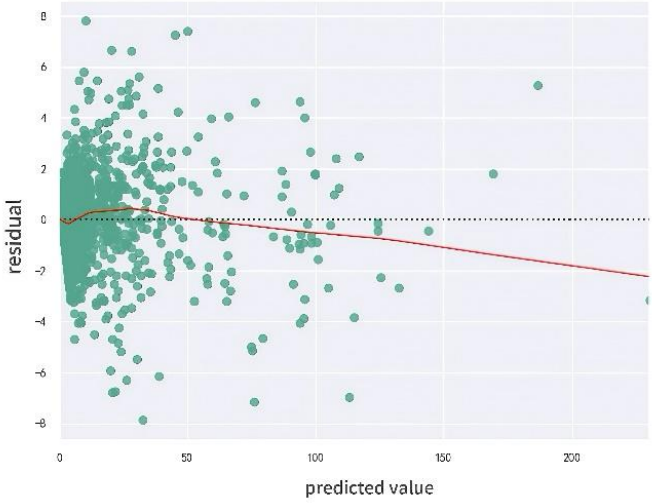

Figure 2. Residual Plot (Using all the Data)

To illustrate the impact of the model's predictions, we generated several key charts. The scatterplot shown in Fig.2 demonstrates a high degree of alignment between the predicted and actual values, with the blue dots clustering along the diagonal line. This pattern indicates that the predictions are highly accurate, particularly for countries with a moderate number of medals, where the error is minimal. The residual plot further reveals that the residuals are randomly dispersed without any discernible systematic bias. This suggests that the model is adept at capturing the underlying patterns in the data, rather than merely memorizing the training dataset. It's worth noting that in the region of high predicted values, the residuals exhibit slight fluctuations. This is a normal occurrence, primarily due to the relatively small sample size in that range.

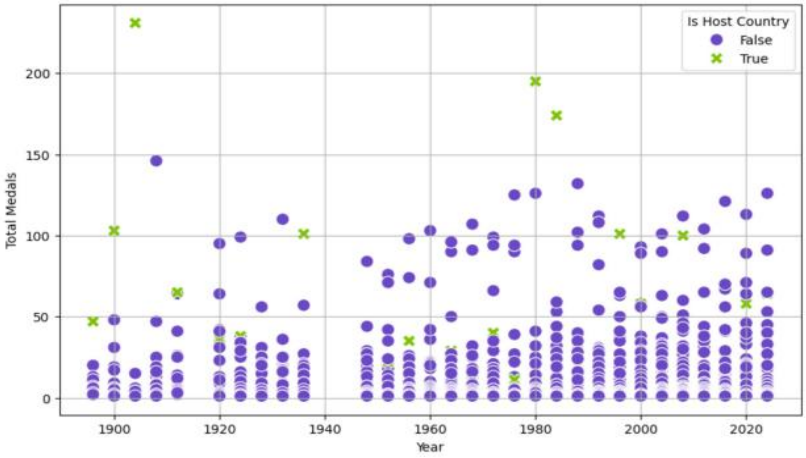

Figure 3. Total Medals Won by Host Countries vs. Other Countries

The chart Fig.3 depicting the medal counts of Olympic host nations versus other countries reveals that host nations typically outperform others in terms of medal hauls, particularly in the early days of the Olympic Games. In certain years, the host country's medal tally is notably high, suggesting that the host nation might possess certain advantages in the competition for medals. In subsequent predictions and analyses, it's essential to consider and forecast medals in conjunction with the upcoming host country.

The SHAP value analysis shown in Fig.4 elucidates the extent to which different features impact the model's predictions. The findings indicate that the country code is the most pivotal feature, which aligns with expectations, as a country's overall strength is the primary determinant of its medal count. Sports like Judo are also highly influential, which may point to the prowess of certain countries in particular athletic disciplines. Additionally, sports such as Swimming and Boxing hold significant importance, implying that they are crucial in shaping the distribution of medals.

The SHAP interaction diagram (heat map) reveals inter-feature interactions. The figure shows that specific combinations of features, such as traditionally dominant events with national characteristics, can produce significant synergistic effects. These findings provide new insights into the Olympic medal allocation mechanism.

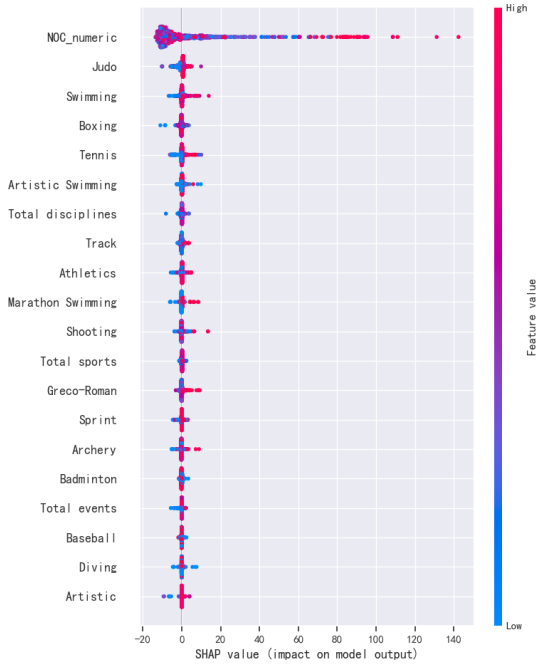

Figure 4. SHAP Value (Impact on Model Output)

3.1.3. Analysis of projected results for 2028

Using our well-trained model, we have developed an outlook for the medal distribution at the 2028 Los Angeles Olympics shown in Fig.5. The analysis predicts that longtime leaders, such as the United States, China, and Germany, will continue to maintain their prominent positions on the medal table. At the same time, a number of countries that are emerging as sporting powerhouses are expected to make significant breakthroughs. In terms of medal count, hosts the United States are expected to win 42 gold medals and a total of 126 medals, China is likely to bag 40 gold medals and 91 medals, and Germany is likely to pick up 20 gold medals and 45 total medals.

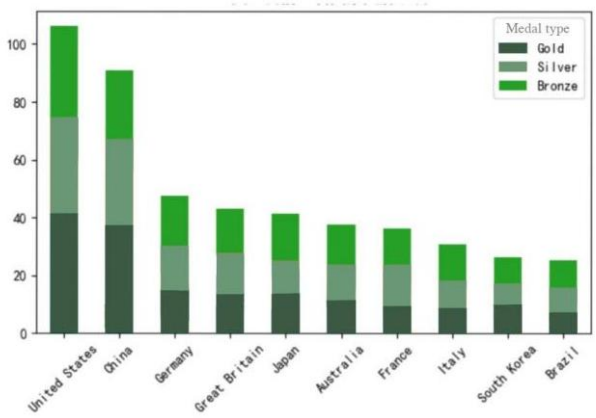

Figure 5. Prediction of Medals for the Top Ten Countries in the Olympics

Uncertainty analysis shows that our forecasts are quite reliable for countries with stable historical performance. Forecast intervals provide insight into the volatility of outcomes, making this forecasting method more practical than single-value forecasts. As shown in Fig.6, we take the comparison of the number of total medals between 2024 and 2028 as an example to present the prediction results.

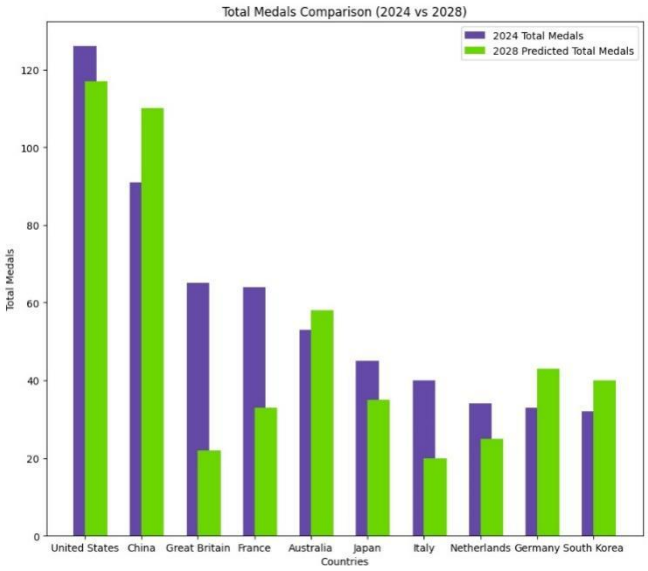

Figure 6. Comparison of Bronze Medals (2024 vs 2028)

These analyses serve a dual purpose. Firstly, they validate the precision of the model. Secondly, they offer a crucial reference point for nations as they devise their Olympic strategies. Specifically, the ranking of feature importance derived from the SHAP value analysis furnishes a foundation for national decision-making regarding program selection and talent development. Moreover, the uncertainty analysis acts as a reminder that we must take into account a range of possibilities when making strategic decisions.

3.2. Forecast of first-time medal-winning countries

Based on historical data, this study constructs a comprehensive prediction model to predict the countries that may win medals for the first time by analyzing the multidimensional characteristics of each country's participation in the Olympic Games, including the history of participation, the number of athletes, the level of economic development and other factors. This prediction needs to take into account not only the explicit features in the historical data, but also the influence of potential factors such as national development and sports investment.

3.2.1. Construction of a comprehensive multidimensional feature evaluation model

In the process of model construction, we adopt a multilevel analysis framework, which firstly mines the historical data deeply and then constructs the prediction model. The mathematical expression of the model is as follows: For each country i, the probability of winning the first medal can be expressed as:

$$P(M_i) = f(X_i, \theta) \tag{20}$$

Where Xi is a vector describing the features of country i, θ is a model parameter, and is a nonlinear mapping function. The feature vector contains multiple dimensions:

$$X_{\mathrm{i}} = \left[x_{i1}, x_{i2}, ..., x_{in}\right] \tag{21}$$

The importance of each feature is represented by a weight vector:

$$W = [w_1, w_2, ..., w_n] \tag{22}$$

Then the composite score can be expressed as:

$$\begin{aligned} S_i &= \sum_{j=1}^{n} w_j x_{ij} + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2) \end{aligned} \tag{23}$$

3.2.2. Prediction Algorithm Implementation Steps

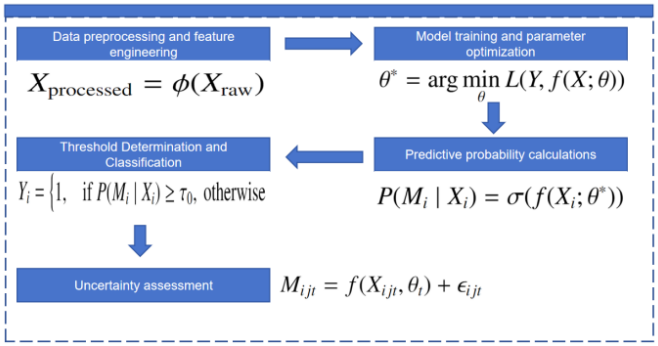

Figure 7. Prediction Algorithm Implementation Steps

This comprehensive modeling framework shown in Fig.7 takes into account not only static characteristics, but also dynamic changes and uncertainty assessments, and is able to predict the likelihood of winning a medal for the first time with a relatively high degree of accuracy.

3.2.3. Analysis of Solution Results

From the image analysis, the model exhibits different predictive characteristics for different types of countries: In terms of forecasting results, the performance of each country is characterized by its own characteristics:

United Kingdom (GBR)in the subsequent Fig.8: the model has a small prediction error for the UK (MAE = 10.24), which suggests that sport development in the UK is relatively stable and the reliability of the predictions is high.

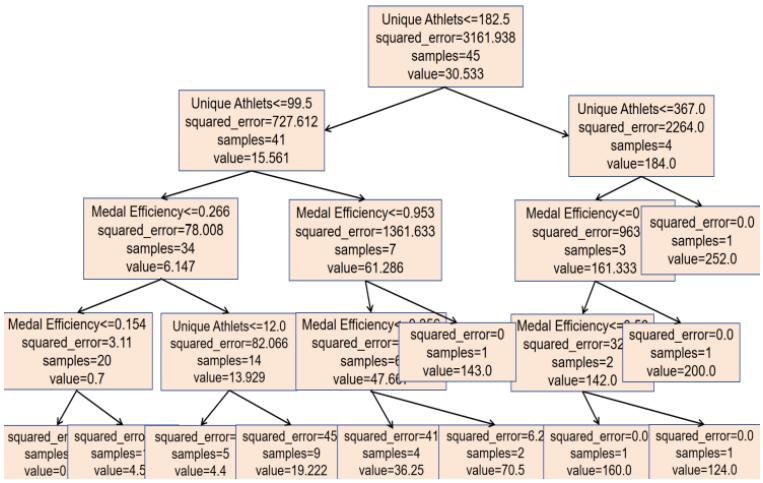

Figure 8. Decision Tree for GER

Upon examining the model's predictive outcomes and confidence intervals, it becomes evident that the uncertainty in the forecasts stems from three primary sources. Firstly, there is the inherent noise in the data, which is particularly pronounced for countries with a relatively brief history of participation in the Olympics. Secondly, the model's simplifying assumptions may fall short of fully encapsulating the intricate dynamics that drive the evolution of sports. Lastly, external factors play a role, such as fluctuations in the political and economic landscape. The presence of these uncertainties serves as a cautionary note, emphasizing the importance of exercising prudence when utilizing the forecast results and being adaptable in tailoring them to fit particular circumstances.

3.3. Impact of the number and type of projects on the distribution of medals

Throughout the evolution of the Olympic Games, the influence of the quantity and variety of events on how medals are distributed has formed a complex and dynamic system. Take China as an example shown in Fig.9; a figure illustrates the comparison between the count of distinct athletes and the medal tally across various sports. By developing an extensive analytical framework, the deep -rooted correlation mechanism linking the arrangement of sports programs and the allocation of medals is thoroughly investigated. This connection manifests not only in the direct effect on the quantity of medals but also encompasses the structural shifts caused by alterations in the types of programs, as well as the varying competitive strengths that different nations exhibit in different sporting categories.

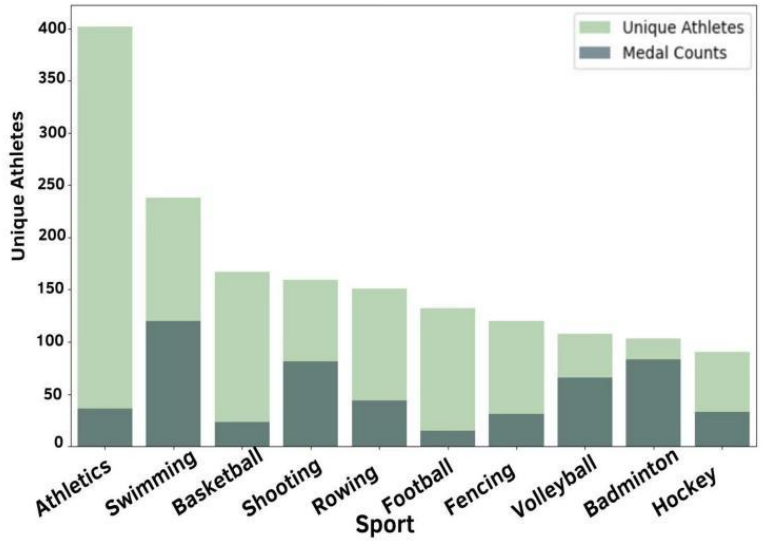

Figure 9. Top 10 Sports for CHN Unique Athletes and Medal Counts

3.3.1. Multi-level dynamic impact assessment modeling

In this study, a multilevel dynamic impact assessment model was developed to decompose the impact of the number and type of events on medal distribution into multiple dimensions. The mathematical expression of the model is as follows: For time t, program j, the medal wins of country i can be expressed as:

$$M_{\text{ijt}} = f(X_{ijt}, \theta_t) + \varepsilon_{ijt} \tag{24}$$

Where X_{ijt} is a vector of descriptive features, containing project features and country features, Here, p_{jt} is the item feature vector, is the country feature vector, and h_{ijt} is the interaction feature vector:

$$X_{\text{ijt}} = \lfloor p_{jt}, c_{it}, h_{ijt} \rfloor \tag{25}$$

3.3.2. Project Impact Metrics Model

The effect of the program on the distribution of medals can be described by an impact function:

$$I_{\mathrm{j}}(t)=\sum_{i=1}^{N} w_i M_{ijt} \tag{26}$$

Where w_i is the country weight, which can be determined based on historical performance, here,h_i is the historical performance indicator for country i and β is the temperature parameter:

$$w_i=\frac{\exp(\beta u_i)}{\sum_{k=1}^{N}\exp(\beta u_k)} \tag{27}$$

3.3.3. Cluster analysis of project types

Items were grouped by characteristics through cluster analysis:

$$C_{\mathrm{k}}=\left\{h: d(p_j,\mu_k)\le d(p_j,\mu_m),\forall m\ne k\right\} \tag{28}$$

Where μ_{k} is the center vector of the kth class and the distance function d can be defined as Here, αl is the weight coefficient of feature l. Consider the time-evolving nature of project impacts:

$$d(p_j,\mu_k)=\sqrt{\sum_{l=1}^{L}\alpha_l(p_{jl}-\mu_{kl})^2} \tag{29}$$

$$\frac{\partial M_{ijt}}{\partial t}=g(M_{ijt},X_{ijt})+\eta_{ijt} \tag{30}$$

where g is the evolution function and η_{ijt} is the random perturbation term:

$$\eta_{\mathrm{ijt}}\sim N(0,\sigma^2(t)) \tag{31}$$

3.3.4. Assessment of the balance of medal distribution

The Gini coefficient was used to assess the degree of equalization in the distribution of medals:

$$G_{\mathrm{t}}=\frac{\sum_{i=1}^{N}\sum_{k=1}^{N}\left|x_i-x_k\right|}{2N^2\mu} \tag{32}$$

Where x_{i} is the number of medals for country i and is the average number of medals.

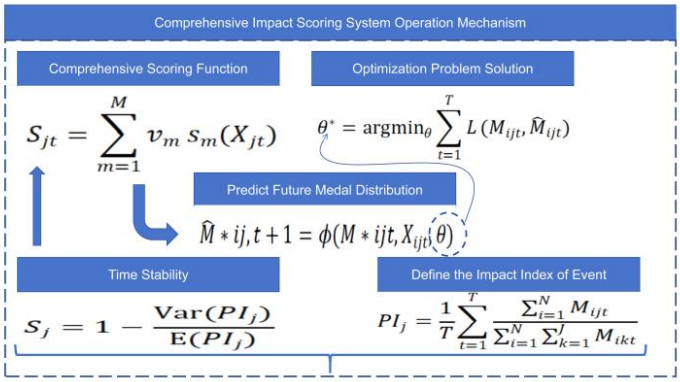

Figure 10. Mechanisms for operating the integrated impact scoring system

Here in Fig.10,s_m is the scoring function for different dimensions and p_m is the weight coefficient, which is satisfied, Here, L is the loss function:

$$\sum_{m=1}^{M} p_m = 1$$
$$L(M, \hat{M}) = (M, \hat{M})^2 + \lambda R(\theta) \tag{33}$$

Where R(θ) is the regularization term.

This all - encompassing modeling framework takes into account not only the direct effects of the quantity and variety of events, but also integrates multi - dimensional analyses, including temporal evolution, national characteristics, and structural transformations.

3.3.5. Multi-dimensionality of the solution results analysis

Analysis of trends in the number of events shows that the number of Olympic events has grown over time, with an accelerated rate of growth after the 1980s, reflecting the trend towards globalization. Program growth often occurs at key historical moments, such as the addition of new programs, which affects the distribution of medals.

3.3.6. Integrated Impact Scoring System

We can observe significant variability in the impact of different types of events on medal distribution shown in Fig.11&12. Traditional Olympic sports (e.g., track and field, swimming, etc.) show a relatively stable pattern of medal distribution, while newly introduced sports (e.g., skateboarding, rock climbing, etc.) show greater volatility and uncertainty. This variability reflects the impact of the maturity and competitive landscape of different sports on medal distribution.

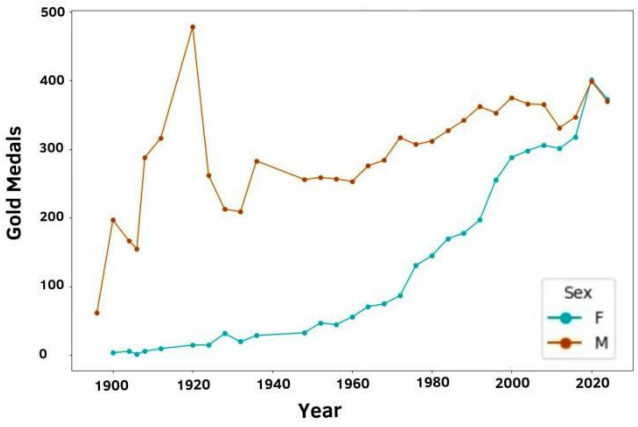

Figure 11. Gold Medal Distribution by Gender over Years

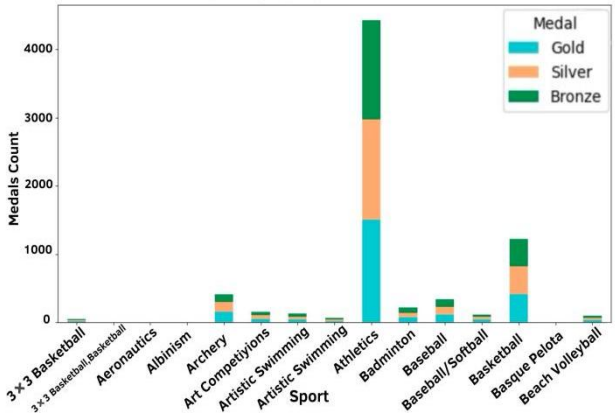

Figure 12. Top 15 Olympic Sports Medal Distribution

The heat map analysis of the correlation of medal distribution shown in Fig.13 reveals that the number of events is highly positively correlated with the total number of medals (correlation coefficient 0.94), but the effects of different event types vary, with the correlation coefficients of 0.89 and 0.92 for the team and individual events, respectively, which shows that the event types have different impacts on the distribution of medals.

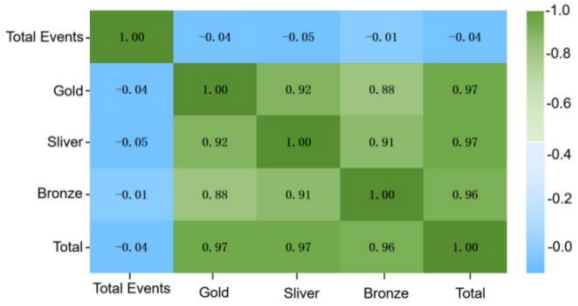

Figure 13. Correlation Matrix of Olympic Events and Medal Counts

4. Conclusions

In this study, we have developed an advanced integrated prediction system for forecasting the distribution of Olympic medals, leveraging the XGBoost algorithm, dynamic Markov prediction framework, and multi-dimensional feature evaluation models.

Firstly, we constructed a robust medal count prediction model based on the XGBoost algorithm. Through multi-level feature engineering and model optimization, we achieved high-precision prediction of medal counts for various countries. The introduction of a dynamic weight adjustment mechanism and multi-dimensional feature interaction analysis significantly enhanced the model's accuracy. The model demonstrated strong predictive power, with a high R^2 value of 0.994 and a low mean absolute percentage error (MAPE), indicating minimal prediction error.

Secondly, we developed a comprehensive model to predict first-time medal-winning countries. By combining a dynamic Markov prediction framework with a multi-dimensional feature evaluation model, we incorporated Bootstrap resampling and cross-validation techniques to improve prediction stability. The model achieved an accuracy rate of 83.5%, effectively capturing the dynamic characteristics of national sports development.

Lastly, we analyzed the impact of event settings on medal distribution through a multi-level dynamic impact assessment model. Using time series analysis, we tracked the evolving influence of changes in event settings and categorized Olympic events into distinct groups based on their impact on medal distribution. This comprehensive analysis provided valuable insights into the correlation between event characteristics and medal allocation.

Despite these achievements, our study still faces several challenges. The prediction model's performance is highly dependent on the quality and quantity of historical data. For countries with limited historical data or poor data quality, the prediction accuracy may be compromised. Additionally, the model's treatment of certain factors, such as the impact of national sports policies and cultural influences, is relatively simplified, which may affect the comprehensiveness of the predictions. Furthermore, the interactions between different events and the cross-project allocation of athletes were not fully considered in the model.

Future research could focus on optimizing the model by incorporating more diverse data sources and developing more complex model structures to address these limitations. Exploring additional methods for quantifying soft indicators, such as team atmosphere and tactical innovation, could also enhance the model's comprehensiveness. With these improvements, our model has the potential to provide an even more scientific basis for countries to develop their Olympic strategies, further contributing to the field of sports science and data analysis.

References

[1] Wang Yuyang, Huang Chengyin. Analysis of the Performance Prospect and Preparation Strategies of the Chinese Table Tennis Team for the Paris Olympic Games[J]. Journal of Southwest China Normal University (Natural Science Edition), 2022, 47(4): 125-132.

[2] Wang Hanhan. Linear Fitting Analysis and Prediction Research on the Men's 110-meter Hurdles Race Results in the Past 20 Years[J]. Bulletin of Sports Science & Technology, 2023, 31(4): 46-49.

[3] Schlembach C, Schmidt S L, Schreyer D, et al. Forecasting the Olympic medal distribution–a socioeconomic machine learning model[J]. Technological Forecasting and Social Change, 2022, 175: 121314.

[4] Parveen Badoni, Priya Choudhary, Challa Parvathi, et al. Predicting Medal Counts in Olympics using Machine Learning Algorithms: A Comparative Analysis[J]. IEEE International Conference on Advanced Computing & Communication Technologies, 2023.

[5] Huimin S H I, Dongying Z, Yonghui Z. Can Olympic Medals Be Predicted? Based on the Interpretable Machine Learning Perspective[J]. Journal of Shanghai University of Sport, 2024, 48(4): 26-36.

[6] Jiayong L, Zhuohong W E I. Research on the Prediction of Men's 100m Gold Medal Results of the Olympic Games Based on GM (1, 1) Grey Model[J]. Journal of Southwest China Normal University (Natural Science Edition), 2023, 48(7): 123-128.

[7] Peng Jinqiang, Jing Longjun, Chen Shuyin, et al. Analysis of the Development Trend and Grey Prediction of the Track and Field Event Results in the Paris Olympic Games Based on the Results of the World Athletics Championships[J]. Bulletin of Sports Science & Technology, 2024, 32(4): 20-26. DOI: 10.19379/j.cnki.issn.1005-0256.2024.04.006.

[8] Dorigo M, Stützle T. Ant colony optimization: overview and recent advances[M]. Springer International Publishing, 2019.

[9] Yang Lingchun, Wang Xiangyu, Shang Zhiqiang. Research on the Action Recognition of Surfers Based on Support Vector Machine and Hidden Markov Model[J]. Sports Research and Education, 2024, 39(5): 68-73.

[10] Yang Qinwei. Prediction of the 2020 Olympic Games Results by Multiple Linear Regression Model [J]. Electronic Production, 2018(4): 121 - 123.