Research on gold medal prediction model based on multi-algorithm fusion

Zhiyuan Xiao, Ziyang Ma, Junjie Wang, Wei Liu *

Shaoxing University Yuanpei College, Shaoxing, China

* Corresponding Author Email: weiliu@zju.edu.cn

Abstract. In this study, a prediction framework based on multi-source historical data is proposed, and feature engineering and regularization techniques are used to construct interpretable machine learning models. After screening key features through Pearson correlation analysis, ridge regression and SHAP prediction models are built respectively, and the main effect correction parameter is innovatively introduced to enhance the model generalization ability. In the process of feature space construction, dynamic weight indicators are generated based on time-series data, and high-dimensional feature dimensionality reduction is accomplished through logistic regression, ultimately forming a predictor system containing four core dimensions. At the algorithmic level, this study compares the performance of L2 regularized ridge regression and game theory-based SHAP prediction model in prediction. And the experiments are parameter tuned, and the results show that the regularized model is effective in suppressing overfitting, and its mean square error (0.281) is reduced by 71.4% compared with the SHAP benchmark model. By constructing a dual assessment system, both the prediction accuracy (R^2>0.95) and the quantitative resolution of feature contribution are ensured. In terms of feature importance analysis, the SHAP value calculation reveals the nonlinear relationship of each dimension, and the probability prediction module realizes the interval estimation of the probability of event occurrence through logistic regression integration. The framework exhibits strong robustness on the validation set. The hybrid modeling approach proposed in this study provides a new technical path for the time-series prediction problem, and its modular design can be extended to other prediction scenarios that require a balance between accuracy and interpretability.

Keywords: SHAP method, Ridge regression model, home field advantage effect, Logistic regression.

1. Introduction

Traditional multi-source time series forecasting studies predominantly rely on static data frameworks for regression modeling, exhibiting limitations in capturing real-time dynamic information. Early approaches, such as linear models based on macroeconomic indicators (e.g., GDP, population size) [1], demonstrate constrained explanatory power due to insufficient variance interpretation and incapability to resolve high-dimensional covariate challenges. While recent machine learning techniques (e.g., Random Forest, XGBoost) improve predictive performance through non-linear feature interactions [2], they remain restricted by static data latency and fail to effectively integrate real-time dynamic factors like spatial weighting effects or participation scale fluctuations.

In this paper, a multi-algorithm fusion prediction framework is proposed to address the challenges of high dimensional data covariance problem and lack of model transparency through regularisation techniques and interpretability analysis. Firstly, Ridge Regression is used to constrain the parameter space and suppress the influence of multicollinearity on the prediction stability; secondly, SHAP (SHapley Additive exPlanations) values are introduced to quantify the feature contributions and enhance the model interpretability; lastly, a probabilistic prediction layer is constructed by combining with logistic regression, which dynamically evaluates the discrete event occurrence probability. In addition, a spatio-temporally weighted dynamic correction module is designed to calibrate the performance gain due to local effects.

The study is based on a multidimensional time series dataset of historical tournaments from 1992-2024, covering features such as participation size, tournament distribution, historical results and home field advantage. By comparing the mean square error (MSE: 0.281 vs. 0.984) of the ridge regression and SHAP models, the obvious advantage of ridge regression in terms of prediction accuracy is verified. The results show that the proposed framework not only outputs highly reliable prediction results, but also provides a data-driven basis for resource allocation strategies through feature contribution decomposition. This study provides a fusion algorithm paradigm for multi-source time series prediction problems that balances accuracy and interpretability, and can be extended to scenarios such as resource scheduling and risk assessment.

2. Research Methods

2.1. Data acquisition and pre-processing

All data used in this study were obtained from the open-source website (https://www.comap.com/contests/mcm-icm). The site provided the data base for this paper. This study enumerates four main real-time dynamic features and explains the reasons for their impact

The prediction model in this paper only uses data from 1992 to 2024 because the data of the event in this period is completer and more consistent, and can better reflect the modernization and development of the event in terms of the competition format, program setting and selection mechanism. Considering that in the absence of major international political or social changes, the number of participating countries usually remains relatively stable, and that changes tend to occur gradually, it is assumed that the number of participating countries in 2028 will remain the same as in 2024.

2.2. Methodology

1) Prediction of competition results without considering host effects

In order to predict tournament results without considering the host effect, this paper considers the use of SHAP prediction model as well as ridge regression prediction model to predict tournament results.

SHAP (SHapley Additive exPlanations) is an explanatory method based on game theory to understand the output of complex machine learning models by calculating the marginal contribution of features. The method uses Shapley values to measure the contribution of each feature to the model's predictions, which is unique and has good theoretical properties. The SHAP method accurately assesses the importance of features (the higher the Shapley value, the higher the contribution of the feature to the prediction) and applies it to predicting the outcome of a race in the year 2028, analyzing the world's most popular races, and providing data to support and inform decision-making for future races in the tournament. Its core strengths are fairness, transparency, and the ability to extract clear feature influences from complex models, providing valuable insights to different stakeholders [3].

In this study, using the characteristic variables such as the results of each country from 1992 to 2024 (total 8 competitions), the number of participating events, the number of participating athletes, the total number of points, and so on, the data set was trained by using XGBoost, and the effect of its Shapley value on the weight was calculated to get the prediction results, and the specific steps are shown as follows:

The first step is to establish the total prediction contribution equation. In this paper, the additive feature attribution method is used to interpret the contribution of each feature variable to the prediction result of the model as the contribution to the final result when the variable participates in the prediction. The “total predictive contribution” of the model can be expressed as the sum of the contributions of each feature, as shown in Equation 1.

$$g(x) = \phi_0 + \sum_{i=1}^{M} \phi_i(x) \cdot 1(x_i), \tag{1}$$

Where $x=(x_1,x_2,...,x_M)$ is an M -dimensional feature variable; $1(x_i)$ is a binary indicator variable that takes the value of 1 to indicate that the ith feature variable is used for prediction, and 0 to indicate that it is not used for prediction; $g(x)$ denotes the outcome of the model prediction for a given country/region.; ϕ_0 denotes the prediction mean, or baseline contribution; and $\phi_i(x)$ denotes the marginal contribution of the feature variable x_i to the prediction outcome.

In the second step, the marginal contribution is calculated. The calculation method of marginal contribution $\phi_i(x)$ is borrowed from the concept of Shapley value in cooperative game theory, which is used to calculate the degree of contribution of characteristic variables to the model prediction results. Its calculation formula is shown in Equation 2.

$$\phi_i(x)=\sum_{S\subseteq F\setminus\{i\}}\frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right)-f_S\left(x_S\right)\right], \tag{2}$$

Where F is the set of all feature variables used by the model, S is a subset of $F\setminus\{i\}$, and x_S is the feature variables encompassed in the subse S ; $x_{S\cup\{i\}}$ is the set of feature variables in the subset S plus the feature variables x_i ; $f_S(x_S)$ is the prediction result of the model trained based on the feature subset S ; $f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right)$ is the prediction result of the model trained based on the feature subset $S\cup\{i\}$; $|F|$ is the total number of feature variables in the feature set; $|S|$ is the number of feature variables in the subset.

In the third step, the XGBoost model was used to train and integrate the SHAP analysis, and finally output the Shapley values of feature variables and performance prediction results. The feature data (including results, number of events, number of athletes, total score, etc.) of eight events from 1992 to 2024 were input into the XGBoost model and trained, and the prediction accuracy of the model was improved by optimizing the hyper-parameters; secondly, on the basis of the trained model, the SHAP framework was used to compute the Shapley values of the feature variables and quantify their contributions to the performance prediction such as the host country effect, athlete size, and other key influencing factors; finally, the model calculated the Shapley values of feature variables and the performance prediction results using SHAP. Secondly, based on the trained model, the SHAP framework is used to calculate the Shapley values of each feature variable and quantify its contribution to performance prediction, such as the host country effect, the size of athletes and other key influencing factors. This step not only provides predicted values, but also reveals the core variables driving the predictions through an interpretable machine learning approach, providing a theoretical basis for strategic decision-making for future events.

The second prediction method uses a ridge regression model, whose central role is to address the problem of multicollinearity in linear regression. Ridge regression constrains the regression coefficients by introducing an L2 regularization term (a penalty term for the sum of squares of the model parameters), which reduces the model variance and improves generalization. In this paper, due to the nearly perfect positive correlation between the feature variables, if multiple linear regression is used directly, the multicollinearity will lead to too much variance in the parameter estimates or even unresolvable. Specifically, multiple linear regression requires that there is no significant correlation between independent variables, and highly correlated variables will undermine this assumption and make the model less stable. Therefore, in this paper, the ridge regression model is used to regularize the highly correlated features, which effectively mitigates the covariance problem while retaining the key information, thus obtaining more reliable achievement prediction results [3].

The ridge regression prediction model process includes data preprocessing, setting an initial regularization factor and optimizing that factor, and then training the final model. Once the model is trained, its performance is evaluated and the results are visualized. The final output includes the importance of features, prediction results and performance evaluation of the model.

In this paper, there are four achievement feature variables, and in the presence of multicollinearity, it is necessary to switch to the ridge regression optimization algorithm for modeling, and the ridge regression equation is shown in Equation 3.

$$\hat{\beta}(k)=\left(X^{T}X+kl\right)^{-1}XY, \tag{3}$$

Where X is the matrix of independent variables (eigenvariables) of size $n\times p$ (n is the number of samples and p is the number of features); Y is the vector of dependent variables (target variables) of size $n\times 1$; β is the vector of coefficients of size X; λ is the regularization coefficients (hyperparameters), which are used to control the strength of the regularization terms; I is the unitary matrix of size $p\times p$; and X^{T} is the transposition matrix of X.

2) Consider the host advantage effect [4]

By hosting the event, countries are able to fully demonstrate their economic and cultural strengths and utilize the home field advantage to achieve better results. Taking 1992 to 2024 as the observation interval, Fig. 1 shows the comparison of the performance of each event host country in terms of achievement with the previous performance for that event [5].

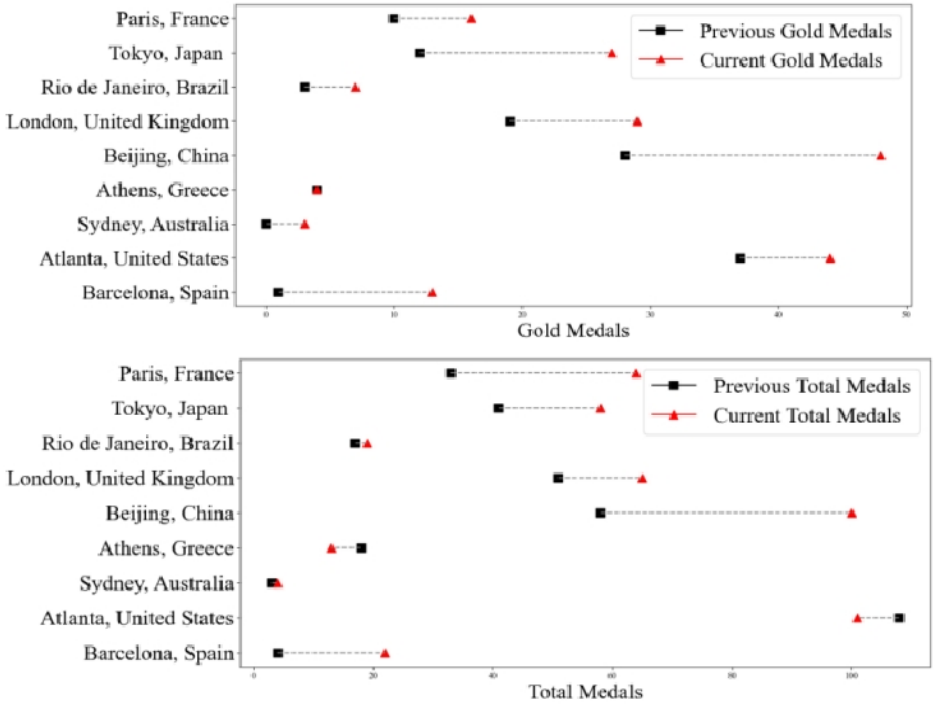

Figure 1. Changes in race results

As shown in Figure 1, with the exception of 1996 and 2004, most countries have improved their performance to some extent after becoming hosts of the competition. Based on this phenomenon, this study hypothesizes that competition host status has an incremental effect on performance. In order to quantify this effect, this study used multiple linear regression to calculate and analyse it [6, 7].

The first step is the selection of the characteristic variables. The selection of variables characterizing the home field advantage effect of the contest is shown in Table 1:

Table 1. Variables Characterizing Home Field Advantage

Notation	Characteristic Variables
K_1	Adjusted sports in the programme
K_2	Additional number of athletes
K_3	Number of previous winners
K_4	Total results of the previous competition
$H_{\text{Total Medals}}$	Number of Overall results at the hosting session
$H_{\text{Gold Medals}}$	Number of champions at the hosting session

As shown in Table 1, the number of sports added by the host country, the number of athletes added by the host country, the number of previous champions and the total performance of the previous edition were selected as characteristic variables in this study. This choice was made because it was considered that the host country is usually able to gain an advantage by increasing the number of participating athletes and adjusting the sports when organizing the competition. In addition, these characteristic variables have a strong correlation with the number of champions and the host country's overall performance, while the correlation between the characteristic variables is weak, making them suitable for use in the analysis of the multiple regression model.

The second step was to build a multiple regression model. By constructing the multiple regression model, the following functional relationships were established: the relationship between the adjusted number of events in the host country, the number of new athletes in the host country, the number of championships in the previous year and the total number of championships in the host country (as shown in Equation 4) and the relationship between the adjusted number of events in the host country, the number of new athletes in the host country, the results of the competition in the previous year and the results of the competition in the host country (as shown in Equation 5).

$$H_{\text{Gold Medals}} = a_1 K_3 + b_1 K_1 + c_1 K_2 + C_1 \tag{4}$$

$$H_{\text{Total Medals}} = a_2 K_4 + b_2 K_1 + c K_2 + C_2 \tag{5}$$

The multivariate regression model was used to quantify the host country effect, and then combined with the SHAP prediction model and the ridge regression prediction model to obtain the final competition score prediction model.

This fusion model ensures the computational efficiency and stability, while taking into account the accuracy of prediction and the interpretability of decision-making, and provides multidimensional analytical support for the prediction of competition scores.

3) Forecast of the number of countries achieving the top three for the first time

The top of the race results table is always in the spotlight, but the race results achieved by other countries are just as important. For example, Albania, Cape Verde, Dominica and St. Lucia finished in the top three for the first time in 2024.

Logistic regression is a widely used statistical model for classification problems, and although it includes “regression” in its name, it is primarily used to solve binary classification problems and can be extended to multivariate classification problems. In this article logistic regression is used to predict how many countries will participate in the next tournament for the first time. Although logistic regression is usually used in classification problems, here it is used to predict a discrete value (the number of countries that will participate for the first time). The computational procedure is as follows:

The first step is to construct a polynomial logistic regression. Logistic regression predicts probabilities by linearly combining features. For multi-categorization problems like this one, multinomial logistic regression can be used. Its mathematical form is shown in Equation (6).

$$P\left(y = k \middle| X\right) = \frac{e^{w_k X}}{\sum_{j=1}^{K} e^{w_j^T X}}, \tag{6}$$

Where $X = [X_{1992}, X_{1996}, \cdots, X_{2024}]$ denotes the number of countries finishing in the top three for the first time in the eight editions of the competition between 1992 and 2024; X is the expanded feature vector; w^k is the weight vector for category k; and K is the total number of categories (number of countries likely to finish in the top three of the competition for the first time).

In the second step, Logistic regression learns the model parameters by maximizing the log-likelihood function. For the multiclassification problem, the loss function (negative log-likelihood) is shown in Equation 7.

$$L(w) = -\sum_{i=1}^{N}\sum_{k=1}^{K}\prod\left(y_i = k\right)\log P\left(y_i = k \middle| X_i\right), \tag{7}$$

Where N is the number of samples, and 0 otherwise; $P(y_i = k|X_i)$ is the probability that sample $i-th$ belongs to category k; $\prod(y_i = k)$ is an indicator function that takes 1 when the true category of sample $i-th$ is k.

In the third step, the predicted probability is calculated. After the training is completed, the model can predict the class probability distribution of the new samples. For the prediction of 2028, assuming that the eigenvalue of 2028 is $X_{2028} = [X_{1992}, X_{1996}, \cdots, X_{2024}]$, The predicted probability distribution for the year 2028 is shown in Equation 8:

$$P(y_{2028} = k | X_{2028}) = \frac{e^{w_k^T X_{2028}}}{\sum_j^K e^{w_j^T X_{2028}}}, \tag{8}$$

The final predicted number of countries with first-time competition results is the category with the highest probability, as shown in Equation 9.

$$\hat{y}_{2028} = \arg\max_k P(y_{2028} = k | X_{2028}). \tag{9}$$

4) Relationship between disciplines and competition score

This study aims to explore the correlation mechanism between the number of competitive sports and tournament results, focusing on analyzing the following three dimensions: firstly, identifying the dominant sports that have the most influence on countries' medal acquisitions; secondly, revealing the intrinsic link between national strategic choices and the cultivation of dominant sports; and finally, quantitatively assessing the influence weights of different athletic elements on the tournament results based on the interpretable SHAP (SHapley Additive exPlanations) value model to quantitatively assess the influence weights of different athletic elements on the outcome of the competition. By constructing a multi-dimensional analysis framework, this study attempts to systematically analyse the role of national sports resource allocation strategies and competition results, and provide a data-driven decision-making basis for the development of competitive sports.

Competition events are closely related to the final competition results, and the contribution of different events to the competition results varies greatly from country to country. In order to find out the most important competition items of each country in the competition, this study combines the machine learning model with SHAP value analysis based on historical data to explore the impact of each item on the competition results, and counts the number of countries participating in each item in the previous competitions from 1992 to 2024, so as to find out the most popular competition items in the world. The specific process is shown in Figure 2:

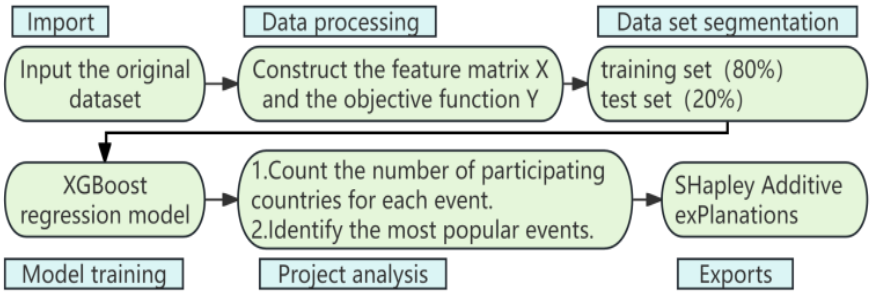

Figure 2. Flowchart for analysis of significant items

As shown in Figure 2, the process of analyzing the tournament based on XGBoost model is demonstrated. First, the raw data are imported, cleaned and normalized, and a feature matrix is constructed, which contains key indicators such as the number of participating countries and the distribution of historical results. The dataset is divided into 80% training set and 20% test set, and the XGBoost regression model is used to optimize the objective function to predict the performance. After the model training was completed, the feature contributions were analyzed by additive

interpretation methods such as SHAP, which identified high heat events with many participating countries as a factor that had a significant impact on the competition results. In this process, the number of countries participating in each event was also counted, and the model results were combined to assess the differences in the international influence of each event program.

There are significant differences in the contribution of different competition programs to the competition performance of each country. Program selection not only affects the competition performance of each country, but also reflects the competitiveness and advantages of each country in specific programs. Similarly, the machine learning model and SHAP value analysis can be used to quantify the impact of each competition item on the competition results, and then analyze the impact of item selection on the results. The specific process is shown in Figure 3:

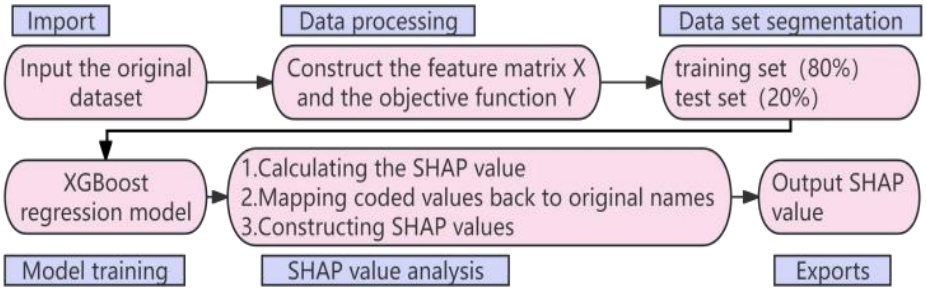

Figure 3. Discipline Selection Impact Results Flowchart

Similar to the flowchart of important project analysis, this study first imports the award data of countries participating in various competitions in past years, cleans and normalizes the raw data, and then divides the dataset according to the ratio of 80% training set and 20% testing set. The key variables are clarified by constructing the feature matrix, and the objective function is defined to guide the optimization direction of the model. After training the model using the XGBoost regression algorithm, SHAP values are calculated and constructed to ensure the interpretability of the results by mapping the coded values back to the original variable names. Finally, the SHAP values are analyzed in depth to quantify the impact of each item on competition performance.

2.3. Indicators for model evaluation

1) Correlation analysis

For the relationship between the characteristic variables and the competition results, the Pearson correlation coefficient can be used. The Pearson correlation coefficient quantifies the linear relationship between continuous variables and assesses the strength and direction of the correlation. By calculating the correlation coefficient, it is possible to specify the positive, negative or uncorrelated relationship between each characteristic and the number of medals [8].

The mathematical expression of Pearson's correlation coefficient formula is shown in Equation 10:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{10}$$

Where X and Y represent two feature variables, respectively.

2) Mean Square Error (MSE)

Both the SHAP prediction model and the ridge regression prediction model show good prediction results. In order to compare the performance of these two models, this paper utilizes these two models to predict the 2024 competition results separately based on the consideration of the home field effect, and evaluates the prediction accuracy of both models by calculating the mean square error (as shown in Eq. 11), so as to determine which model is more advantageous [9].

$$MSE = \frac{1}{n}\sum_{i}^{n}\left(I_i - \hat{I}_i\right)^2, \tag{11}$$

Where denotes the projected competition score for each country, I_i denotes the predicted value, $\hat{I}_i$ denotes the true value, and MSE denotes the mean square error (the smaller the value of MSE, the better the result).

3. Results and analysis process

3.1. Relevance results

The correlation of the characteristic variables in each predictive model was obtained through Pearson's correlation coefficient, as shown in Figure 4:

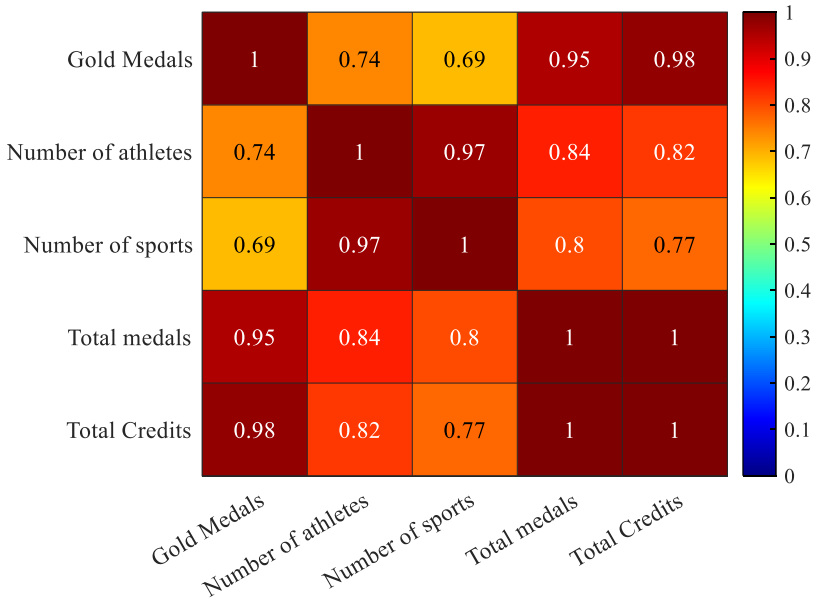

Figure 4. Pearson Correlation Coefficients of Each Feature Variable

As shown in Figure 4, there is generally a strong positive correlation between athlete scale, number of sports events, and total points with the number of gold medals and total medal count. The athlete scale shows a moderate positive correlation with gold medals (0.74) and total medals (0.84), meaning countries with more athletes tend to win more medals. The number of sports events has a weaker correlation with gold medals (0.69), but a stronger correlation with total medals (0.80), indicating that countries participating in more events typically win more medals. Total points have an almost perfect positive correlation with both gold medals and total medals, as the gold medals and total medals play a dominant role in determining the total points.

3.2. Comparison of forecasting models

3.2.1. Solve the SHAP prediction model

Through the SHAP prediction model, the data set of the characteristic variables was substituted into the model and the results of the Shapley values of each variable were calculated, as shown in Tables 2 and 3:

Table 2. Projected number of gold medals

Characteristic Variables	Shapley
Historical number of medals	0.351
Historical Olympic Points	3.039
Number of Athletes	0.049
Number of Events	0.119

Table 3. Projected Total Medals

Characteristic Variables	Shapley
Historical Gold Medal Count	0.381
Historical Olympic Points	7.982
Number of Athletes	0.072
Number of Events	0.163

Based on the Shapley value analysis results in Tables 3 and 4, historical Olympic points contribute the most to predicting the number of gold medals and total medals (with Shapley values of 3.039 and 7.982, respectively), making it the core driving factor. Historical medal count and historical gold medal count provide secondary contributions to predicting the gold medal count and total medal count (with Shapley values of 0.351 and 0.381, respectively), while the number of participating athletes and the number of events have smaller contributions (with Shapley values below 0.2). By calculating the feature vector weights and performing XGBoost prediction, an initial model for predicting gold/total medals can be obtained. Combining this with the home field advantage effect will provide the final prediction result.

3.2.2. Solve the ridge regression model

Substituting the feature variable datasets, the model is then subjected to data preprocessing, feature engineering, and model training, and finally the ridge regression equations for the number of gold medals and the total number of medals in the 2028 Olympic Games are obtained, as shown in Equations 12 and Equations 13:

$$Y_{\text{Gold Medals}} = 15.13X_1 + 0.07X_2 - 0.32X_3 - 9X_4 + 1.58 \tag{12}$$

$$Y_{\text{Total Medals}} = 20.42X_1 + 0.3X_2 + 0.13X_3 - 5.3X_5 + 4.92 \tag{13}$$

Where $Y_{\text{Gold Medals}}$ is the predicted number of gold medals, $Y_{\text{Total Medals}}$ is the predicted number of total medals, X_1 is the total number of Olympic points, X_2 is the number of events entered, X_3 is the number of athletes, X_4 is the number of total medals, and X_5 is the number of gold medals.

3.2.3. Home field advantage effect results

The number of gold medals and the total number of medals in the previous Olympics can be regarded as the predicted values without the effect of home field advantage, while the number of gold medals and the total number of medals in the host Olympics reflect the predicted values under the effect of home field advantage. Through the least squares solution, this study obtained the multiple regression equations for predicting the number of gold medals and the total number of medals under the effect of home field advantage, as shown in Equation 14 and Equation 15:

$$H_{\text{Gold Medals}} = 0.006Y_{\text{Gold Medals}} + 0.047K_1 - 0.137K_2 + 7.344 \tag{14}$$

$$H_{\text{Total Medals}} = 0.007Y_{\text{Total Medals}} - 0.096K_1 + 0.14K_2 - 3.43 \tag{15}$$

The test obtained the multiple regression equation R^2 of gold medals and medals are 0.975, 0.994 respectively, R^2 is very high, the regression effect is good, so that this study gets the number of gold medals and medals under the influence of the organizer effect.

3.2.4. Mean square error comparison results

By calculation, the F of SHAP prediction model for predicting the number of Olympic gold medals is 0.790, and the MSE of SHAP prediction model for predicting the number of Olympic medals is 0.984; the MSE of ridge regression prediction model for predicting the number of Olympic gold medals is 0.262, and the MSE of ridge regression prediction model for predicting the number of Olympic medals is 0.281. The results show that the MSE of ridge regression prediction model for predicting the number of Olympic gold medals and the number of medals is smaller than that of SHAP

prediction model, so this study choose the ridge regression model as the final prediction model. The results show that the *MSE* of the ridge regression prediction model in predicting the number of gold medals and the number of medals is smaller than that of the SHAP prediction model, so this study choose the ridge regression model as the final prediction model.

3.3. Predictions of the outcome of the competition for each participating country

Through the ridge regression prediction model, the number of gold medals and medals for the non-host is firstly predicted, and then the prediction data for the host (USA) is further calculated by combining the home field advantage effect. The specific flow of the algorithm is shown in Algorithm 1:

Algorithm: Ridge Regression Algorithm

Start

Input: X : Feature Matrix

Y : Target variable

α : Regularization coefficient

For i =1,2,3 …… n

Normalized feature matrix

Transpose a matrix

Calculate: $\mathrm{w}_i = \left(X_i^T X_i + \alpha I\right)^{-1} X_i^T Y_i$

Output: w : Regression coefficient

End

The process of Algorithm 1 has been implemented through the program code, which finally calculates and produces the predicted number of gold medals and medals for the world countries in the 2028 Olympic Games in Los Angeles, USA, as shown in Figure 5 and Figure 6:

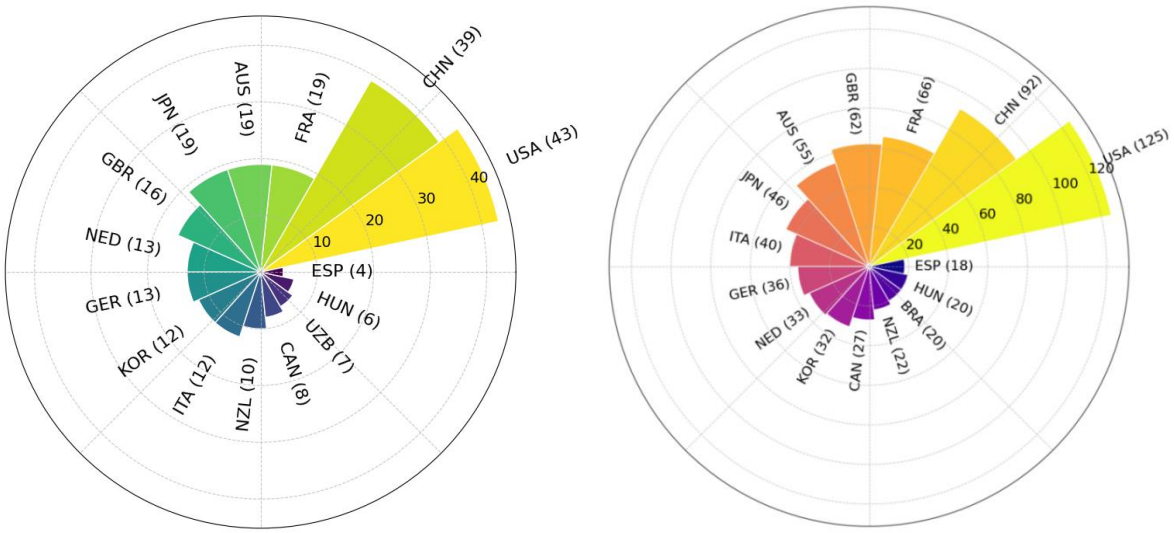

Figure 5. 2028 Summer Olympics Gold Medal Table (left) and Gold Medal Table (right)

The visual presentation of the data in this study shows (as shown in Figures 5 and 6) that the number of gold medals for each country is characterized by a significant unbalanced distribution. It is worth noting that, based on the analysis of the competition performance prediction model, although the 2028 Olympic Games are expected to welcome 216 participating national/regional delegations, the quantitative prediction results show that only 34.3% of the participating subjects (about 74 countries/regions) are actually competitive for medals.

3.4. Prediction of the number of countries achieving their first top-three finish in a competition.

r substituting the data on first-time medalists in the Summer Olympics from 1896 to 2024 into the logistic regression model, the probability distribution of the number of countries winning first-time medals in the 2028 Olympics was obtained by maximizing the log-likelihood function to learn the model parameters, as shown in Figure 6:

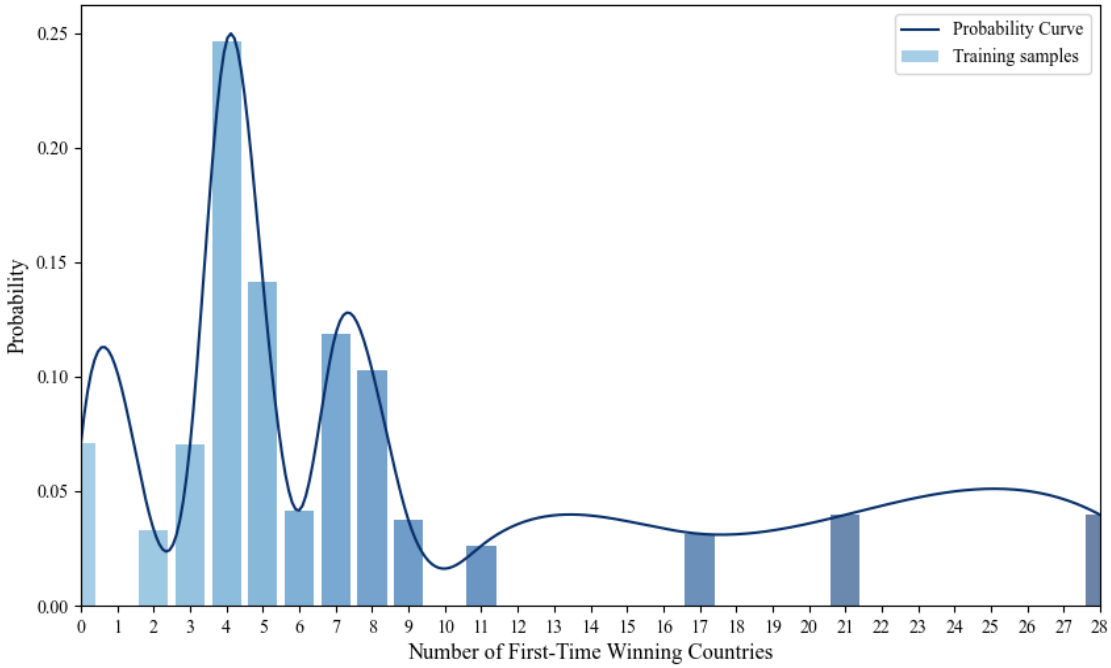

Figure 6. Probability of First-Time Medal-Winning Countries in 2028 Olympics

The probability distribution of the number of countries winning first-time medals at the 2028 Olympic Games shows that the most likely outcome is 4 countries winning first-time medals with a probability of 24.67%. 5 and 7 countries have a probability of 14.17% and 11.87%, respectively, indicating that these outcomes also have a high probability of occurring, and that low-probability events (e.g., 9, 10, and more countries) have a low probability of occurring.

3.5. Most Important Olympic Events

The most important competition reveals are obtained by covering the key steps of data processing, model training, SHAP value analysis and analysis of the number of countries participating in the project, and the results are shown in Table 4:

Table 4. Projected Total Medals

Sports Event	SHAP Value
Athletics	9.685803937
Wrestling	9.635639061
Swimming	9.087853294
Sailing	8.496754212
Shooting	8.489501564
Boxing	7.110579924
Gymnastics	6.840733535
Tennis	6.62137932
Badminton	6.607880376
Basketball	6.173330001

As shown in Table 6, the higher SHAP value means that the sport is more important and popular, and the importance of Athletics, Wrestling, Swimming and other competitions in descending order.

3.6. Olympic programmes for national representatives

Based on the medal data of each country in each sport in the previous Summer Olympics from 1992 to 2024, we have preprocessed the data, trained the model and analyzed the SHAP values, and finally came up with the sports with the highest SHAP values in each country, and the results are shown in Table 5:

Table 5. Projected Total Medals

Country	Strengths Campaign	SHAP Value
USA	Wrestling	3.526
AUS	Swimming	2.623
KEN	Athletics	2.461
KOR	Archery	2.436
CHN	Badminton	2.233
ETH	Athletics	1.997
JAM	Athletics	1.924
GBR	Athletics	1.602
GER	Athletics	1.584
CUB	Wrestling	1.557

As shown in Table V, the advantageous program of the United States is wrestling, the advantageous program of Australia is swimming, and the advantageous program of Kenya is track and field. Through the analysis of SHAP value, the advantageous and disadvantageous items of each country can be clarified, so as to reasonably choose the participating items in the Olympic Games and prioritize the participation in the advantageous items of the country. In addition, further intensive training for the advantageous programs, while the disadvantageous programs are targeted to improve, in order to increase the number of medals won [10].

4. Results and analysis process

In this paper, we successfully constructed a multi-algorithm fusion Olympic medal analysis model. The number of gold medals and medals of 2028 Los Angeles Olympic Games are predicted by ridge regression algorithm with some credibility, the logistic regression model provides some references for the prediction of the countries that won medals for the first time, and the combination of SHAP worths the most popular competitions around the world. Through the results of the study, this paper finds that: the top three countries predicted in the gold medal list of the 2028 Summer Olympics are the United States 43, China 39 and France 20, and the top three countries predicted in the medal list are likewise the United States 130, China 95 and France 70; the probability distribution of the number of countries that will win medals for the first time in the 2028 Olympics shows that the most probable outcome is that four countries will win medals for the first time, with a probability of 24.67 per cent, while the probabilities of five and seven countries are 14.17 per cent and 11.87 per cent, respectively, indicating that these outcomes also have a high probability, but low-probability events such as nine, 10 and more countries winning medals for the first time have a low likelihood of occurring; furthermore, the importance of the competition events globally is, in descending order, for the sports of Athletics, Wrestling, Swimming, and so on. Finally, the research in this paper provides new methods and ideas for the prediction and analysis of Olympic medals, and future research can further go on to optimise the model, introduce more dynamic factors, and improve the accuracy and practicality of the model. The same research method can also be applied to other neighbourhoods for talent development and competition research.

References

[1] Yuan, J. A preliminary study of Olympic gold medal prediction model in the era of big data: Taking the results of the Athletics World Championships as an example [J]. Sports Science and Technology Literature Bulletin, 2021, 29 (06): 132 - 134.

[2] 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.

[3] Liao, L., Zhao, Z., Li, Z., et al. Logistic regression and SHAP analysis for modeling and validation of femoral head necrosis after internal fixation of femoral neck fracture [J/OL]. Chinese Tissue Engineering Research, 2025 - 03 - 03.

[4] Xue, Y., & Yang, W. Influencing factors and suggestions for home field advantage in the Winter Olympics [J]. Chinese Sports Coach, 2022, 30 (01): 28 - 30+63. DOI: 10.16784/j.cnki.csc.2022.01.004.

[5] Yang, Y., & Zhu, F. Research on logistics demand forecasting in Guangxi based on ridge regression model [J]. Logistics Technology, 1 – 9.

[6] Tian, H., He, Y., Wang, M., et al. Medal prediction and participation strategy of Chinese athletes in the 2022 Beijing Winter Olympics: Analysis based on the effect of home field advantage in the Olympic Games [J]. Sports Science, 2021, 41 (02): 313 + 22.

[7] Feng, D. An analytical study of home field advantage in the 2002-2018 Winter Olympics [C] // Chinese Society of Sports Science. Compilation of Abstracts of Papers from the 11th National Sports Science Conference. Beijing Sport University; 2019: 3.

[8] Zhang, W.-F. Statistical characterization of Spearman's parsimonious correlation coefficient and Gini gamma correlation coefficient [D]. Guangdong University of Technology, 2020.

[9] Wu, J., Jing, B., Jing, J., et al. Study on MRI denoising based on nonlocal mean and linear least mean square error estimation [J]. China Medical Devices, 2025, 40 (02): 35 - 39+66.

[10] Yan, S., & Zhao, Y. Intercontinental distribution and trend prediction of track and field medals in the 27th to 31st Olympic Games [J]. Journal of Xichang College (Natural Science Edition), 2019, 33 (03): 68 - 74.