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

Medal Prediction Models Based on LASSO Regression and Random Forest Algorithm

Xinran Chen, Xuming Yan *, Rongtao Zhang

Soochow University, Suzhou, China

* Corresponding Author Email: 16665256115@163.com

Abstract. Medal prediction serves as a critical research direction in sports science and data analysis, holding significant implications for optimizing resource allocation and strategic decision-making in competitive sports. This study proposes an innovative hybrid predictive model that integrates hierarchical clustering, LASSO regression, and random forest algorithms. By constructing a purely competition-endogenous multidimensional competitiveness indicator system, the model overcomes the limitations of conventional approaches that rely heavily on external factors. The methodology begins with establishing feature-based indicators to categorize participating nations into three distinct clusters through hierarchical clustering, reflecting their respective stages of sports development and establishing an optimized differentiated modeling framework. For countries at different developmental stages, LASSO regression and random forest algorithms are strategically applied, achieving both model robustness and systematic exploration of feature importance. Empirical results demonstrate the model's capability to accurately forecast medal distributions for the 2028 Los Angeles event, with predictions aligning closely with historical trends and prediction errors confined within a margin of 2 medals. This research provides a quantifiable decision-making tool that substantially enhances the scientific basis for event resource allocation and policy formulation in competitive sports systems.

Keywords: hierarchical clustering; lasso regression; random forest.

1. Introduction

Medal distribution prediction has always been an important research direction in the field of sports science and data analysis. With the globalisation of competitive sports, the competition among countries in terms of investment of sports resources, optimisation of training system and technological empowerment has become increasingly fierce, and the establishment of an accurate medal prediction model can not only provide a basis for strategic decision-making for sports policy makers, but also provide theoretical support for event organisers to optimise the allocation of resources.

In recent years, the methodological paradigm in medal prediction has exhibited a notable transition from conventional econometric analysis toward intelligent modeling.While researchers such as Schlembach [1], Badoni [2], Tchamkerten [3], and Yeh [4] have successfully introduced machine learning algorithms (e.g., random forests, convolutional neural networks) into this domain, their implementations generally neglect the critical issue of multicollinearity among variables during feature engineering. Although studies by Scelles N [5], Otamendi F J [6], and Li F [7] have analyzed and quantified the marginal effects of macroeconomic variables such as population size and GDP, they exhibit a critical oversight by neglecting the significance of event-specific characteristics in competition outcomes. Concurrently, while scholars including Wunderlich F [8], Baumer B S [9], and Wilkens S [10] have concentrated on predictive modeling for individual sports disciplines, their research frameworks fail to adequately address the heterogeneity inherent in national sports development models.

Through the systematic critical analysis of the core literature in the past five years, this paper is based on the theoretical gaps and innovates as follows: First, it introduces systematic hierarchical clustering to reveal the heterogeneous hierarchical characteristics of the national sports development model and establishes various prediction frameworks to provide quantitative decision support for the

differentiation strategy of competitive sports resource allocation; Secondly, constructing a purely endogenous multidimensional competitiveness index system of the event, discarding the exogenous dependence of economic, population and other macro variables, and deeply mining the micro data of the event itself for prediction; finally, designing a hybrid architecture of LASSO regression and Random Forest algorithm, which effectively avoids the autocorrelation problem while specifically analysing the importance weights of each feature.

2. Model

2.1. Hierarchical Clustering

Hierarchical clustering is a flexible and intuitive clustering method for exploratory data analysis. Its core idea is to cluster samples based on sample similarity or distance metrics, and to recursively group the samples in a dataset hierarchically, creating a tree structure. It does not require us to specify the number of clusters in advance, but can gradually merge or split the clusters from fine-grained subdivisions to large clusters until a certain stopping condition is met.

For all countries that have ever won a medal, the bottom-up strategy is applied to unfold the systematic hierarchical clustering, and the specific process is shown in Fig. 1.

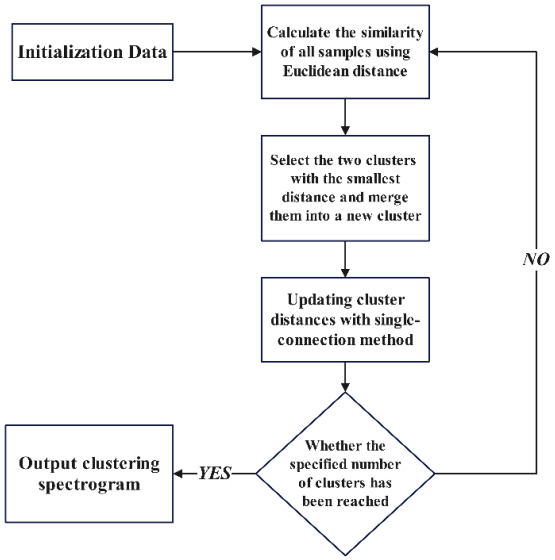

Figure 1. Operational procedures for Hierarchical Clustering

The elbow rule is used to estimate the optimal number of clusters, starting from 1 cluster and gradually increasing the number of clusters, for each number of clusters k, the clustering quality index SSE is calculated under the number of clusters, i.e. For each number of clusters k, calculate the clustering quality index SSE under the number of clusters, i.e. the sum of the squares of the distances from each data point to the cluster centre of each data point, and use the number of clusters as the horizontal coordinates and the SSE value as the vertical coordinates in the graph, in which the declining trend of SSE is observed, when the rate of decline slows down significantly, the corresponding point is called the "elbow". When the rate of decline slows down significantly, this point is called the 'elbow' and the corresponding number of clusters is usually considered to be the optimal number of clusters. When the rate of decline slows down significantly, this point is called the 'elbow', and the corresponding number of clusters is usually considered to be the optimal number of clusters.

In order to categorise all medal-winning countries according to their level of sporting development, the level of sporting development of a country is measured by the following four indicators: the total number of medals won historically, the total number of medals won at the most recent event and the total number of athletes competing at the most recent event, and the number of stable events in which each country competes.

2.1.1. Total number of medals awarded in history

The total number of medals won historically is the sum of the total number of medals won in each event in which a country has competed in its history, which reflects both the number of events in which each country has competed in the macro-time dimension and the number of medals won by each country in each event at the micro-level, which in turn reflects the overall level of a country's sporting performance. Therefore, the total number of medals won by each country in history is included as one of the assessment indicators for the classification.

2.1.2. Total number of medals won in the most recent event

The results of the most recent event are an important indicator of a country's recent sporting level, reflecting the overall level of active athletes in that country, and are another important indicator to distinguish from the total number of medals won historically.

2.1.3. Total number of athletes participating in the most recent event

In order to classify all the countries that have won medals, and to build separate models based on the different classifications to predict the number of medals each country might win in the future, the recent level of sport in each country is particularly important. Therefore, in addition to the total number of medals won by each country in the most recent event, the total number of athletes fielded by each country in the most recent event was selected as one of the evaluation indicators for the classification, and the number of active athletes supplemented the previous evaluation indicator to reflect the recent level of sport in the country.

2.1.4. Number of projects with stable entries

Sports in which countries have participated at least twice in the last three editions are defined as projects with stable entries. This indicator reflects the stability and continuity of countries in the design of their participation programmes. Stable participation in multiple sports provides an effective reference for assessing a country's potential to win awards and is one of the most important factors in measuring a country's level of sporting excellence.

2.2. LASSO Regression

LASSO regression is a technique for regularising least squares regression. The aim of the regularisation is to reduce overfitting by limiting the complexity of the model. This is done by adding a penalty term to the objective function, which takes the form of the sum of the absolute values of all the regression coefficients. The objective of LASSO regression is to minimise the following objective function:

$$Loss\ Function = \sum_{i=1}^{n}(y_i - y_i)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right|, \tag{1}$$

where y_i is the data obtained after normalising the original data using the normalisation equation:

$$y_i = \frac{x_i - \mu}{\sigma}, \tag{2}$$

where μ is the mean of the original data and σ is the standard deviation of the original data.
In equation (2), the first part on the right side of the equals sign is the residual sum of squares of OLS, i.e, the sum of squares of the error term, which is used as a measure of the model's fit; and the second part is the LASSO regularisation term, where β_j is the penalty regression coefficient and λ is the penalty parameter for the strength of the regularisation.

Based on data on the number of prizes won by each country in previous editions of the competition, the following five indicators have been selected to predict the total number of medals won by each country in the next edition of the competition.

2.2.1. Host effect

A comparative analysis of the results of the host country of previous events and its participation in the previous event shows that the number of gold medals won by the country as host of the current event, the total number of medals won, its ranking in the medal table and the number of athletes participating in the event have, on average, increased significantly. This shows that the host effect is crucial for predicting medals. The relevant data for six representative countries are shown in TABLE I.

Table 1. Host Effect

Symbol	FRA	US	ITA	JP	GR	CHN	Ave
Increase in gold medals	4	22	5	12	2	16	15.67
Increase in total medals	-3	54	11	11	3	37	32.50
Improvement in ranking	5	0	2	5	2	1	3.08
I Increase in athletes	179	308	196	173	320	212	252.83

During the competition, the host country has certain advantages in terms of home field advantage, government support, athletes' momentum, etc. because it is the host country, and under the support of good competition environment and resources, athletes from this country are more likely to overperform and achieve better results than usual. Therefore, a dummy variable $Host_{t,i}$ is introduced to indicate whether the i-th country of the t-th competition is the host country, which takes the value of 1 for the host country and 0 for the other participating countries, in order to capture the effect of the host country on winning medals.

In addition, the host country in four years' time and the host country four years ago have an effect on medals won, for example Team GB, as the host of London 2012, had a highly successful event in Beijing in 2008, winning 47 medals, a possible reason being that as the host of the next event, the country's athletes will begin their training and preparation earlier, with the aim of achieving better performances when they benefit from the host effect. Similarly, for the host country four years ago, it may be intuitively felt that the intensive preparation and training it did four years ago to win more medals as the host of the previous event may have had a lasting positive effect on the event four years later, even if the country is no longer the host [1]. Continuing with the dummy variables from above, $Host_{t-1,i}$ is used to denote the host of the previous event, $Host_{t+1,i}$ denotes the host of the subsequent event, and together they are vectorised to represent the full host effect of the current event:

$$\overrightarrow{Host_{t,i}} = (Host_{t-1,i}, Host_{t,i}, Host_{t+1,i}). \tag{3}$$

2.2.2. Total number of projects offered in this event

The total number of projects directly determines the total number of medals awarded, and the more sports a country has, the greater its chances of winning.

2.2.3. Number of participating athletes from each country

The total number of participating athletes is an important predictor of the number of medals, as countries sending more athletes are likely to win more medals.

2.2.4. Overall level of athletes

In order to reasonably quantify the overall level of athletes in each country, AHP-EWM was used to develop the evaluation analysis. First, the hierarchical analysis method is used, and based on common

sense, it is subjectively determined that the level of winning a gold medal is higher than that of winning a silver medal, and the level of winning a silver medal is higher than that of winning a bronze medal. Based on the above comparisons, a judgement matrix was constructed using hierarchical analysis to derive the evaluation weights for gold, silver and bronze medals respectively:

$$\omega = (0.6480, 0.2299, 0.1222). \tag{4}$$

where the consistency ratio of the judgement matrix is CR = 0.004 and the consistency is acceptable. Secondly, the entropy weighting method is used to objectively analyse the weights of gold, silver and bronze medals. Collect the attached data and classify and summarise the number of gold, silver and bronze medals won by participating athletes from different countries, noting that the total number of athletes is n, forming a raw data array with n rows and 3 columns. The normalisation performed on the data is shown below:

$$x_{ij} = \frac{x_{ij} - \min\{x_j\}}{\max\{x_j\} - \min\{x_j\}}, \tag{5}$$

where x_{ij} indicates the original award-winning data of the i-th athlete on the j-th indicator, $\tilde{x}_{ij}$ indicates the normalized data, $\max\{x_j\}$ and $\min\{x_j\}$ indicate the maximum and minimum values of all data on the j-th indicator, respectively.

Then, the proportion p_{ij} of the award data of the i-th athlete in the j-th index is calculated, and the specific formula is as follows:

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}. \tag{6}$$

According to the concepts of self-information and entropy in information theory, the information entropy E_j of the j-th index is calculated. The specific formula is as follows:

$$E_j = -\frac{\sum_{j=1}^{n} p_{ij} \ln p_{ij}}{\ln n}. \tag{7}$$

where $p_{ij} = 0$, to avoid the case of $\ln 0$, it can be calculated as a very small positive number (e.g. 10^{-10}).

Then, the redundancy d_i of the j-th index is calculated, which can reflect the amount of information of this indicator in the comprehensive evaluation. The specific formula is as follows:

$$d_j = 1 - E_j. \tag{8}$$

Finally, the weight σ_j of the j-th index is derived as follows:

$$\sigma_j = \frac{d_j}{\sum_{i=1}^{3} d_j}. \tag{9}$$

The combination weight of each indicator is calculated by combining the two aforementioned methods. For each indicator, two types of weights are calculated. To enhance the model's accuracy, the average of the weights of the two types of indicators is selected as the combination weight λ_j of each indicator. This approach serves to minimise error, as illustrated below:

$$\lambda_j = \frac{\omega_j + \sigma_j}{2}. \tag{10}$$

2.2.5. Comprehensive national strength (total number of medals won in the most recent event)

A review of the literature shows that background factors such as a country's total population and GDP are also important in predicting the number of prizes it will win, on the grounds that preparing for the next event usually requires a large amount of human and economic resources, so demographic and economic data are essential to the prediction. However, as the prediction model was only intended to use data from previous years, the level of national power of each country was predicted by reference to the winning record of the most recent event, which can be interpreted as the results of the most recent event may to some extent reflect the recent level of demographic and economic development of each country. In addition, the most recent event has a special significance for predicting the next games, which is actually called 'momentum', if a country has achieved very good results in the most recent event, then it is very likely that they will continue the excellent training methods, invest more external support and expect to achieve even better results.

2.3. Random forest

Random forest is an integrated learning method, mainly used for classification and regression tasks, which improves the accuracy and stability of the model by constructing multiple decision trees and combining their predictions, the basic principles of which are described below and visualised in Fig. 2.

2.3.1. Ensemble Learning

Random forests are a typical example of ensemble learning, which achieves improved prediction performance by combining multiple weak learners, such as decision trees. Each decision tree can make predictions independently, and the random forest arrives at the final prediction by voting (classification problem) or averaging (regression problem) these predictions.

2.3.2. Bagging

Random forests use bootstrap aggregating by sampling the original data set with put-back, i.e. each data point can be repeated multiple times for selection, thus generating multiple training sets, each of which is used to train an independent decision tree. The advantage of this method is that by training multiple trees, the variance of a single tree can be reduced, thereby reducing the risk of overfitting.

2.3.3. Random feature selection

When constructing each decision tree, Random Forest not only randomly samples the training data, but also randomly selects some of the features for decision making each time a node is split.This random selection improves the diversity of the model, thus reducing the possibility of overfitting.

The performance and generalisability of the model can be evaluated through cross-validation after modelling is complete. This is a common method of model evaluation in machine learning where the dataset is divided into smaller subsets and the model is repeatedly trained to obtain a more comprehensive and robust estimate of model performance. Specifically, the data is first divided into k subsets (often referred to as "folds"), and then k model trainings and simulations are performed,

each using k-1 subsets as training data and the remaining 1 subset as test data. This results in k model performance scores, which are then averaged to produce the final performance score.

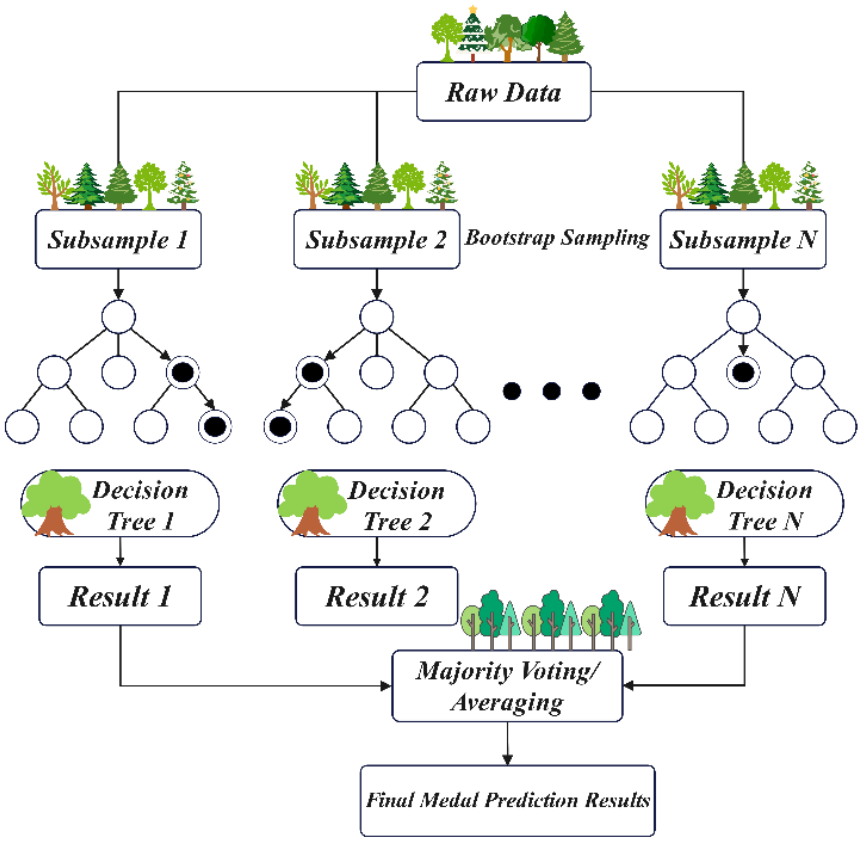

Figure 2. Random Forest Schematic

3. Results and Analysis

3.1. Hierarchical Clustering

The best clustering number 3 was obtained by the elbow method and the SSE decreasing trend graph is shown in Fig. 3. Hierarchical clustering divides all medal-winning countries into three categories and the results are shown in Fig. 4. It shows that cluster 1 contains 128 countries, cluster 2 contains 11 countries and cluster 3 contains 1 country. The result can be explained as follows: Cluster 1 contains 128 countries with a low and medium level of sport; Cluster 2 represents countries with a high level of sport, which have bright performances in the medal table, including the well-known Australia, Canada, China, France, Germany, Great Britain, Hungary, Italy, Japan, Netherlands and Sweden; Cluster 3 refers to countries with an excellent level of sports, and only the United States is included in this category, because it has a long history of sports and a deep sports culture, and the US team has not missed any of the normal events except for the event in Moscow in 1980, and such a long and stable history of participation has given it the opportunity to accumulate a large number of medals in all previous events. At the same time, the United States has a perfect sports selection mechanism and advanced scientific training methods, so it has world-class competitiveness in a number of sports and is able to send high-level athletes to participate in the competition and win gold and silver in these programmes.

Based on the clustering results, all medal-winning countries were classified into two categories: 128 countries in cluster 1 were classified as countries with a low to medium level of sport, and 12 countries in clusters 2 and 3 were classified as countries with a high level of sport. Categorising the predictions for countries with different levels of sport significantly improves the accuracy of the predictions.

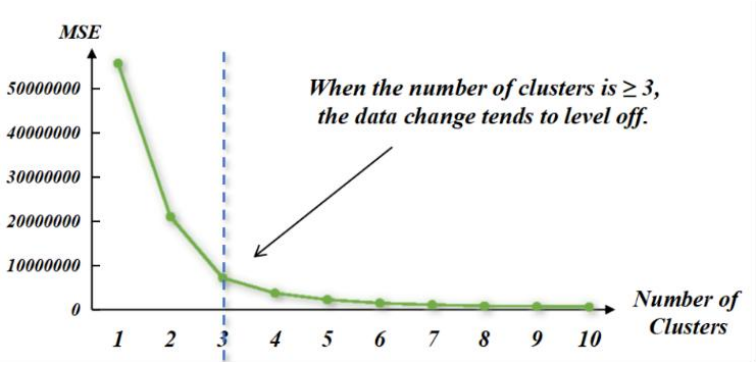

Figure 3. SSE Decreasing Trend Graph

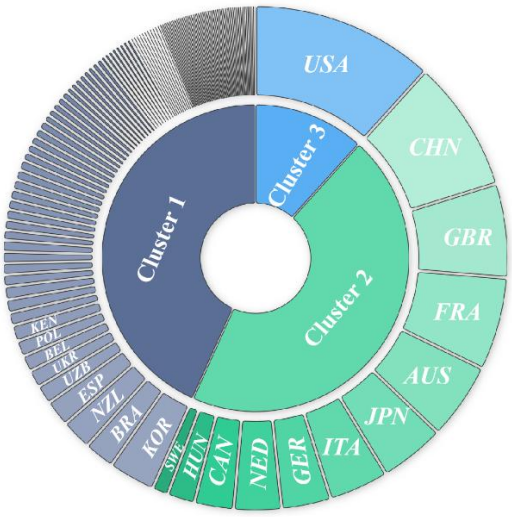

Figure 4. Clustering Results

3.2. LASSO Regression

For the development of a prediction model, LASSO regression was applied to countries with a high level of sport performance. The appropriate regularisation parameter λ = 0.02 was determined by cross-validation, and the minimum characteristic coefficient of the objective function was calculated in MATLAB and fitted to the total number of medals to obtain the regression equation:

$$y = 58.078 + \vec{n} \cdot \overrightarrow{Host_{t,i}} - 0.175 \times Pro + 0.45 \times Lev + 0.006 \times Ath + 0.527 \times Str. \tag{11}$$

where $\vec{n} = (-4.846, 20.178, 5.072)^T$, Pro represents the total number of projects offered in this event, Lev represents the overall level of athletes, Ath represents the number of participating athletes from each country, Str represents the comprehensive national strength.

After testing, the R^2 of the prediction model for the total number of medals is 0.969, which is close to 1. It can be considered that the model has a good fit.

The predicted results of the model are shown in Fig. 5, where the United States is predicted to win the most medals, with a predicted 147 medals, which is a reasonable prediction given its host effect and strong sporting prowess. The second highest ranked country is China, which is predicted to win 92 medals, which is basically the same as the results of the last edition of the event, and the predicted results of the remaining countries also vary less, making the predicted results more accurate.

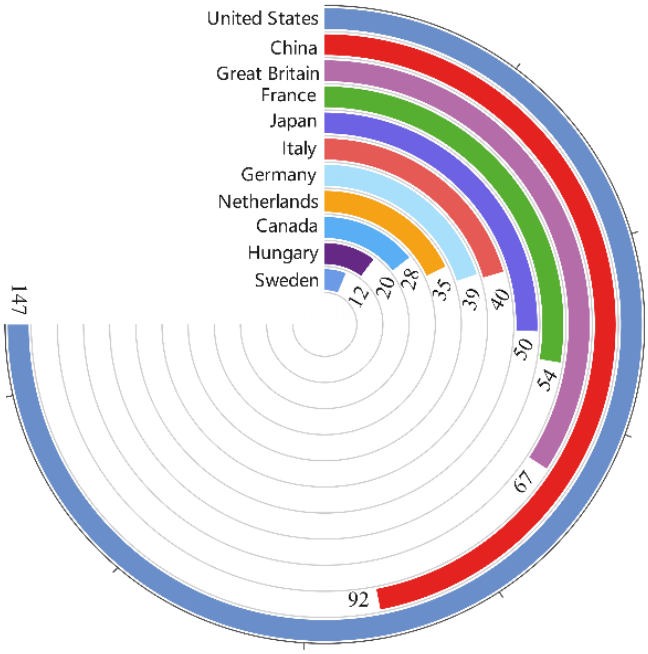

Figure 5. Predicted Number of Medals

3.3. Random Forest

A random forest prediction model was built for low and medium level countries in sport, and the dataset was divided into 70% training set and 30% test set, and the model was trained by grid search to obtain the optimal combination of parameters, as shown in TABLE II.

Table 2. Random Forest Model Parameter

Parameter name	Parameter value
Split ratio	0.7
Cross-validation	10
Maximum depth of the decision tree	10
Maximum number of leaf nodes	50
Number of decision trees	300

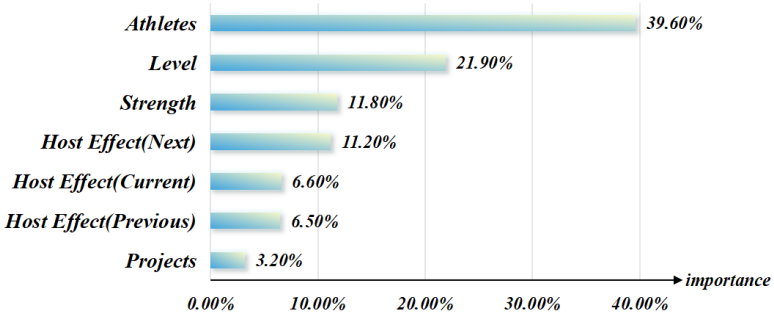

Figure 6. Indicator Importance Ranking Chart

The model is trained to predict the number of medals, while the importance ranking of each feature indicator can be obtained as shown in Fig. 6.

The most important indicator is the number of participating athletes, which accounts for 36.9% of the importance of this feature, meaning that the more athletes a country sends, the better its chances of increasing its total number of awards. The number of athletes also reflects, to some extent, a country's investment in sports development, including funding, training facilities, coaching teams, etc., all of which contribute greatly to its award-winning performance.

Second, it is the level of the athletes. The level of the athletes is the key factor that determines their performance in competition. Better skills, more physical fitness and more experience in the competition will give them a better chance of winning the award. Finally, there is the overall national strength, the host effect and the total number of projects offered. Most of the countries with a low or medium level of sport have a relatively weak national strength at a given level and are less likely to be hosts than the strongest countries in the sport, so the contribution of national strength and the host effect to the number of medals they win is not significant. The least important indicator is the total number of events, which accounts for only 3.2%, suggesting that its direct impact on the number of medals is relatively small. One possible explanation is that while the addition of a competition programme may increase a country's chances of winning medals, the allocation of resources and inputs to the additional competition programme must also be considered.

Table 3. Results of the Model Evaluation

	MSE	RMSE	MAE	MAPE	R^2
Training Set	4.476	2.116	1.200	50.983	0.950
Cross-validation set	8.964	2.917	1.95	111.161	0.907
Validation Set	3.855	1.936	1.442	7.174	0.881

The evaluation results of the model are shown in Table III, where the MSE, RMSE and MAE of the training set, cross-validation set and test set are small, and R^2 is close to 1. The model performs well. In addition, the MAE value of the model is 1.442, which is rounded up to give a prediction error interval of ±2 for the model.

According to the model, the prediction of the winners of the next event can be obtained for low and medium level countries in sports, and the prediction results for some representative countries are shown in Fig. 7.

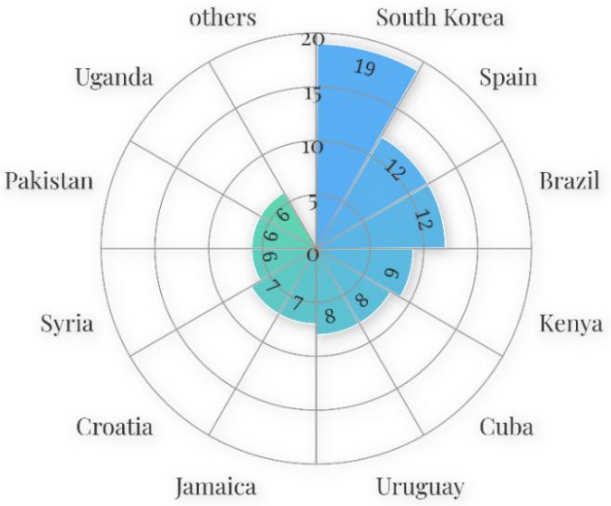

Figure 7. Predicted Number of Medals

To extend the dimension of prediction, in the same way as for the prediction of the total number of medals, the gold, silver and bronze medal data is used to predict the number of gold, silver and bronze medals each country will win at the next event, giving a complete predicted medal table. The predicted medal table for all medal-winning countries at the next event is shown in TABLE IV.

Table 4. Predicted Medal Table

NOC	Gold	Silver	Bronze	Total
United States	69	40	38	147
China	40	27	25	92
Great Britain	19	23	25	67
Australia	17	21	18	56
France	16	19	16	54
Japan	15	24	21	50
Italy	13	12	15	40
Germany	13	11	15	39
Netherlands	10	14	11	35
Canada	9	7	12	28
Hungary	7	6	7	20
...	...	...	...	...
Czechia	0	0	1	1
Namibia	0	0	1	1
Sudan	0	0	1	1

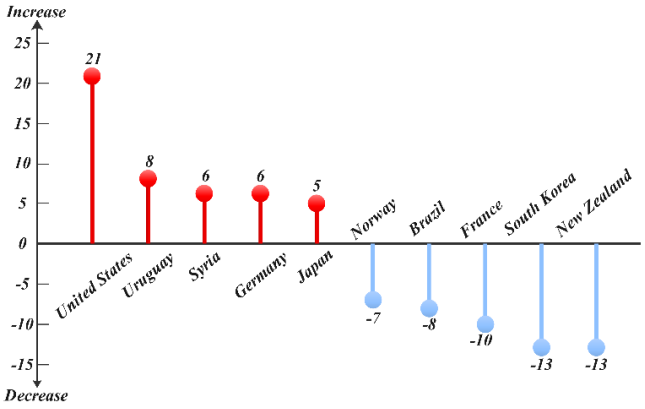

Figure 8. Visualisation of Progressing and Regressing Countries

Comparing and analysing the data of the predicted medal table and the medal table of the last event, it can be learnt that the countries with greater progress include the United States, Germany, Japan, etc., and the countries with greater regression include South Korea, New Zealand, France, etc., and the detailed progress and regression is shown in Fig. 8.

3.4. Model Performance Evaluation

3.4.1. Sensitivity Analysis

The number of decision trees in the random forest tends to have a significant effect on the model, with too few affecting the goodness of fit and too many causing overfitting. The number of decision trees in the random forest model was adjusted and classified into simple, medium and complex trees, the model was reconstructed and cross-validated tenfold and the results are shown in Table V.

Table 5. Cross-validation Results

TREES	MSE	RMSE	MAE	R^2
simple	27.347	4.112	1.003	0.673
medium	14.167	3.628	1.447	0.883
hard	22.928	4.787	1.882	0.901

The TABLE V shows that the range of variation of the MAE caused by the factor of the number of decision trees does not exceed 30%, so the effect on the number of medal predictions only fluctuates around one, proving that this model is more stable.

3.4.2. Robustness test

In reality, the level of athletes is very much affected by the pre-game state, so to test the robustness of the model, china and Spain are taken as high level countries and medium-low level countries, respectively, and their athletes are allowed to fluctuate up and down by 5% and 10%, and the percentage of fluctuation in the difference of the predicted results is calculated, as shown in Fig. 9. The results show that the error coefficient of medal prediction basically stays within 5%, so the model has a strong anti-interference ability to external unstable factors such as athletes' level, and passes the stability test.

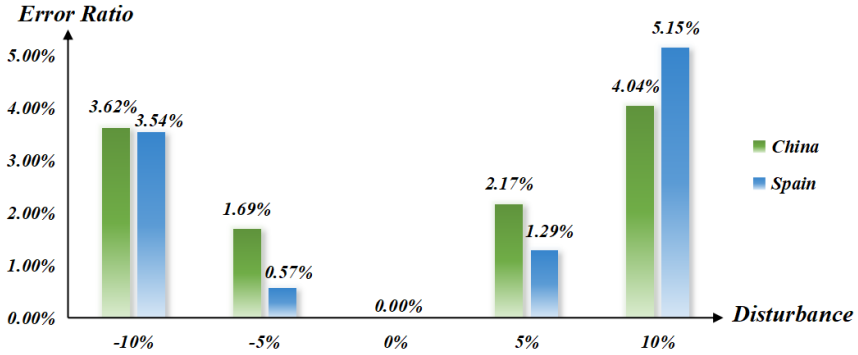

Figure 9. Stability Testing

4. Conclusion and Outlooks

4.1. Conclusion

Based on Hierarchical Clustering, LASSO Regression and Random Forest, this study constructs a sports medal prediction model with strong interpretability. Through the self-constructed multidimensional sports competitiveness evaluation system, the study innovatively quantifies and downscales the core event indicators, and presents the predicted results of the events in 2028. The clustering analysis of countries based on the indicators accurately classifies the participating countries into three differentiated stages of development; the analysis of the model summarises the importance weights of each sports indicator and quantifies the promoting or inhibiting effects of the explanatory variables on the results of the study; and the error intervals and accuracy of the model are explored through cross-validation. This study provides a quantifiable decision support framework for optimising resource allocation in competitive sport. After transformation through feature engineering, the hybrid modelling paradigm of this study can also be extended to the fields of economic forecasting and business competitiveness assessment.

4.2. Outlooks

Although LASSO regression has been used for variable screening in the feature engineering stage of this study, there is still the problem of potential omitted variables, which is manifested in the fact that the heterogeneous characteristics of the dominant items in each country have not been fully taken into account, resulting in the limited ability of the predictive model to capture the specialised sporting performances, so that in the future the dominant item weight matrix can be introduced to establish the predictive sub-models of the item-specific characteristics.

References

[1] 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.

[2] Badoni P, Choudhary P, Rudesh C P, et al. Predicting Medal Counts in Olympics Using Machine Learning Algorithms: A Comparative Analysis[C]//2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech). IEEE, 2023: 116-121.

[3] Tchamkerten A, Chaudron P, Girard N, et al. Career factors related to winning Olympic medals in swimming[J]. PLoS One, 2024, 19(6): e0304444.

[4] Yeh C C, Peng H T, Lin W B. Achievement Prediction and Performance Assessment System for Nations in the Asian Games[J]. Applied Sciences, 2024, 14(2): 789.

[5] Scelles N, Andreff W, Bonnal L, et al. Forecasting national medal totals at the Summer Olympic Games reconsidered[J]. Social science quarterly, 2020, 101(2): 697-711.

[6] Otamendi F J, Doncel L M, Martín-Gutiérrez C. Meeting expectations at the 2016 Rio Olympic games: country potential and competitiveness[J]. Social Science Quarterly, 2020, 101(2): 656-677.

[7] Li F, Hopkins W G, Lipinska P. Population, economic and geographic predictors of nations' medal tallies at the Pyeongchang and Tokyo Olympics and Paralympics[J]. Frontiers in Sports and Active Living, 2022, 4: 931817.

[8] Wunderlich F, Memmert D. Forecasting the outcomes of sports events: A review[J]. European journal of sport science, 2021, 21(7): 944-957.

[9] Baumer B S, Matthews G J, Nguyen Q. Big ideas in sports analytics and statistical tools for their investigation[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2023, 15(6): e1612.

[10] Wilkens S. Sports prediction and betting models in the machine learning age: The case of tennis[J]. Journal of Sports Analytics, 2021, 7(2): 99-117.