Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 6, No. 1, 2025

Research on Olympic Medal Prediction Based on Random Forest Regression

Ye Zhu[1, *, #], Yuhang Zheng[2, #], Yufei Mao[3, #]

[1]College of Transportation Engineering, Chang'an University, Xi'an 710064, China

[2]School of Electrical and Control Engineering, Chang'an University, Xi'an 710064, China

[3]School of Highway, Chang'an University, Xi'an 710064, China

[*]**Corresponding author**: Ye Zhu (Email: 2022902533@chd.edu.cn)

[#]**These authors contributed equally.**

Abstract: The distribution of Olympic medals serves as a critical measure of national sports competitiveness. This study proposes a predictive model based on Random Forest Regression and feature engineering, trained on historical Olympic data (1896-2020) and validated using 2024 Paris Olympics data. Key features, including athlete participation, host country status, historical medal averages, and elite athlete count, were selected through data preprocessing, with model stability assessed via 5-fold cross-validation. Results demonstrate that the Random Forest Regression model achieved a coefficient of determination (R^2) of 0.96, with mean squared errors of 0.94 (gold medals) and 3.09 (total medals), outperforming multiple linear regression and support vector regression models. Quantitative analysis reveals a 5%-10% host country advantage in medal probability and a 70% breakthrough potential for non-winning countries through sports diplomacy and elite coaching. These findings provide a data-driven framework for optimizing resource allocation and strategic planning in international sports governance.

Keywords: Random Forest, Logistic Regression, Pearson Correlation Analysis, Feature Engineering Contents.

1. Introduction

The Olympic Games, which originated in ancient Greece, have transformed into a premier global event that celebrates athletic excellence across a multitude of sports. Influenced by a country's sports traditions, resource investments, and training systems, the medal table serves as a clear measure of national sports competitiveness. A notable example is the 2024 Paris Olympics, where the United States led in total medals, while China and the United States tied for the most gold medals[1].

Early studies by Xiong[2] identified economic strength, political stability, and sports infrastructure as primary drivers of Olympic success, noting that socialist systems and hosting advantages further amplify medal outcomes. These findings align with Andreff's[3] econometric analysis, which highlighted GDP per capita, population, and host-country effects as statistically significant predictors of medal totals in both Summer and Winter Olympics. Andreff's model for Winter Games additionally incorporated snow coverage and winter sports facilities, demonstrating how geographic and economic resources shape competitive advantages. a study by Clarke[4] emphasized that host nations typically secure 1.8% more medals due to familiarity with local conditions and heightened national motivation. This host effect is quantitatively supported by regression analyses showing a 5–10% increase in medal probabilities for local athletes[5]. Conversely, economic downturns, such as Spain's reduced sports funding post-2008, correlate with declining medal rankings, underscoring the financial dependency of competitive success.

This paper constructed a model to predict the number of MEDALS of each country and evaluate the uncertainty and accuracy of the model prediction as well as indicators to measure the performance of the model. Then the model predicted the medal table of the 2028 Olympic Games and the confidence interval of the predicted results. This model studied the relationship between Olympic event and the number of MEDALS won by each country. Explore the impact of the host country's selection of events on the medal performance of each country. The change of medal number in each country was analyzed. Finally this paper use the model to uncover novel insights into the Olympic medal count and explain how it can serve as a reference for national Olympic Committees.

2. Preliminary

2.1. Assumption

(1) The ability and trend of countries to win MEDALS are consistent with history, regardless of special events[6].

(2) The various events are independent of each other. The results of one event do not affect those of the others.

(3) Athletes performed normally during the competition, and the participation of athletes from various countries was stable.

(4) A great coach can achieve similar improvement effects on athletes from different countries and in different sports, and such effects remain consistent over different time periods, being minimally affected by changes in the external environment[7].

2.2. Random Forest

Random Forest, a cornerstone of modern machine learning, is an ensemble algorithm that combines multiple decision trees to enhance predictive accuracy and model robustness. Developed by Leo Breiman and Adele Cutler, this algorithm leverages two core principles: Bootstrap Aggregating (Bagging) and random feature selection, which collectively mitigate overfitting and improve

generalization[8]. By aggregating predictions from hundreds to thousands of trees, Random Forest achieves superior performance across classification and regression tasks, as evidenced by its widespread adoption in fields such as bioinformatics, finance, and medical diagnosis. Figure 1 shows the structure of Random Forest.

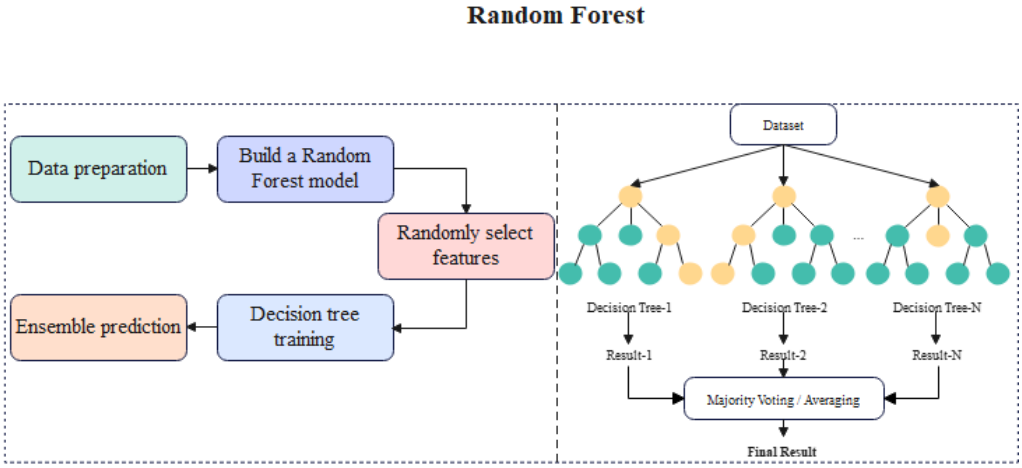

Figure 1. Random Forest

2.3. Notations

Table 1 systematically summarizes the key symbols and their definitions, including variables such as the annual number of athletes per country , participated events, and host country status. These symbols quantify critical factors in historical Olympic data, such as elite athlete participation and historical medal averages, establishing a structured foundation for the Random Forest regression model. The standardized design of the table enhances the interpretability of model variables and supports consistent feature engineering and data preprocessing, ensuring methodological rigor and reliability in subsequent analyses.

Table 1. Notations

Symbols	Definitions
N_Ac,t	Number of athletes competing in Country c in year t
N_Ec,t	The number of events in which Country c competed in year t
Hc,t	Whether country c is the host country in year t
G_Ac,t	Number of elite athletes competing in Country c in Year t (athletes who have won gold MEDALS)
T_Ac,t	Number of elite athletes competing in Country c in Year t (athletes who have won MEDALS)
H_A_Tc,t	The average number of MEDALS won by country c in year t in the previous three Olympic Games
H_A_Gc,t	In year t the average number of gold MEDALS won by country c in the previous three Olympic Games
Gc,t	Number of gold MEDALS in country c in year t
Gc,t	The total number of medals of country c in the t-th year

3. Medal counting model based on random forest regression

Through checking for gaps and making up for omissions, this study aggregating athlete information data with medal statistics to ensure the data accuracy and consistency. The random forest regression model is used to predict and compare the data of 2024 with the real value under the training set. The model performance is predicted by the prediction set and the uncertainty is analyzed by the 5-fold cross test.

3.1. Data Preprocessing

By analyzing the given data, the present research find that the number of MEDALS in each country is affected by a variety of factors, such as: The number of gold MEDALS may be related to the number of athletes participating, the number of sports participating, whether it is the host country, the type of sports participating, the number of elite athletes participating, the number of competitive sports participating, and the number of historical gold MEDALS[9].

Through data mining and feature engineering analysis, the analysis select the number of participating athletes, the number of participating events, whether it is the host country, the number of partici-pating elite athletes and the average number of gold MEDALS won in the first three sessions as the characteristic variables of the number of gold MEDALS in each country.

$$G_{c,t} = \left(N\_A_{c,t}, N\_E_{c,t}, H_{c,t}, G\_A_{c,t}, H\_A\_G_{c,t} \right) \tag{1}$$

Similarly, it can be concluded that the characteristic variables of the total number of medals in each country are the number of athletes participating in the competition, the number of participat-ing projects, whether it is the host country, the number of outstanding athletes participating in the competition, and the average total number of medals won in the first three sessions[10].

$$T_{c,t} = f\left(N\_A_{c,t}, N\_E_{c,t}, H_{c,t}, T\_A_{c,t}, H\_A\_T_{c,t}\right) \tag{2}$$

Through correlation analysis, the model can get the degree of correlation between each characteristic variable and the dependent variable. It can be found that most of the characteristic variables have a strong correlation with the dependent variable. Only whether it is the host country has a weak correlation with the dependent variable. It may be because the variable only has a positive impact on the host country of each Olympic Games, and the impact on other countries is relatively small. It is shown in Figure 2 and 3.

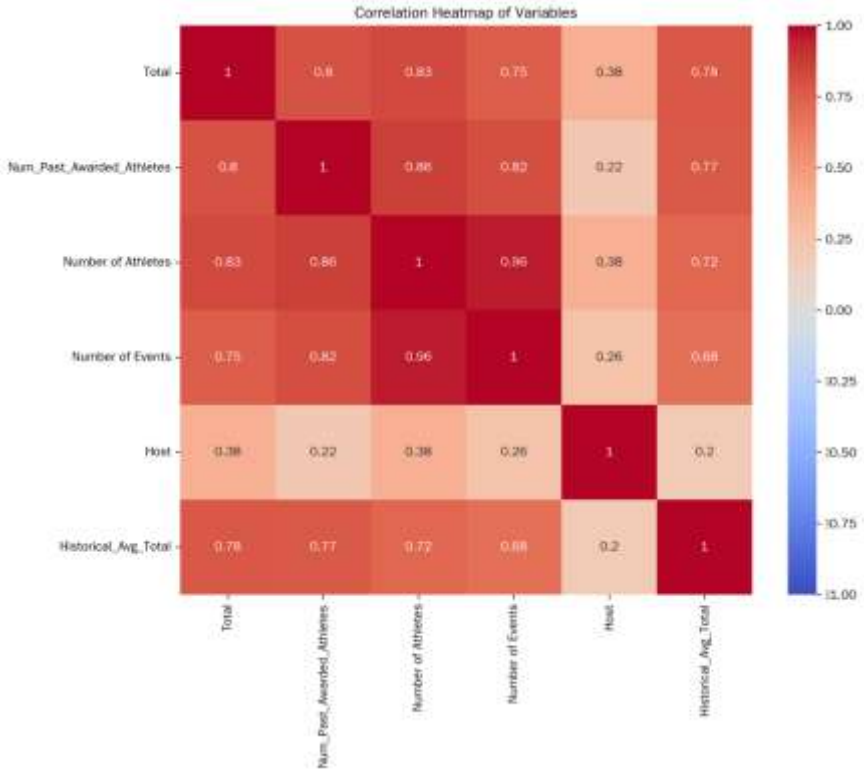

Figure 2. Correlation between gold medals and feature variables

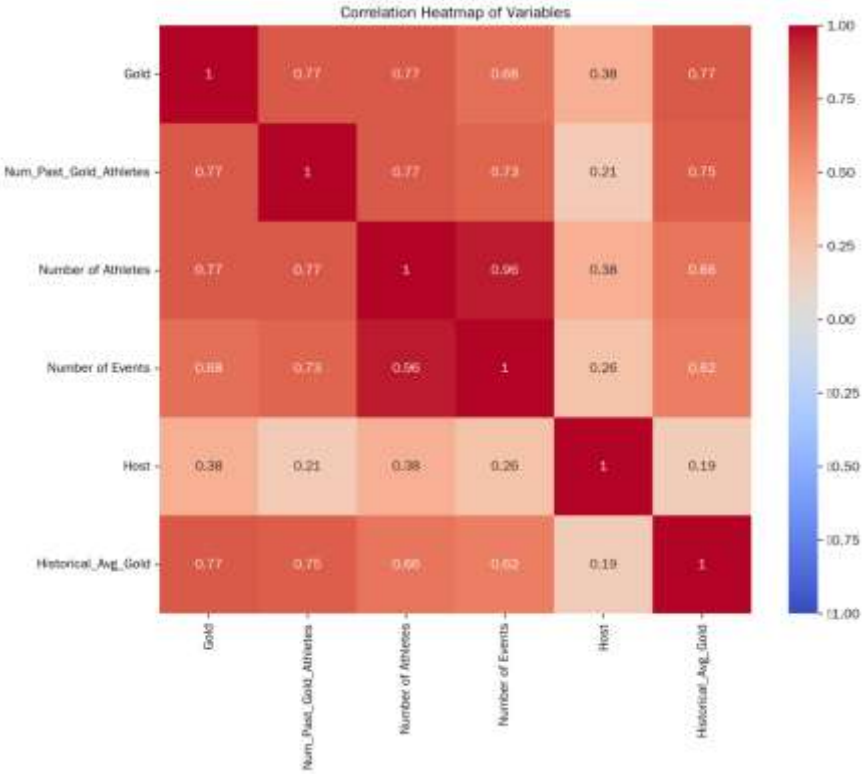

Figure 3. Correlation between total medal count and characteristic variables

3.2. Regression Model

By training different regression models with eigenvalues: multiple linear regression, random forest regression and support vector regression (SVR), it is found that the random forest regression model has the largest coefficient of determination (R2) among these models, and it can be shown that the model has a good fitting effect on the data. Therefore, this paper choose the random forest regression model and the random forest regression model.

Table 2. Coefficient of determination of different model fit

	Multiple linear regression	Random forest regression	support vector regression
R2	0.77	0.96	0.94

The stochastic forest regression model is an ensemble learning model, which is composed of several decision trees. During construction, multiple sample subsets were randomly selected from the original data set through bootstrap sampling, and a decision tree was built for each subset. When nodes split, some features are randomly selected to determine the optimal splitting condi-tions, so as to reduce the correlation between decision trees and improve the generalization ability of the model. Finally, the average of the prediction results of all decision trees is taken as the final prediction value, effectively avoiding the overfitting problem of a single decision tree, and greatly improving the generalization ability and prediction accuracy of the model.

Assume the independent variable vector X=(x1 , x2 , x3 ...). The dependent variable y represents the total number of gold MEDALS or MEDALS for each country in the same expression (1) and (2).

The ultimate purpose of the model is to predict the 2028 medal table, so as to prevent insufficient training data from leading to underfitting of the fitted model, which can only achieve a local optimal solution and reduce the prediction accuracy of the 2028 medal table. In order to improve the fitting degree of the model, this study adopted a special data set partitioning method, using most of the data in 2020 and before as a training set for the fitting training of the model, so that it can learn the patterns and rules in the historical data. The 2024 data is used as a test set to evaluate the predictive performance of the model. Specifically, Olympic data from 1896-2020 were selected as the training set. Using the 2024 Olympic data as a test set. This partitioning method not only ensures that the training set has enough data to train the model, but also can use the relevant test set to accurately evaluate the performance of the model.

Based on the divided training set, the random forest regression model is trained in Figure 4 and 5.

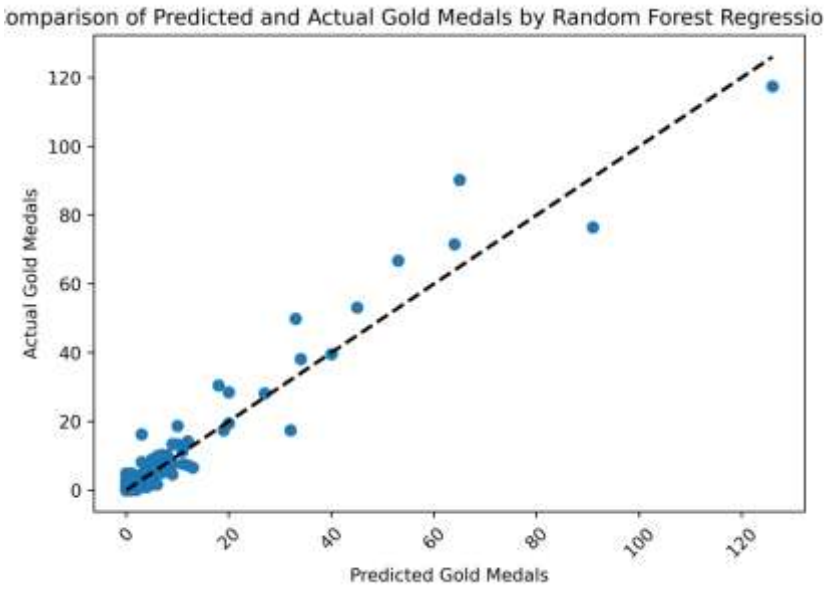

Figure 4. Gold medal from RF

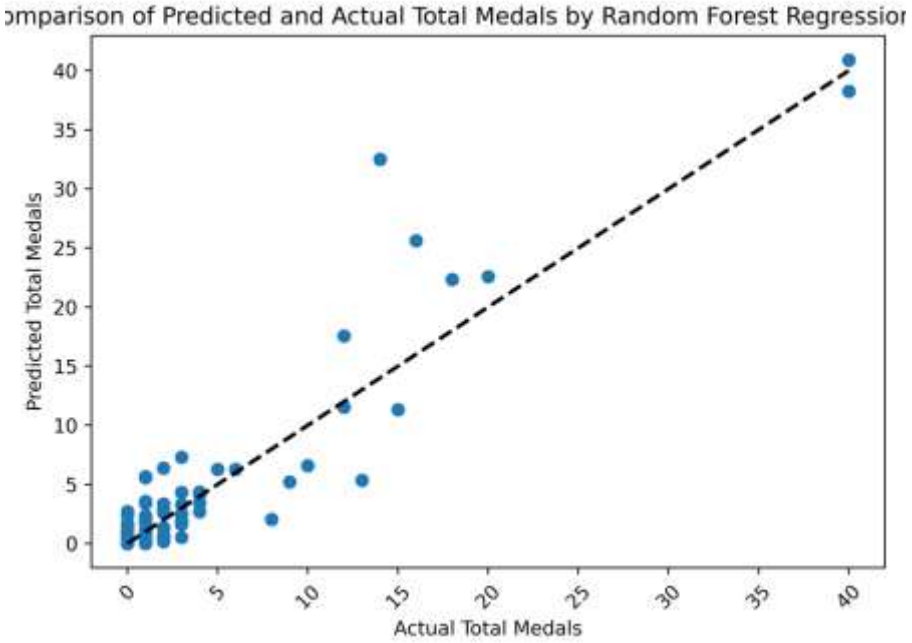

Figure 5. Total medal from RF

After the training of the model, the independent variable data from the test set (2024 data) is input into the trained random forest model, and the model outputs the predicted value of the number of gold MEDALS and the total number of MEDALS for each country in the 2024 Olympic Games. It is shown in Figure 6 and 7.

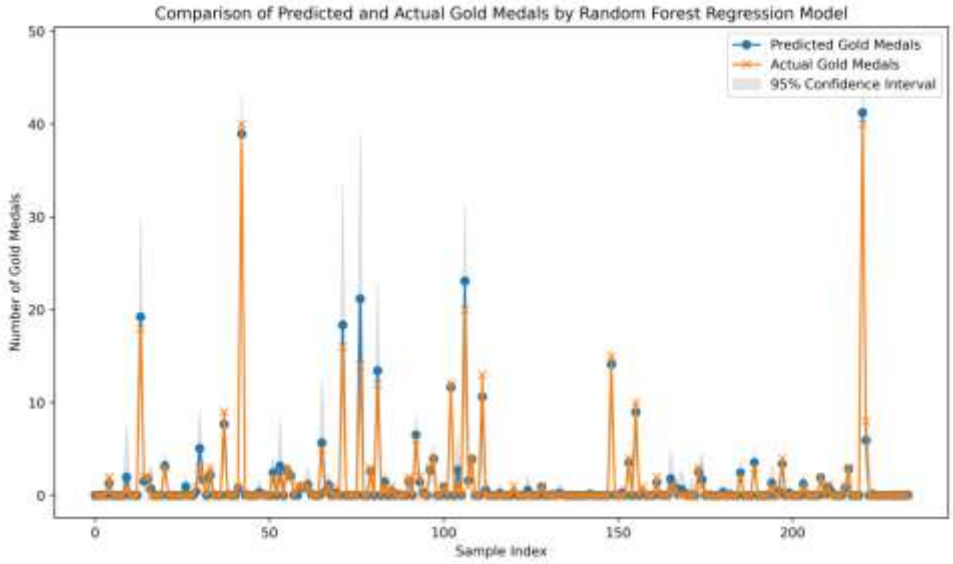

Figure 6. Predicted and actual gold MEDALS for each country in 2024

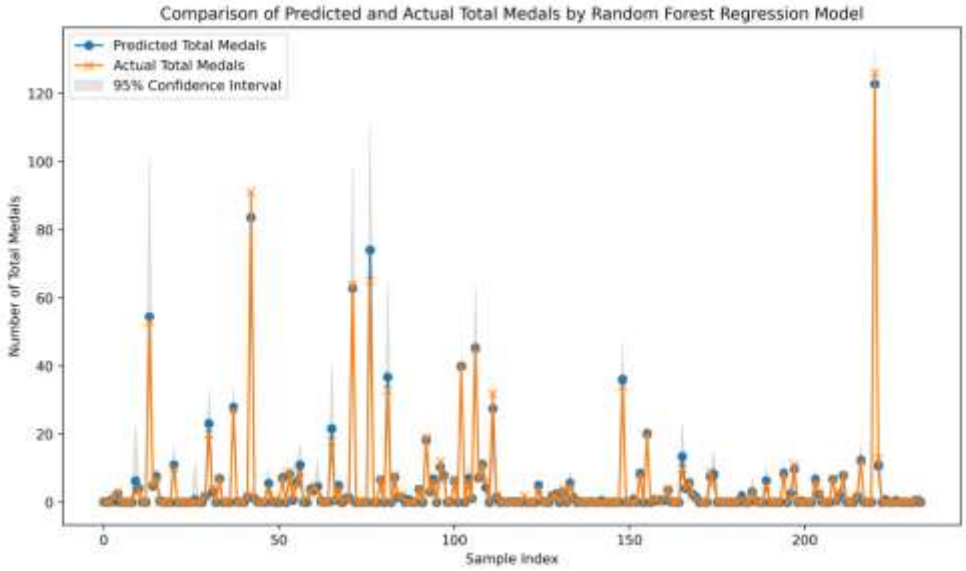

Figure 7. Comparison of projected and actual total MEDALS for each country in 2024

The random forest regression model was evaluated using test set data, and the performance of the model was measured by mean square error (MSE), mean absolute error (MAE), determination coefficient (R2) and root mean square error (RMSE).

Table 3. Coefficient of determination of different model fit

	MSE	MAE	R2	RMSE
Gold_mdl /Total_mdl	0.94/3.09	0.36/0.75	0.96/0.98	0.97/1.76

Table 3 shows that both the gold medal regression model and the total medal regression model get a good performance on the test set. The uncertainty and accuracy of the model are estimated by the method of 5-fold cross test. Table 4 shows that the reflects that the model has a good precision in regression fitting.

Table 4. Accuracy of Model by Cross-validation

	MSE	MAE	R2	RMSE
Gold_mdl /Total_mdl	0.94/3.09	0.36/0.75	0.96/0.98	0.97/1.76

3.3. Strengthen the degree of sports diplomacy

According to the medal table, it can be seen which countries are powerful in sports, which countries are outstanding in a particular event. Countries with weak sports can send coaches and athletes to strong sports countries for training and exchange; A country wishing to win an Olympic medal in a particular sport may apply for a friendly match against a country that has won a medal in that sport at the previous Olympics.

National Olympic Committees can actively promote domestic sports activities and cooperation, strengthen communication and coordination with international sports organizations, and strive for a favorable international environment for the development of the country's sports industry. Through international sports cooperation and exchanges, they can enhance the country's influence on the world sports stage and create conditions for winning more Olympic medals. It is shown in Figure 8.

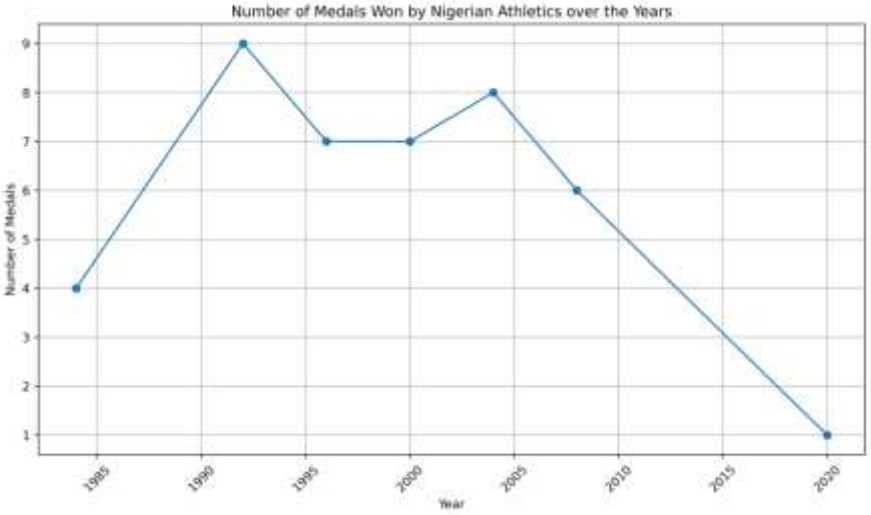

Figure 8. Number of medals won by Nigerian Athletics over the years

4. Model evaluations

4.1. Sensitive analysis

The model uses most of the data in 2020 and before as the train-ing set, and the data in 2024 as the test set. If the data partitioning is changed, such as increasing or reducing the amount of data in the training set, it may affect the model's learning of historical data patterns and rules, and then affect the model's prediction performance and stability. If the training set data amount is too small, the model may not be able to fully learn the complex relationship in the historical data, resulting in the prediction accuracy of 2028 medal table is reduced. If the training set data is too large, more noisy data may be introduced, which also affects the model performance. When predicting the eigenvalues of countries in 2028, methods such as weighted moving average method are used. Different weight Settings or prediction methods may lead to different eigenvalue prediction results, which will affect the final medal number prediction. For example, when calculating the predicted value of the number of participating athletes and the number of participating events, if the weighted average weight is changed, the corresponding characteristic value calculated in 2028 will change. After substituting the random forest regression model, the predicted gold medal number and total medal number in 2028 will also change. The regression coefficient and intercept in a logistic regression model determine the specific form of the model and the prediction results. These parameters are estimated by training data. If the training data changes or different estimation methods are used, the regression coefficients and intercepts obtained will be different, which will affect the accuracy of the model's prediction of the probability of non-winning countries winning MEDALS for the first time.

In the analysis of the relationship between events and MEDALS won by countries, the country with the largest increase in gold MEDALS is selected for analysis and is the host country of the year. If the data screening criteria are changed, such as selecting a country with a larger increase in the number of gold MEDALS but not the host country of the year, or taking into account changes in the number of silver and bronze MEDALS, the relationship between the resulting events and the number of MEDALS may be different.

4.2. Strength and Weakness

By comparing multiple linear regression and support vector regression models, the coeffi-cient of determination is up to 0.96, and the fitting effect is good; In the test set, the MSE and MAE of the regression model of gold medal and total medal are lower and higher, and the prediction accuracy is higher. Although the error index increases under the 5-fold cross test, it still maintains a certain level and has a certain generalization ability. It can calculate the importance of features and clarify the influence of each factor on the number of MEDALS. On some characteristic value using a weighted average of the forecast, correlation and data time series, the reasonable method; The predicted results can reflect the changing trend of sports competition level in different countries and are related to actual factors. During data preprocessing, the missing value is filled with 0, and the abnormal value is deleted. The selection of characteristic variables may not be comprehensive; The error index of 5 fold cross test increases, and there is uncertainty in the model. The weighted average of the set of subjective weighting, affects the characteristic value and medal predictions; Comparing only 2024 and 2028 data, without taking into account more factors, the forecast results are unstable.

5. Conclusion

This paper synthesizes the key findings and practical implications derived from applying the Random Forest regression model to predict and analyze Olympic medal patterns. By leveraging feature engineering to identify critical variables, the study demonstrates that non-winning nations exhibit significant medal potential, with some achieving predicted medal probabilities exceeding 70%. This underscores the importance of targeted investments in sports infrastructure and talent development for emerging nations. The analysis further reveals the substantial influence of hosting the Olympics on medal counts, highlighting the "host advantage" as a critical factor in enhancing national performance. These findings not only advance the methodology for predicting Olympic success but also offer policymakers and sports federations a data-driven framework to optimize performance and foster global competitiveness in international sporting events.

References

[1] Piepiora P A, aplová, Petra, Zimoń, Pawe, et al. On research into the relationship between personality traits and the sporting level of competitive, professional and elite athletes[J]. Frontiers in Psychology, 2024.

[2] Tsukahara Y, Mountjoy M, Takazawa Y, et al. Analysis of the Characteristics of Patients Visiting the Tokyo 2020 Olympics Polyclinic[J] Clinical Journal of Sport Medicine, 2025, 35(1):7.

[3] Tomioka N. Sakurai Medal No. 46 Akimotoite, a new mineral discovered in a shocked meteorite[J]. Japanese Magazine of Mineralogical and Petrological Sciences, 2024, 53(1):4.

[4] Cutler L, Lewis W. Portraits of Fatherhood: Depictions of Fathers and Father–Child Relationships in Award-Winning Children's Literature[J]. Children's Literature in Education, 2024, 55(4).

[5] Lvarez-Yates T, Iglesias-Caamao M, Cuba-Dorado A, et al. Explanatory Model for Elite Canoeists' Performance Using a Functional Electromechanical Dynamometer Based on Detected Lateral Asymmetry[J]. Symmetry (20738994), 2024, 16(3).

[6] Leck R. The William Houston gold medal of the Royal College of Surgeons of Edinburgh 2022: Two case reports[J]. Journal of orthodontics, 2024, 51(3):292-306.

[7] Lim H B T, Ong N C H. The Relationship between Perceived Parenting Styles and Youth Athletes' Sporting Achievement in Singapore[J]. Youth (2673-995X), 2024, 4(1).

[8] Cheng S C, Chang C, Chen Y C, et al. Human gut microbiome determining athletes' performance: An insight from genomic analysis[J]. Ecological Genetics and Genomics, 2025, 34.

[9] Nci H, Inci F. Depression, Anxiety, and Pain Catastrophizing in Migraine Patients[J]. Acta Medica Alanya, 2023.

[10] Mohamed A E, Ashraf G, Azmi A A B M. Contribution of Fluid Substitution and Cheetah Optimizer Algorithm in Predicting Rock-Physics Parameters of Gas-Bearing Reservoirs in the Eastern Mediterranean Sea, Egypt[J]. Natural resources research, 2023, 32(5):1987-2005.