Stacking integration algorithm Based on Regression models and Informer for Olympic Medal Standings Forecasting

Kaidi Wang [1], Yunqing Guo [2,*], Shiyuan Chen [3]

[1] Electronic and Information Engineering, Liaoning Technical University, Huludao, China, 125105

[2] Faculty Electrical and Control Engineering, Liaoning Technical University, Huludao, China, 125105

[3] Safety Science and Engineering College, Liaoning Technical University, Huludao, China, 125105

* Corresponding Author Email: 18749081324@163.com

Abstract. As the most influential sports event in the world, the Olympic Games is a wind vane for evaluating a country's sports strength, resource allocation, development strategy and competitive level. With the end of the 2024 Paris Olympics and the approach of the 2028 Los Angeles Olympics, scientific prediction of the number of medals has become a topic of great interest. However, challenges such as the emergence of first-time medal-winning countries, the host country effect, and the impact of additional events complicate the prediction process. This study proposes a comprehensive medal prediction framework that integrates multiple data science methods. The study introduces a novel country classification method based on K-means++ clustering, which categorises the participating countries into sports powerhouses and emerging countries, thus greatly improving the model's adaptability and prediction accuracy. In addition, the study adopts a hybrid modelling approach that combines multiple regression models (e.g., Random Forest, XGBoost) with time series models (Informer) through stacked integration to improve robustness and generalisation. For prediction uncertainty, the framework calculates 95% confidence intervals and incorporates prediction bias analysis to validate model stability. This study provides a robust and reliable tool for Olympic medal forecasting, offering valuable insights to policy makers and sports analysts.

Keywords: Random Forest, XGBoost, Informer, Mixed-effects model.

1. Introduction

The Olympic Games represent the pinnacle of international sports competition, reflecting not only a nation's sporting prowess, but also its ability to allocate resources, implement development strategies and achieve athletic excellence. As the 2024 Paris Olympics draw to a close and the 2028 Los Angeles Olympics approach, the ability to scientifically and accurately predict medal results is increasingly in the spotlight. Accurate medal forecasts are critical to understanding the dynamics of global sporting performance and to informing strategic decision-making. However, the forecasting process is fraught with complexities, including the rise of first-time medallists, the dominance of the host nation, and the impact of new sports introduced in each Olympic cycle. These factors introduce a great deal of uncertainty, making traditional forecasting methods seem overwhelming.

To address these challenges, this study proposes a comprehensive medal forecasting framework using advanced data science techniques. The innovation of this study on the national medals dataset is the introduction of a country classification method based on K-means++ clustering [1], which categorises the participating countries into sports powerhouses and emerging countries. This classification method enhances the model's ability to adapt to data heterogeneity and improves prediction accuracy.

In terms of innovation in model construction, the study used a stacking approache. combining regression models (e.g., Random Forest. XGBoost) and time series models (Informer) through a stacked integration method [2-5]. This fusion creates a mixed-effects model that significantly improves robustness and generalization. Additionally, to address forecast uncertainty, the framework computes 95% confidence intervals and incorporates prediction bias analysis to validate model stability. By integrating these methods, this study provides a robust and reliable tool for Olympic

medal forecasting, offering valuable insights to policy makers, sports analysts and stakeholders. The findings contribute to the growing field of sports analytics and lay the foundation for future research in predictive modelling of international sporting events.

2. Feature engineering

The Olympic medals data over the years show that some countries have long and well-resourced sports systems and therefore have won a large number of medals in international competitions such as the Olympic Games. On the other hand, some countries tend to rank lower in the medals table for various reasons, notably limited resources and sports systems that are still under construction. This disparity is reflected not only in the number of medals but also in the dominance of certain sports. Because of this, we find that the level of sports development varies considerably from country to country.

Data discrepancies, if not handled properly, may have a significant impact on the construction of subsequent prediction models. For example, large differences between data may trigger oscillation problems during model training, leading to difficulties in the smooth convergence of the model, which in turn leads to distorted prediction results. These problems will not only reduce the validity of the prediction, but also weaken the generalisation ability of the model, which cannot be of strong practical guidance.

The dataset used in this study was obtained from official and authoritative sources, including the official Olympics website [6] and athlete profiles [7-8]. These sources provide reliable and comprehensive records of Olympic medal distributions and athlete achievements. To ensure data consistency, we preprocessed the raw data by removing duplicates and filling missing values using interpolation techniques. The cleaned dataset was then used to train and validate our prediction models.

2.1. K-means++ cluster analysis

K-means++ algorithm is an improvement of K-means algorithm, which discards the idea of random initialisation when initialising cluster centres, and its principle of choosing cluster centres is to initialise cluster centres as far away from each other as possible.

We consider the classification of countries with different levels of sports development in order to solve the above problems. Firstly, we count the number of gold, silver and bronze medals of each country, and then use these three elements as classification indexes to classify countries into sports strong and sports weak countries, where the K value is set to 2.

2.2. Derive clustering centres

By using the number of gold, silver and bronze medals of each country as an indicator, an algorithm is used to classify the countries into two categories: strong sports countries and weak sports countries.

We derived the clustering centre as shown in Table.1.

Table 1. Cluster centres

Cluster centres				
	Gold medal	Silve medals	Bronze medals	classification
Cluster centre 1	318.3636	288.0000	289.7272	Maturing sports countries
Cluster centre 2	15.0784	17.0523	20.2941	Emerging sports countries

Explanatory note on country classification categories:

Strong sports countries are generally characterised by high production of gold, silver and bronze medals. Weak sports countries are generally characterised by low production of gold, silver and bronze medals.

2.3. Classification results show

The clustering results using K-means++ for established sports countries are shown in Table 2, and the clustering results for emerging sports countries are shown in Table 3 (due to space constraints, this paper only shows clustering results for 20 countries):

Table 2. Maturing sports countries

Maturing sports countries	
1	United States
2	Soviet Union
3	China
4	Great Britain
5	France
6	Italy
7	Germany
8	Japan
9	Hungary
10	Australia

Table 3. Emerging sports countries

Emerging sports countries	
11	Serbia and Montenegro
12	Sri Lanka
13	Sudan
14	Taiwan
15	Tanzania
16	Togo
17	Tonga
18	Turkmenistan
19	Virgin Islands
20	Zambia

Through cluster analysis, we were able to classify countries into two categories. This categorization provides a basis for subsequently improving the accuracy of the prediction model. For sports-developed countries, due to their strong sports capital investment, coaching resources, and economic development, they have a great advantage in many sports, and we will focus on this point. As for the emerging sports countries, although they are backward in sports resources and weak in economic strength, they have certain potential in certain projects, and we will focus on their project potential and development trend. By building specific prediction models for these two countries separately, we can effectively avoid errors in the prediction results caused by large data differences, thus improving the accuracy and effectiveness of the models.

3. Fundamentals of Regression Integration Modelling

This study required the development of a medal prediction model to predict the number of medals for different countries. To solve this problem, we developed two models to predict the number of medals for different countries based on the classification of old and emerging sports countries: a regression model and a time series model. After feature engineering the data, we found that the variables to be considered were continuous, categorical and time series variables, so we built regression and time series models for different variables and weighted and fused the two models together for prediction [9].

3.1. Selection of regression models

After experiments we found that when selecting a single tree model to train on the dataset, although it has a good fit on the training set, it performs poorly on the new dataset, i.e., an overfitting state occurs. In order to improve the generalisation ability of the model, we consider combining multiple tree models by using Stacking to integrate the model, using the predictions of multiple tree models as new inputs and fusing the respective predictions with a new model. This model significantly improves the generalisation ability of the model and performs well on new datasets.

3.2. Selection of base model

3.2.1. Linear Regression

Linear regression is a statistical method used to model the linear relationship between input characteristics and a continuous target variable [10-11]. It minimises the error between predicted and actual values by fitting a straight line (or hyperplane). The basic formula is as follows:

$$\hat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b \tag{1}$$

where $w_1, w_2, \ldots, w_n$ are the weights of the model, $x_1, x_2, \ldots, x_n$ are the input features, and b is the bias term. Linear regression optimises the model parameters by minimising the loss function (usually the mean square error) with the objective function:

$$Obj = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

3.2.2. Random Forest

Random Forest is an integrated learning method based on decision trees. It improves the generalisation of the model by constructing multiple decision trees and voting (classification) or averaging (regression) their results. The objective function for its optimisation is as follows

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) \tag{3}$$

where L is the loss function, usually the mean square error.

3.2.3. XGboost

XGBoost is an efficient implementation of GBDT that significantly improves the performance and speed of the model by introducing techniques such as regularisation, parallel computation and second-order derivative optimisation. The model is efficient, flexible, and supports custom loss functions. With strong generalisation and overfitting resistance, it is widely used in data science competitions. The optimisation objective of the model is:

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \\ &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \end{aligned} \tag{4}$$

3.2.4. SVR (Support Vector Regression)

SVR is a regression method based on Support Vector Machines (SVMs) [12]. It does this by finding a hyperplane such that most of the data points fall within a certain tolerance range and is able to deal with non-linear relationships by means of a kernel function while minimising the error. Its mathematical formula is as follows:

$$Obj = \sum_{i=1}^{n} |y_i - \hat{y}_i|^2 + C \sum_{i=1}^{n} \text{ò}_i \tag{5}$$

Where 1 represents the error term, it is the hyperparameter that regulates the complexity of the model.

3.2.5. KNN (K-Nearest Neighbors)

KNN is an instance-based learning method [13]. It performs classification or regression by calculating the distance between the samples to be predicted and the K nearest neighbours in the training set. And it has high computational complexity, especially on large datasets, with no assumptions on data distribution. Its specific principle is as follows:

$$Obj = \sum_{i=1}^{n} \left(y_i - \frac{1}{K} \sum_{k \in N_i} y_k \right)^2 \tag{6}$$

where N_i refers to the K nearest neighbours to sample i and $\frac{1}{K} \sum_{k \in N_i} y_k$ is the average of these neighbours, which will be used as the predicted value.

3.3. Adjustment of base model parameters

For the model we have selected, we have mainly used the following to improve the efficiency and accuracy of the algorithm:

(1) Early stopping method: monitor the performance of the validation set during the training process, and stop the training early when the performance no longer improves [14].

(2) Stratified K-fold cross-validation: Ensure that the distribution of categories in each subset is consistent with the original dataset through stratified sampling, to improve the reliability of model evaluation, here we choose 4-fold stratified cross-validation [15].

(3) Grid search tuning: Grid search tuning is a systematic traversal of hyper-parameter combinations, which is suitable for the case of small parameter space, and can find the best parameter combinations. We optimize the parameters such as depth, number and learning rate of the decision tree to find the optimal parameters [16].

By using these 3 tuning methods, we can better improve the performance and generalization of the model to enhance the performance of the model on this dataset and obtain more accurate prediction results. By the above method, we obtained the optimal parameters of the model as shown in the Table.4 and Table.5.

Table 4. Optimal parameter combinations for the base model (partial)

Model	Random Forest	XGBoost
Optimal Parameters 1	Learning rate=0.1	Learning rate=0.05
Optimal Parameters 2	Depth =10	Depth =10
Optimal Parameters 3	Number of decision trees=200	Number of decision trees=200

Table 5. Base model at optimal parameters for each type of indicator

Indicator	Random Forest	Linear Regression
MAE	1.233251	1.188022
MSE	5.557710	4.557710

Then, we visualise the two base model training processes as shown in Figure 1.

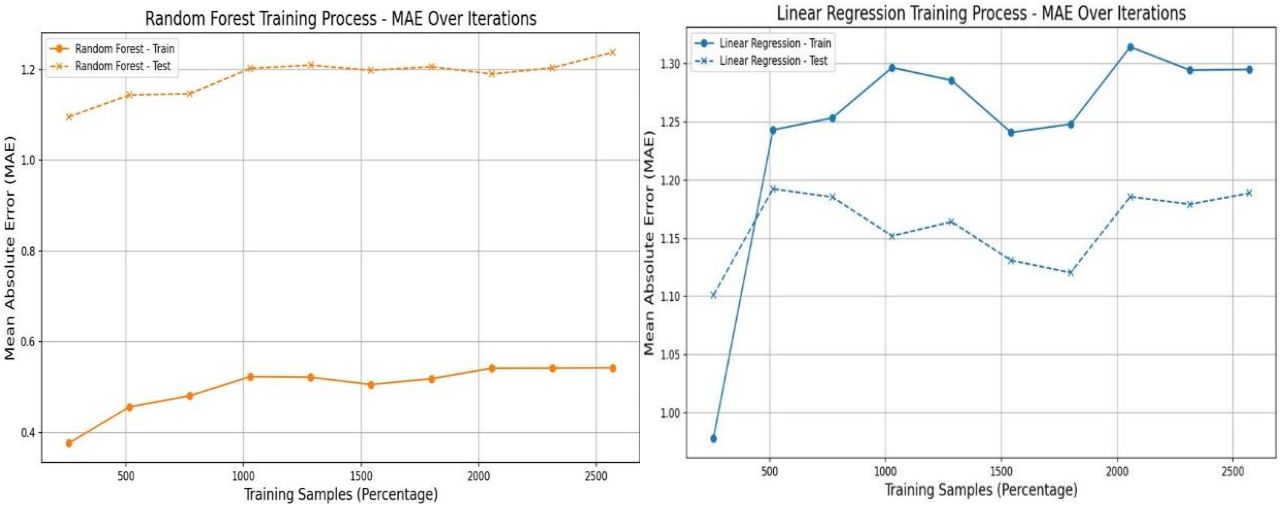

Figure 1. Iterative diagram of the base model

3.4. Stacking Integration Model

Considering the limited size of the medium dataset, the tree model is prone to overfitting problem, we adopt the stacked integration architecture. This architecture not only enhances the robustness of the models, but also effectively avoids the risk of overfitting. When integrating multiple base models, we choose Linear as the meta-model to fuse the prediction results of each base model. Specifically, we use the outputs of the five trained and optimized base models as inputs to Linear to train the model for efficient model integration. Figure 2 shows the principle of the model structure:

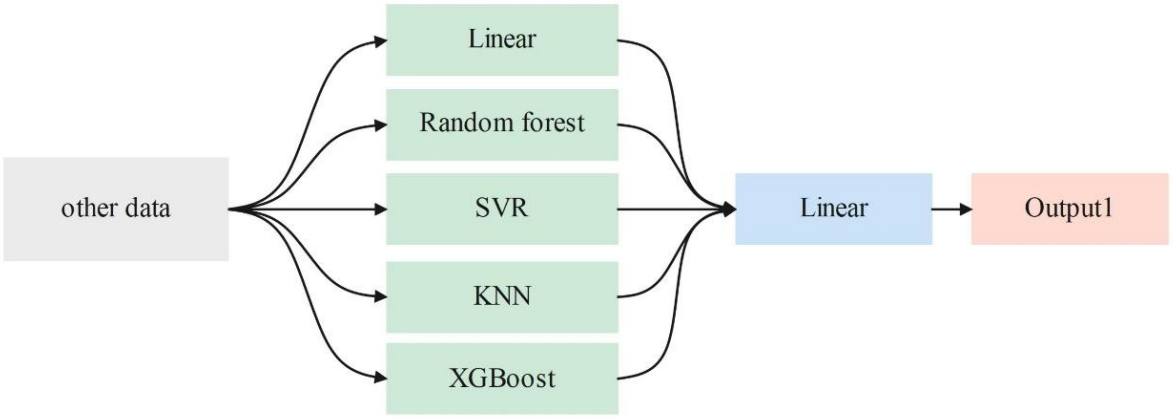

Figure 2. Schematic diagram of Stacking Integration Model

4. Time series modelling in Informer

Although directly using the time series data of the Olympic Games historical medal table to predict the future medal table ignores key information such as athletes' performance, programme competitiveness, external factors, etc., it is difficult to capture structural changes and cannot reveal the intrinsic drivers of changes in the number of medals, which leads to inaccurate prediction results. However, these time-series data can provide an auxiliary role for the model, which makes the model better in terms of refinement. At the same time, we consider that the variability of the number of Olympic medals may include cyclical changes in the athletes' sports status, etc., which has a complex seasonal nature, so we consider using the Informer model as a time-series prediction model [17].

Informer is an efficient and powerful time series prediction model, especially suitable for processing long series data. With ProbSparse self-attention mechanism and generative decoder, Informer significantly outperforms traditional Transformer models in terms of computational efficiency and prediction performance [18]. The workflow is shown in Figure 3.

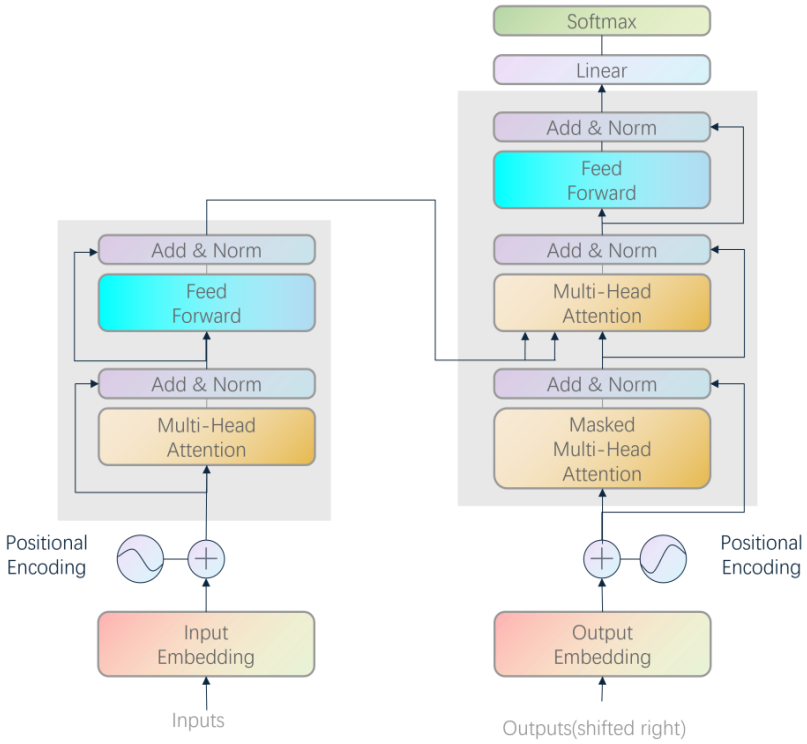

Figure 3. Informer model overview

5. Mixed effects model

Although the two prediction models are different in their underlying principles, both are able to show excellent performance when targeting specific problems. Therefore, considering this, we decided to hybridise the two models to obtain a higher performance hybrid predictive model. By complementing the strengths of both the models through the hybrid model, a better performance is obtained as compared to a single model. Figure 4 shows a schematic of the structure of our model:

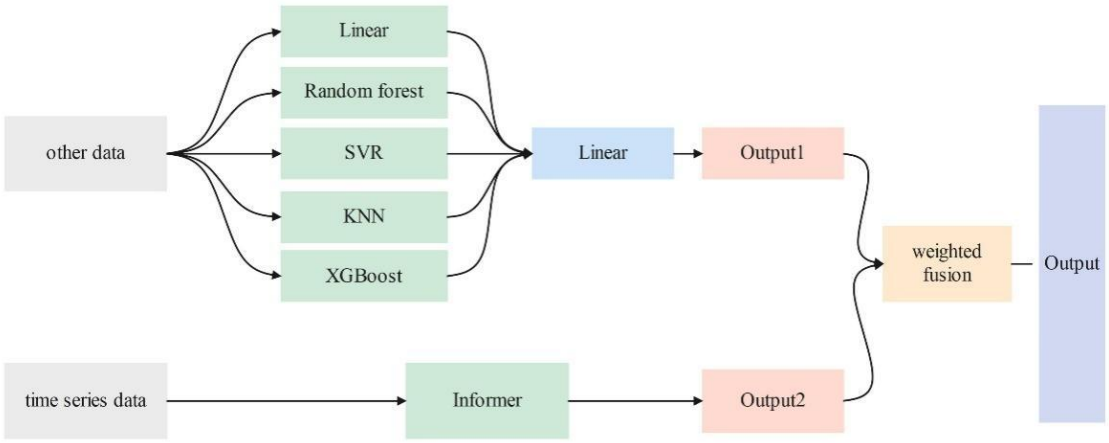

Figure 4. Mixed effects model overview

We then trained the hybrid model and calculated the values of MSE and MAE and found that it outperformed the linear regression and decision tree models alone on the dataset, while avoiding overfitting of the data and obtaining better predictions. The performance of the hybrid model is shown in Table.6.

Table 6. Mixed Model Performance

Indicator	MAE	MSE
Mixed effects model	1.24769362	11.41196783

6. Results

6.1. Calculating uncertainty measures

We calculate the standard deviation from the prediction results of each base learner, and this result reflects the confidence interval of the predicted values [19]. Meanwhile, when calculating the confidence intervals, we chose a confidence level of 95%, and finally, the visualization intuitively shows the effect and uncertainty of the prediction. The calculation is obtained as shown in Figure 5:

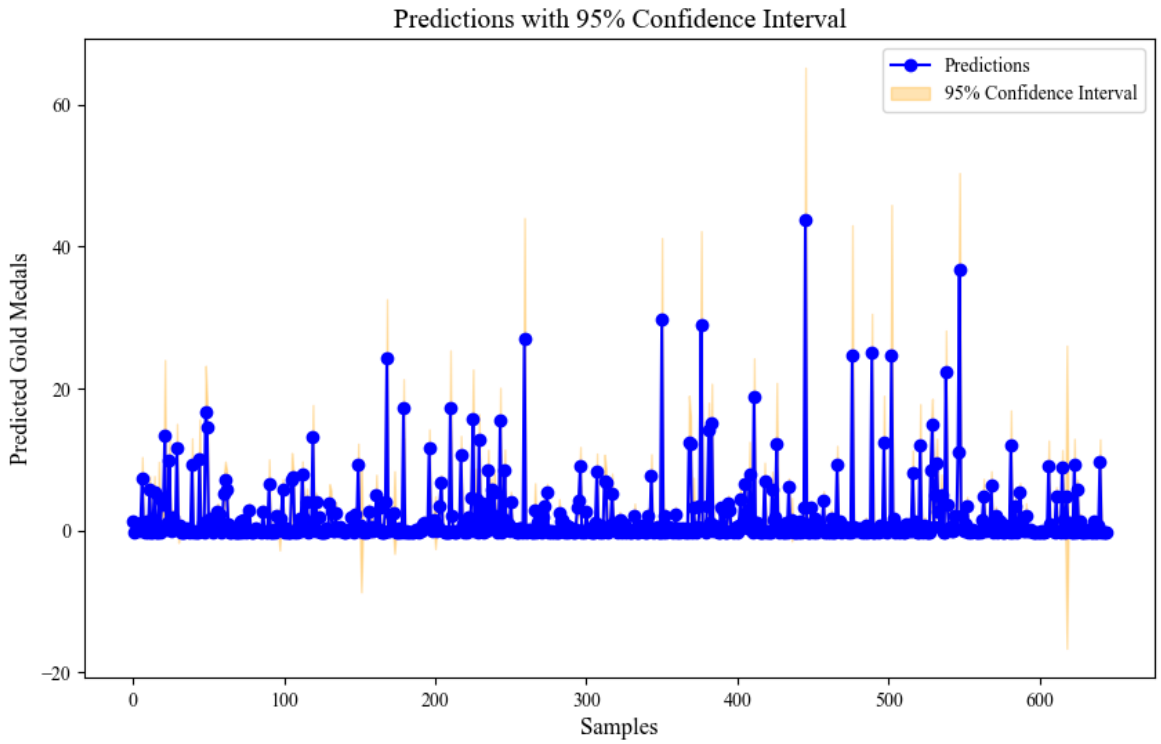

Figure 5. Predicted values with 95 per cent confidence intervals

Figure 5 shows that the uncertainty of the model's prediction for some samples is large, i.e. the standard deviation is too large, which also reveals the model's missing accuracy when dealing with some data. The reason for this could be white noise or problems with the structure of the data. However, overall, the model shows good performance in predicting most of the samples.

6.2. Analysis of experimental results

In predicting the medal table for the 2028 Olympic Games in Los Angeles, we assumed that there were no major changes in participants compared to the 2024 Games and analysed the new and deleted events for the 2028 Games and took this change into account in the model. In addition, we added points for the United States since it is hosting in 2028. Finally, we input these data into the constructed prediction model and ended up with the results shown in Table.7.

Table 7. Olympic Medal Projections

NOC	Gold medals	Total medals
United States	49	148
China	35	85
Japan	18	43
Australia	18	50
France	19	61
Netherlands	17	39
Great Britain	15	67
South Korea	10	30

By observing the table above, we can see that the United States, as the host country in 2028, has seen a substantial increase in the number of medals compared to 2024. In comparison, countries such as China and Japan may see a small decline in their results. The reason for this may lie in the cancellation of sports such as weightlifting and boxing at the 2028 Los Angeles Olympics, and the disappearance of these sports may lead to a drop in China's gold medal count. In conclusion, the 2028 Olympic Games in Los Angeles are full of variables and different countries will be affected by various factors at that time.

7. Conclusions

This study proposes a comprehensive Olympic medal prediction framework using advanced data science techniques. A novel K-means++ clustering method classifies countries into sports powerhouses and emerging nations, improving model adaptability and accuracy. A hybrid approach combines regression models (e.g., Random Forest, XGBoost) and time series models (e.g., Informer) via Stacking integration, creating a robust mixed-effects model. The framework calculates 95% confidence intervals and incorporates forecast bias analysis to address uncertainty and validate stability. The findings provide valuable insights for policymakers and sports analysts, advancing sports analytics and laying the groundwork for future research. Future work could integrate socio-economic and athlete performance data to enhance accuracy. This study highlights the transformative potential of data science in sports analytics and beyond.

References

[1] Ikotun A M, Ezugwu A E, Abualigah L, et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data[J]. Information Sciences, 2023, 622: 178-210.

[2] Koopialipoor M, Asteris P G, Mohammed A S, et al. Introducing stacking machine learning approaches for the prediction of rock deformation[J]. Transportation Geotechnics, 2022, 34: 100756.

[3] Sun Z, Wang G, Li P, et al. An improved random forest based on the classification accuracy and correlation measurement of decision trees[J]. Expert Systems with Applications, 2024, 237: 121549.

[4] Zhang J, Ma X, Zhang J, et al. Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model[J]. Journal of environmental management, 2023, 332: 117357.

[5] Zhu Q, Han J, Chai K, et al. Time series analysis based on informer algorithms: A survey[J]. Symmetry, 2023, 15(4): 951.

[6] Olympics.com, "Paris 2024 Medals," Olympics.com, 2024. [Online]. Available: https://olympics.com/en/paris-2024/medals. [Accessed: Oct. 10, 2023].

[7] Olympics.com, "Lang Ping Biography," Olympics.com, 2024. [Online]. Available: https://olympics.com/en/athletes/ping-lang. [Accessed: Oct. 10, 2023].

[8] USA Gymnastics, "Bela and Martha Karolyi Coaching Team," USA Gymnastics Hall of Fame, 2024. [Online]. Available: https://usagym.org/halloffame/inductee/coaching-team-bela-martha-karolyi/. [Accessed: Oct. 10, 2023].

[9] Wood S N, Augustin N H. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling[J]. Ecological modelling, 2002, 157(2-3): 157-177.

[10] Montgomery D C, Peck E A, Vining G G. Introduction to linear regression analysis[M]. John Wiley & Sons, 2021.

[11] Ali P, Younas A. Understanding and interpreting regression analysis[J]. Evidence-Based Nursing, 2021, 24(4): 116-118.

[12] Sun Y, Ding S, Zhang Z, et al. An improved grid search algorithm to optimize SVR for prediction[J]. Soft Computing, 2021, 25: 5633-5644.

[13] Bansal M, Goyal A, Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning[J]. Decision Analytics Journal, 2022, 3: 100071.

[14] Bai Y, Yang E, Han B, et al. Understanding and improving early stopping for learning with noisy labels[J]. Advances in Neural Information Processing Systems, 2021, 34: 24392-24403.

[15] Mahesh T R, Geman O, Margala M, et al. The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification[J]. Healthcare Analytics, 2023, 4: 100247.

[16] Belete D M, Huchaiah M D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results[J]. International Journal of Computers and Applications, 2022, 44(9): 875-886.

[17] Zhou H, Zhang S, Peng J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(12): 11106-11115.

[18] Zhang J, Li X, Tian J, et al. An integrated multi-head dual sparse self-attention network for remaining useful life prediction[J]. Reliability Engineering & System Safety, 2023, 233: 109096.

[19] Smithson M. Confidence intervals[M]. Sage, 2003.