A Medal Prediction Model Based on XGBoost-Logistic Regression and Quantification of the "Great Coach" Effect

Yisheng Gao *, #, Yixiao Mou #, Shengdi Xu #

School of Management, Shandong University of Technology, Zibo, China

* Corresponding Author Email: 13210839959@163.com

#These authors contributed equally.

Abstract. The paper selected sports that have consistently appeared in the Olympic Games since 2000, created a feature system, considering factors such as historical data, the elite athletes, gender ratios, and hosting country in the games. By comparing the performance of different models in this input feature the XGBoost performed exceptionally well in forecasting with its highest R^2 value which indicates that it is good at explaining the fluctuations in the data and the lowest MSE value which reflect that it has less variance in its predicted values. Furthermore, paper aimed to quantify the contribution of "great coach" on the number of medals won. Using the coaching information, the paper reconstructe the feature system and applied Ridge Regression to assess the impact of coaches. The effect of Coach Lang Ping on the Chinese and American women's volleyball teams was quantified as 6.432 and 6.913, respectively. Similarly, Coach Béla Károlyi contribution to Romanian and American gymnastics medals was 4.6123 and 4.3333, respectively. Then employed the CUSUM method to identify key breakpoints in the medal sequence and identified 3 countries (Italy and United States) that could benefit from employing "great coaches." The results highlighted the host country and elite athletes as particularly influential factors, offering valuable insights for national Olympic committees in their decision-making processes.

Keywords: XGBoost, Medal Prediction, Ridge Regression, CUSUM, "great coach".

1. Introduction

Olympic medal list prediction is an important research direction in the field of sports science and data analysis. In recent years, scholars have gradually shifted from traditional statistical methods to the application of complex algorithms (e.g., XGBoost, Random Forest), and tried to combine multi-dimensional features (e.g., athlete level, host country effect) to improve prediction accuracy. Meanwhile, the effect of "great coach" on sports performance has gradually become a hot research topic, but how to quantify its contribution is still controversial. The core of the study focuses on analyzing the application of data-driven models and the assessment method of coaching effect, which provides a theoretical framework for subsequent studies.

In terms of forecasting methods, traditional models are mostly based on linear regression or time series analysis. For example, Scelles et al [1]. predicted the medal distribution of the Summer Olympic Games by constructing a linear relationship between national sports inputs and medal outputs, but their model had limited ability to capture nonlinear features. With the rise of machine learning technology, integrated learning models with strong interpretability have gradually become a research hotspot. Shi et al [2]. proposed a gradient boosted tree (XGBoost) prediction model based on an interpretable machine learning framework, combining features such as historical medal data, athletes' level, and host advantage, which significantly improves prediction accuracy and stability, and provides methodological references for the subsequent research. The model provides a methodological reference for the subsequent research. In the analysis of influencing factors, the effect of "great coaches" has gradually attracted attention, and Kuang et al [3]. pointed out through case study that the psychological capital and leadership of coaches have a significant impact on team athletic performance. For example, during her coaching of the Chinese and American women's volleyball teams, Lang Ping significantly improved the team's performance through tactical innovation and team cohesion construction.

Therefore, this paper focuses on the research of Olympic medal prediction model and “great coach” effect. The application of machine learning methods such as XGBoost significantly improves the prediction accuracy, while the synergistic framework of Ridge Regression-CUSUM provides a new paradigm for quantifying the coaching effect, which supports the research of Olympic medal prediction and “great coach” effect. The Ridge Regression-CUSUM synergistic framework provides a new paradigm for quantifying the coaching effect, which supports the prediction of Olympic medals and the study of the “great coach” effect.

2. Materials and methods

2.1. Data preprocessing

All data about the Olympic medals as well as athletes and coaches are sourced from the website(https://www.contest.comap.com/undergraduate/contests/mcm/login.php). From 1932 to 2024, there are some sports that have been removed over time, such as Art Competitions, Baseball, etc., and others that have only existed for one term, such as Breakdancing, Rock Climbing, etc., so the research only counting the ones that have been included in the Olympics through the 2000, and that will continue to exist for a long period, such as the 100 meters of men's track and the 4x100 relay race. It should also be noted that volleyball, basketball, soccer and other team sports do not need to be dealt with when counting athletes, but when calculating the number of medals need to be viewed as a whole, there are some sports with missing values, such as men's table tennis singles and the team event of the 2024 Olympic Games have missing data, which was improved according to the actual situation.

For the study of the “great coach” effect, the research takes the coaching information of Lang Ping and Béla Károlyi as the examples to analyze the “great coach” effect.

In China, Lang Ping's first stint as head coach of the Chinese women's volleyball team was from 1995-2000, winning a silver medal at the 1996 Atlanta Olympics. With Lang Ping taking the head coaching position again from 2013-present, she led the Chinese women's volleyball team back to prominence and won the gold medal at the 2016 Rio Olympics. In the United States, Lang Ping was the head coach of the U.S. women's volleyball team from 2005-2008. The U.S. women's volleyball team has achieved remarkable successes, including winning the silver medal at the 2004 Beijing Olympics.

In Romania, Béla Karolyi coached the Romanian women's gymnastics team in 1976, and after Béla's tenure, the Romanian women's gymnastics team made significant improvements and won the 1976 Montreal Olympic gold medal. Subsequently, Bella Karolyi joined the U.S. women's gymnastics team as head coach in 1981, after serving as the U.S. women's gymnastics team's performance has made great progress, and won the 1996 Atlanta Olympic Games gold medal.

3. XGBoost-based medal prediction models

3.1. Feature Selecting

In the process of establishing the Characterization system, first of all, apart from considering the number of athletes, it also needs to consider the level of participating athletes, this research regarded the athletes who have won medals in two consecutive Olympic Games as elite athletes, and the gender ratio of participating athletes is also used as one of the features[4-5], and whether or not the country is the host country has an important impact on the prediction of the number of medals, which is used as a 0/1 variable.

Taking the gold medals earned by China in swimming in 2012 as an example, our features are set up as shown in the Table 1.

Table 1. The features about the China's gold medals in swimming in 2012

characteristic	label	example
0	Host Country	0
1	Total Medals in 4 years ago	3
2	Total Medals in 8 years ago	2
3	Total Medals in 12 years ago	0
4	Participants 4 years ago	42
5	Participants 8 years ago	28
6	Participants 12 years ago	21
7	Elite atheletes	1
8	Male won Medals 4 years ago	0
9	Male won Medals 8 years ago	0
10	Male won Medals 12 years ago	0
11	Number of Gold 4 years ago	1
12	Number of Gold 8 years ago	1
13	Number of Gold 12 years ago	0
14	Number of Silver 4 years ago	4
15	Number of Silver 8 years ago	1
16	Number of Silver 12 years ago	0
17	Number of Bronze 4 years ago	0
18	Number of Bronze 8 years ago	0
19	Number of Bronze 12 years ago	0

It should be noted that countries that have not won a medal for a long time should be excluded from the medal table prediction as some countries have not won a medal so far. After deriving the features used for prediction, the research standardized the metrics using the Z-score, which converts data points into standard deviation units, enabling comparisons between different data sets. After normalization, the Z-score indicates the degree of deviation of a data point from the mean; the greater the deviation, the greater the absolute value of the Z-score. If Z=0, the data point is equal to the overall mean; if Z>0, the data point is greater than the overall mean; and if Z<0, the data point is less than the overall mean. The Visualization is shown in Fig.1.

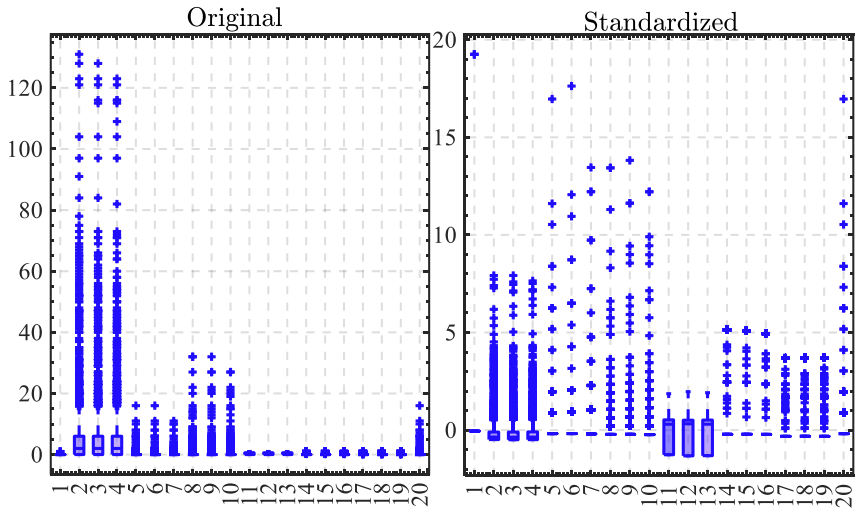

Figure 1. Original Feature Distribution and Normalized Feature Distribution

As can be seen from the images, the original feature distribution scatter shows relationships between features, although some relationships are relatively weak. The feature distributions after Z-score normalization, on the other hand, demonstrate more complex relationships, with some pairs of variables exhibiting non-linear relationships, suggesting that models are needed to capture the interactions between these variables. Meanwhile histograms on the diagonal provide the distribution

of each variable, helping to identify biases and outliers in the data. The pattern and density of the distribution of the scatterplot allows us to further assess the correlation between the variables.

3.2. XGBoost-based medal prediction models

XGBoost model is used for prediction, The XGBoost adopts the idea of integration, and it can be used to solve both classification and regression problems [6-8]. XGBoost can automatically select features and combine them through multiple tree models by gradient boosting tree algorithm to gradually improve the model performance, and XGBoost is highly efficient and flexible, which can effectively deal with large-scale data and complex features (e.g., the data of Olympic athletes in this paper). Take one of the countries for example, let ${N_t}^{Gold}$,${N_t}^{Silver}$,${N_t}^{Bronze}$ be the number of gold, silver and bronze medals won by the country in t Olympics, and let ${n_{i,t}}^{gold}$,${n_{i,t}}^{silver}$,${n_{i,t}}^{bronze}$represent the number of gold, silver, and bronze medals won by the country in the ith sport in the t Olympics respectively, where, taking ${n_{i,t}}^{gold}$ as an example, the calculation formula:

$$ {n_{i,t}}^{gold} = {\beta_{i0}}^{gold}\vartheta + {\beta_{i1}}^{gold}{a_{i,t-1}}^{all} + {\beta_{i2}}^{gold}{a_{i,t-1}}^{all} + \cdots + \beta_p a_{ip} + \epsilon_i \tag{1} $$

ϑ indicates whether it is the organizing party; ${a_{i,t-1}}^{all}$ represents the Total Medals in last Olympics; a_{ip} represents the feature p.

The formula for silver and bronze medals is similar to the above, where indicators 4, 5, 6, 13, 14, and 15 are individual feature variables specific to gold, silver, and bronze medals, and the rest are common indicator variables for all data.

The goal of the XGBoost model is to learn the model by minimizing the objective function, which consists of two parts: the loss function and the regularization term, and the training goal for each tree is to reduce the residuals (i.e., prediction error) of the model and reduce the loss by gradual adjustment. In this case, the objective function is as follows:

$$ \mathcal{L}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{2} $$

$\mathcal{L}(\theta)$ is the objective function, i.e., the prediction of the number of gold medals in a given country ${N_t}^{Gold}$; $L(y_i, \hat{y}_i)$ is the loss function, which is set to the root mean square error $MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ in this paper; and $\Omega(f_t)$is the regularization term, which prevents the model from being overfitted, and is commonly of the form:

$$ \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3} $$

T is the number of leaves in the tree; γ is the regularization parameter that controls the complexity of the tree (the larger γ, the shallower the tree); λ is the parameter of the L2 regularization term (the larger λ, the simpler the model); and w_j is the weight of the leaf nodes in the tree.

The research set the ratio of training set and test set as 7:3, firstly, the prediction results of XGBoost are compared with other models, including the three evaluation indexes of R^2, MAE and MSE, and the comparison results are shown in Fig.2.

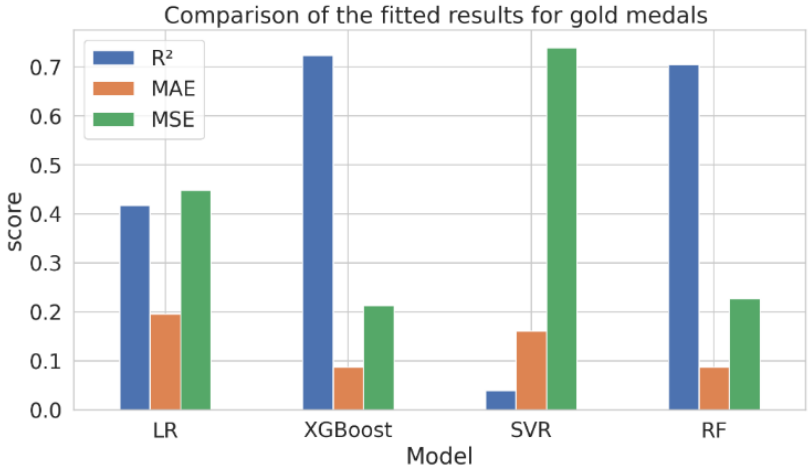

Figure 2. Model Comparison

From the figure it can be seen that the comparison of the different models for the prediction of gold medals where XGBoost has the highest R^2 value which is close to 0.7 indicating that the model is able to explain about 70% of the fluctuations in the data whereas the SVR model has the lowest R^2 value indicating that it is weak in explaining the fluctuations in the data. The MAE value of XGBoost is the lowest indicating that it has a smaller mean absolute error in its prediction value and also the XGBoost also has the lowest MSE value, indicating that it has less variance in its predicted values. In summary, XGBoost performs well in all three-evaluation metrics, which verifies the performance of our constructed model, and subsequent predictions will also use XGBoost as the model. Where the comparison of RMSE changes between the training set and the test set is shown in Fig.3.

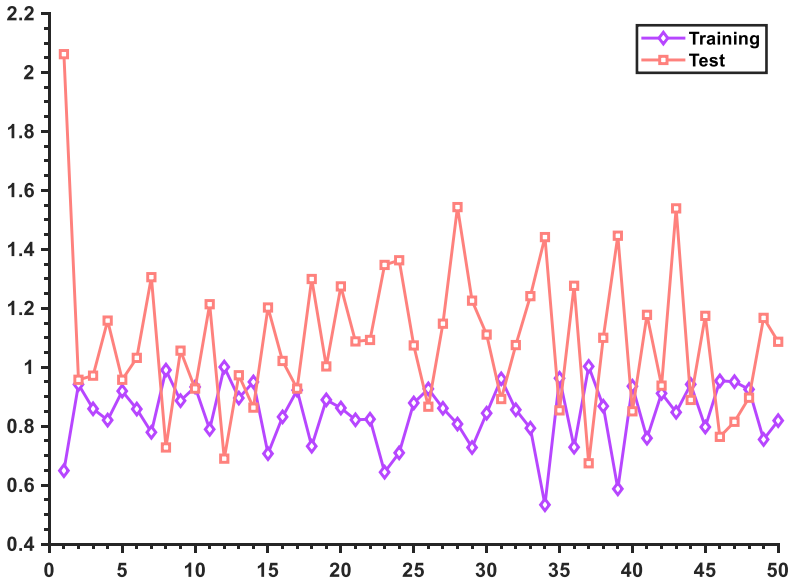

Figure 3. Comparison of RMSE

As can be seen from the figure, the RMSE stays around 1, which is a better case, and then comparing the actual situation of the training set and the test set as well as the residual effect, the effect is as shown in the Fig.4 and Fig.5.

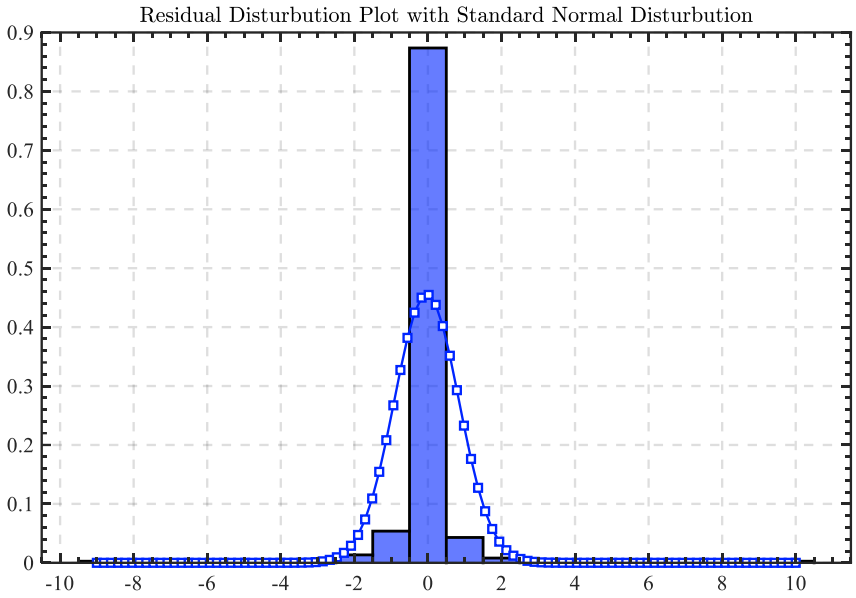

Figure 4. Training Set

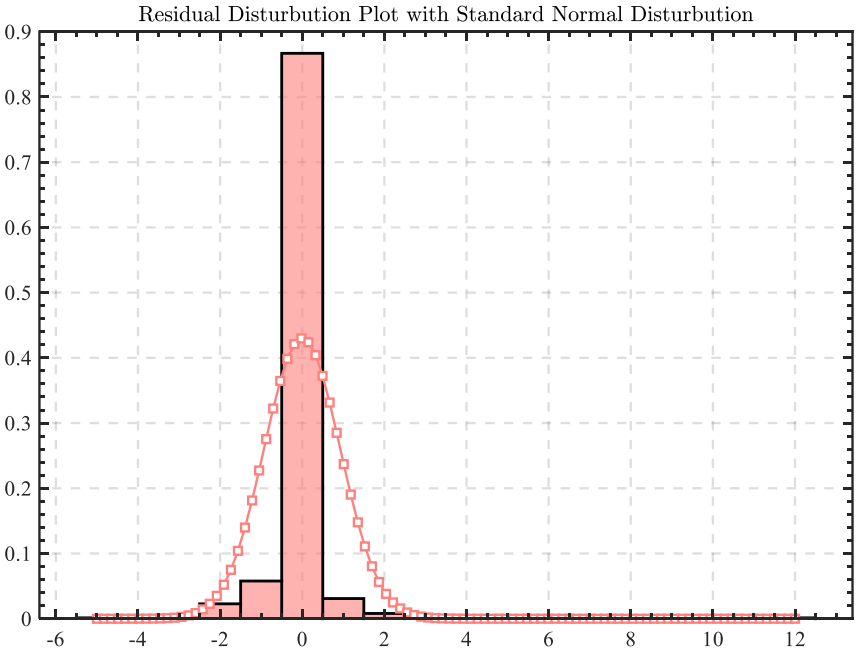

Figure 5. Test Set

As shown in the figure, the distribution of residuals in the training set shows an extremely sharp peak and decays quickly to close to 0, which indicates that most of the residuals are concentrated in a very small range and most of the residuals for most of the data points are very small, indicating that the model fits very accurately on the training data. The distribution of residuals in the test set is much looser and looser relative to the training set, with lower peaks and flatter distributions, indicating that the model's errors in the test set are more widely distributed than in the training set, which may be due to the fact that some of the countries that do not have medals have an impact on the model's predictions.

The resulting predictions for the 2028 Olympic Games in Los Angeles, CA are shown in the Table 2.

Table 2. 2028 Olympic Medal Table Projections

Country	Gold	Silver	Bronze	Total
United States	41	45	43	129
China	40	27	24	92
Japan	20	13	12	47
Australia	18	19	17	53
France	15	25	20	60
…	…	…	…	…
Slovakia	0	0	1	1
Zambia	0	0	1	1
Taiwan	0	1	2	2
Virgin Islands	1	0	1	2

4. The contribution of the “great coach” based on Ridge Regression

4.1. Quantitative modeling of the “great coach” effect based on Ridge Regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity [9]. It is a variant of linear regression that prevents overfitting by adding an L2 regular term to the loss function, a method that is suitable for cases where there is multicollinearity between the input variables.

Since women's volleyball is a high-intensity confrontation sport, the study assume that the longest career of women's volleyball players is 2 Olympic cycles, and the players in the team who can participate in the Olympic Games with the team for many times are regarded as elite players, and whether there is a “great coach” is used as a 0/1 input variable, based on the above, the study extracted the following feature.

Take the data of Chinese women's volleyball team in 2004 Athens Olympic Games as an example, the features are shown in the Table 3.

Table 3. The features about the Chinese women's volleyball team in 2004 Olympic

Feature	implication	example
1	Number of core athletes that year	1
2	Number of gold medals 4 years ago	0
3	Number of silver medals 4 years ago	0
4	Existence of “great coaches”	0
5	Number of bronze medals 4 years ago	0
6	Number of gold medals 8 years ago	0
7	Number of silver medals 8 years ago	9
8	Number of bronze medals 8 years ago	0
9	Change in number of entries from 8 years ago to 4 years ago	3
10	Change in number of gold medals 8 years ago to 4 years ago	0
11	Change in silver medals 8 years ago to 4 years ago	-9
12	Change in number of bronze medals 8 years ago to 4 years ago	0

The features are then normalized using Z-score, and the comparison of the features before and after normalization is shown in Fig.6.

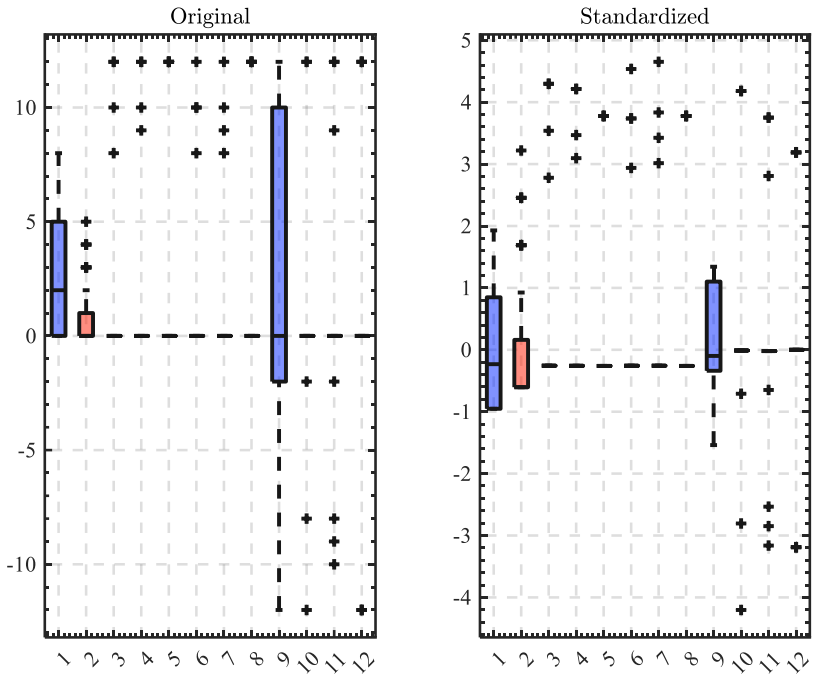

Figure 6. Women's Volleyball Feature Before and After Normalization

After standardizing the features, a linear regression model was constructed using the total number of medals as the output variable and the feature as the input variables, with the formula:

$$u_i = \lambda_0 + \lambda_1 V_{i1} + \lambda_2 V_{i2} + \cdots + \lambda_p V_{ip} + c_i \tag{4}$$

Where u_i is the observed value of the output variable, i.e., the number of medals of the country in that sport in the current Olympics; $V_{i1}, V_{i2,} \ldots, V_{ip}$ is the observed value of the input variable, i.e., the feature; $\lambda_0, \lambda_1, \ldots, \lambda_p$ are the parameters, where λ_0 is the parameter corresponding to the "Great Coach " is the parameter corresponding to the "great coach", which needs to be further calculated; c_i is the error term, which indicates the difference between the predicted value and the predicted value of the model.

The goal of Ridge regression is to minimize the loss function, which is given by:

$$L(\lambda) = \sum_{i-1}^{n} (u_i - \sum_{j=1}^{p} V_{ij}\lambda_j)^2 + \psi \sum_{j-1}^{p} \lambda_j^2 \tag{5}$$

Among them, ψ is the regularization parameter of the model, the idea of ridge regression is to minimize the residuals squared and SSR while minimizing the sum of the squares of the regression coefficients, thus preventing the regression coefficients from being too large, reducing the complexity of the model, and improving the model's generalization ability. The estimated formula for ridge regression can be obtained by deriving the loss function so that it is zero, and deriving the loss function with respect to λ to obtain:

$$\frac{\partial L(\lambda)}{\partial \lambda} = -2(U - V\lambda) + 2\psi\lambda \tag{6}$$

Making the derivative zero gives the estimation formula for ridge regression as

$$\beta_{\text{ridge}} = (V^T V + \psi I)^{-1} V^T U \tag{7}$$

Where I am the unit matrix.

Finally, the research chooses the data of Lang Ping on Chinese and American women's volleyball and Béla Károlyi on Romanian and American women's gymnastics to be fitted, and set the significance level of 0.05 as the criterion of whether the regression coefficients are significant or not, and the regression coefficients in Ridge regression usually have a p-value corresponding to the regression

coefficients to test whether the regression coefficients are significant or not. If the p-value of the regression coefficient is less than the significance level, then the regression coefficient is considered to be statistically significant, and the final fitting results of the effect feature of the “great coach” are shown in the Table 4.

Table 4. Feature of the "Great Coach" effect

National		Coef	Std_err	[0.025	0.975]
China Volleyball	const	3.51	1.231	0.692	8.641
	x1	6.43	3.52	-2.312	13.23
USA Volleyball	const	8.75	3.743	3.721	20.72
	x1	6.91	6.177	-7.486	23.91
Romania Gymnastics	const	7.888	2.181	2.955	12.82
	x1	4.612	5.116	-5.961	17.18
USA Gymnastics	const	12.12	2.543	6.373	17.87
	x1	4.333	4.869	-6.806	15.22

It can be observed that the coef values are all positive, with Lang Ping's influence effect on Chinese and American women's volleyball being 6.432 and 6.913, respectively; while Béla Károlyi's influence effect on Romanian and American women's gymnastics is 4.6123 and 4.3333, respectively.

4.2. CUSUM-based “Great Coach” Hiring Method

Based on the above analysis, the study give the basis for the introduction of master teachers, i.e., when a country's medal sequence has a very flat part (no medals), then the introduction of master teachers can be considered, and for the influence of master teachers, the study use here the average influence obtained from the previous modeling as a criterion, and use the CUSUM model to identify the flat sequence. The main advantage of the CUSUM method is that it is less restrictive on the distribution of the series, and can be used to analyze the variance of the series when the data do not conform to the normal distribution [10].

The CUSUM (Cumulative Sum) is a statistical method for detecting change points in a data series. It monitors the degree of deviation in the data by calculating the cumulative sum of the data series and triggers an event when the cumulative sum exceeds a preset threshold. The basic principle of CUSUM is to calculate the difference between each data point and the expected value and accumulate the difference values, and when the cumulative sum exceeds the set threshold, it is considered that the data has changed significantly. Generally, if the medal sequence is smooth for three consecutive years, then it is already time to consider hiring a master teacher to improve performance, so the study search for a flat sequence by finding the largest interval point and hiring a master teacher if the interval point is greater than 4. The study chooses the gymnastics teams of Italy and USA to analyze, and their medal sequence charts are shown in Fig.7 and Fig.8.

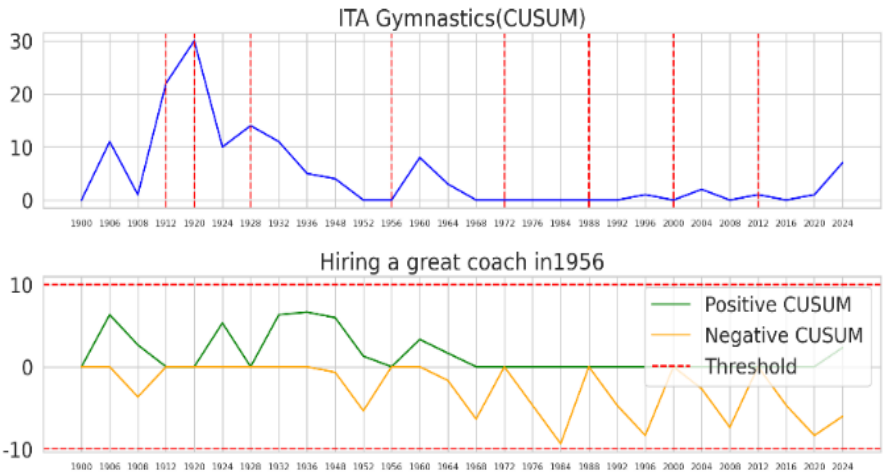

Figure 7. Italian Gymnastics CUSUM Chart

It can be observed from the images that Italy hired a "great coach" in 1956, and this event had a significant impact on the gymnastics performance, which fluctuated a lot and remained in a negative CUSUM state most of the time before 1956, and stabilized and remained in a whole CUSUM state most of the time after the hiring of the "great coach". After the hiring of the "great coach", the performance of gymnastics gradually stabilized and remained in the whole CUSUM state most of the time.

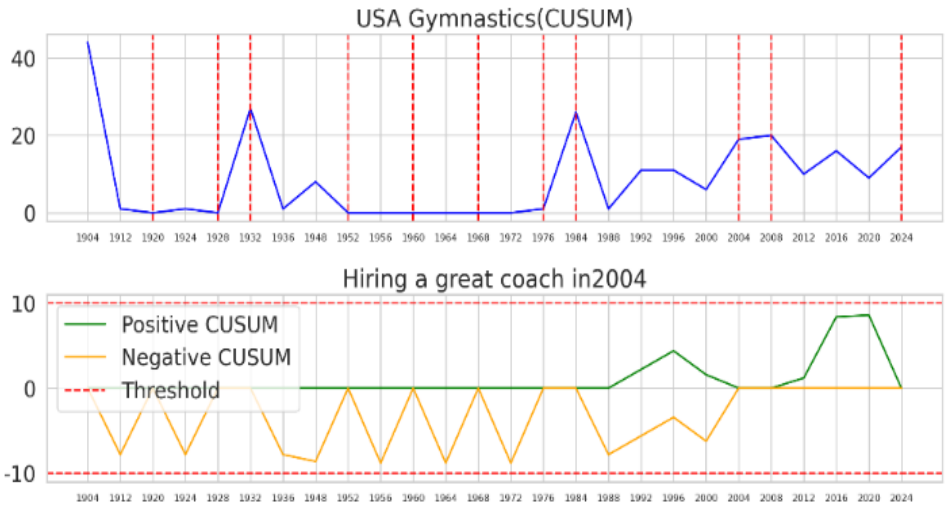

Figure 8. USA Gymnastics CUSUM Chart

Through the image, it can be observed that the United States hired a "great coach" in 2004, before 2004, the two CUSM lines basically fluctuate around zero, which indicates that the U.S. Gymnastics did not achieve much during this period of time, but after 2004, the CUSUM began to rise and reached a peak in 2008, and then fell back, but still remained at a high level, which indicates that the introduction of the "great coach" has brought a long-term positive impact on the U.S. Volleyball team. After 2004, CUSUM began to rise, and reached a peak in 2008, then fell back, but still maintained at a high level, the surface of the introduction of "great coach" for the U.S. volleyball team to bring a long-term positive impact.

5. Conclusion

This paper centers on the prediction of the medal table of the 2028 Los Angeles Olympic Games and the influence of the "great coach" effect. Through model comparison, the study chose XGBoost to predict the medal rank of the 2028 Olympic Games, and the results showed that XGBoost performed well in predicting the number of medals, and demonstrated a high degree of accuracy and stability. through the analysis of the prediction results, the study predicted the medal rank of the 2028 Los Angeles Olympic Games which provided a valuable reference for the Olympic Committee.

For the "Great Coach" effect, the study takes Lang Ping, who coached the Central American women's volleyball team and Béla Károlyi who coached the Romanian and U.S. women's gymnastics teams' analysis as the example for analysis. The study uses Ridge Regression to quantify the effect, and the CUSUM model is used to determine which countries can bring positive benefits to their programs by hiring "Great Coach". Subsequently, the data of the Olympic Games were analyzed from the geographical perspective, the time perspective and the impact of "Great Coach", and provided certain references for each country and the Olympic Committee.

In summary, this paper proposes a comprehensive model to predict the medal table of the 2028 Los Angeles Olympic Games, analyzes in depth the factors affecting the number of medals and the effect of "great coaches", and analyzes the Olympic Games data visually from three perspectives: geographical perspective, time perspective, and the effect of "great coaches". The data is visualized and analyzed, providing scientific decision-making support for countries and Olympic committees. In the future, more models can be combined to make more accurate predictions.

References

[1] Scelles N, Andreff W, Bonnal L, Andreff M, & Favard Forecasting national medal totals at the Summer Olympic Games reconsidered [J]. Social Science Quarterly, (2020). 101 (2). 697 – 711.

[2] SHI Huimin, ZHANG Dongying, & ZHANG Yonghui. Can Olympic Medals Be Predicted?Based on the Interpretable Machine Learning Perspective [J]. Journal of Shanghai University of Sport, (2024). (04) 26 - 36.

[3] KUANG Hong-da, XU Li-ping, & Ll Lin-ying. China Sport Science and Technology [J]. China Sport Science and Technology, (2018). (01) 57 - 63+128.

[4] SONG Zhigang, JAO Fangqian & WU Xiangwei. Study on the Characteristics of China's Awards in the Tokyo Olympics andthe lmplications for the Preparation of the New Olympic Cycle[J]. Journal of Beijing Sport University (2024). (07), 146 - 156.

[5] TIAN Hui, HE Yiman, WANG Min, Ll Juan, YU Peiyang, Ql Shunhong, & TlAN Ye. Medal Forecasts and Competition Strategies for Chinese Athletes in theBeijing 2022 Winter Olympic Games-Based on Olympic Home Advantage [J]. China Sport Science. (2021). (02). 3 - 13+22.

[6] SHUI Yingyi, ZHANG Qi, LI Gen, ZHANG Shihao & Wu Shang. A Review of Research on Social Network Influence Prediction Based on Multi-Class Features [J]. Frontiers of Data & Computing (2025) (01), 2 - 18.

[7] Zhang Y, & Chen L.A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-Sampling Method [J]. Theoretical Economics Letters, (2021).11, 258 - 267.

[8] Waberi A.D, Mwangi R.W. and Rimiru R.M. Advancing Type II Diabetes Predictions with a Hybrid LSTM-XGBoost Approach [J]. Journal of Data Analysis and Information Processing, (2024)12, 163 - 188.

[9] Khalaf, G Improving the Ordinary Least Squares Estimator by Ridge Regression [J]. Open Access Library Journal, (2022).9: e8738.

[10] Degang Li. Baidu Index Change Point Analysis Based on CUSUM Method [J]. Statistics and Applications, 2024, 13 (2), 521 - 527.