Predicting Olympic Medal Counts by Bootstrap-Enhanced Gradient Boosting Machine with Log-Transformed K-Means Clustering

YiRan Zhang *

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, 150081

* Corresponding Author Email: zzzzzyr1527@outlook.com

Abstract. Olympic medal counts serve as a critical indicator of a nation's comprehensive sports strength. Accurate predictions offer distinctive perspectives for strategic resource allocation. To handle the strongly skewed distribution featuring long tails in medal data and capture complex nonlinear relationships among variables, this study proposes a hybrid methodology combining log-transformed K-means clustering and Bootstrap-Enhanced Gradient Boosting Machine (BEGBM). Countries are stratified into five strength tiers (e.g., elite to bottom) via K-means clustering on log-transformed medal features, reducing skewness bias. A robust BEGBM model trained on bootstrap samples (B=150) incorporates fluid sports variables and clustered tiers, with missing values imputed by random forests. The work presented here is the first to integrate tiered and mindful of uncertainty GBM into medal prediction. The model achieves RMSE=1.16 (gold)/1.91 (total), MAE=0.33/0.81, and R^2=0.50/0.61. It further provides 95% CIs for the 2028 Los Angeles Olympics (e.g., USA: [35.8, 42.7] gold medals) and identifies nations with significantly changing medal potential. This data-based method offers sports authorities with a dependable tool for optimizing resource investment.

Keywords: Gradient Boosting Machine, Bootstrap Sampling Algorithm, Log-transformed k-means Clustering, Prediction Model.

1. Introduction

During competitions, people often concentrate on athletes' showings while also considering the outcomes. Additionally, medals with significant historical importance for the country are highly valued [1]. Medals counts may relate to factors such as the athletic strength of different countries and various sports events. In previous studies, Tchamkerten et al. applied logistic regression model to examine the relationship between swimmers' medal wins and factors like nationality and event type [2]. Csurilla et al. employed survival analysis to study the determinants of success sustainability over seven consecutive Summer Olympics from 1996 to 2021 [3]. Sekitani et al. used a DEA model with restricted multipliers to predict medal counts [4].

While existing methods offer partial solutions, critical gaps persist: Logistic regression cannot simulate nonlinear interactions such as those between nationality and events. Tripepi et al. pointed out that linear regression analysis requires the dependent variable to be continuous [5]. DEA models display limited adaptability to dynamic, multi-factor sport systems [6]. Additionally, survival analysis did not fully incorporate real-time changing data of individual competitive states.

Based on this, this study pioneers a Bootstrap-Enhanced Gradient Boosting Machine (BEGBM) framework. The data were transformed through logarithmic transformation in k-means clustering to clearly classify countries by different levels of national competitiveness tiers, and to quantify features such as the number of gold medals, total medals, event typologies, and the correlation between events and medal counts. To address the issues of capturing nonlinear variables and solving multi-dimensional interaction effects, this approach aims to achieve more accurate predictions of medal counts.

2. The basic fundamental of models

2.1. The structure of log-transformed k-means clustering

Log-transformed K-means clustering is an improved algorithm combining traditional K-means clustering, mainly used to solve the problem of inaccurate clustering results caused by skewed data distribution. Its core process includes two stages: logarithmic transformation and K-means clustering.

The implementation of log-transformed K-means clustering involves three key steps:

(1) The medal records of each country-session are transformed into a data point through log-transformed for cluster analysis. The transformation formula is:

$$x' = \log(x + 1) \tag{1}$$

Where x is the Olympic medal counts, and the $x + 1$ operation avoids the problem of undefined logarithm for 0 values. This transformation can compress the range of extreme values and reduce their impact on clustering results.

(2) Determine the number of clusters k [7], And randomly select k data points from the transformed dataset as the initial clustering centers $c_1, c_2, \ldots, c_k$.

(3) Calculate the Euclidean distance between each data point and each cluster center, and assign the data point to the nearest cluster:

$$d(x_i', c_j) = \sqrt{\sum_{m=1}^{n}(x_{i,m}' - c_{j,m})^2} \tag{2}$$

where $d(x_i', c_j)$ represents the Euclidean distance between the i-th feature x_i' after logarithmic transformation and the j-th cluster center c_j, $x_{i,m}'$ is the m-th eigenvalue of the i-th transformed data point, $c_{j,m}$ means The m-th eigenvalue of the j-th clustering center, and n represents feature dimension.

Then recalculate the center of each cluster, that is, the mean of all data points in that cluster. And repeat these steps until the cluster centers no longer undergo significant changes.

$$c_j = \frac{1}{n_j}\sum_{i \in A_j} x_i' \tag{3}$$

where A_j is the set of data points in the j-th cluster, n_j represents the number of data points in this cluster.

Calculate the sum of squared errors (SSE) corresponding to different k values to evaluate the clustering effect.

$$SSE = \sum_{j=1}^{k}\sum_{i \in A_j} d(x_i', c_j)^2 \tag{4}$$

As k increases, SSE will monotonically decrease, and the maximum value of k should not exceed the square root of the sample size. Take the value of k when the aggregation degree of each cluster no longer shows a significant improvement.

2.2. The structure of Bootstrap-Enhanced Gradient Boosting Machine

BEGBM is used to predict the number of medals and calculate confidence intervals to identify countries with changes in the number of medals. The structure is shown in Figure 1.

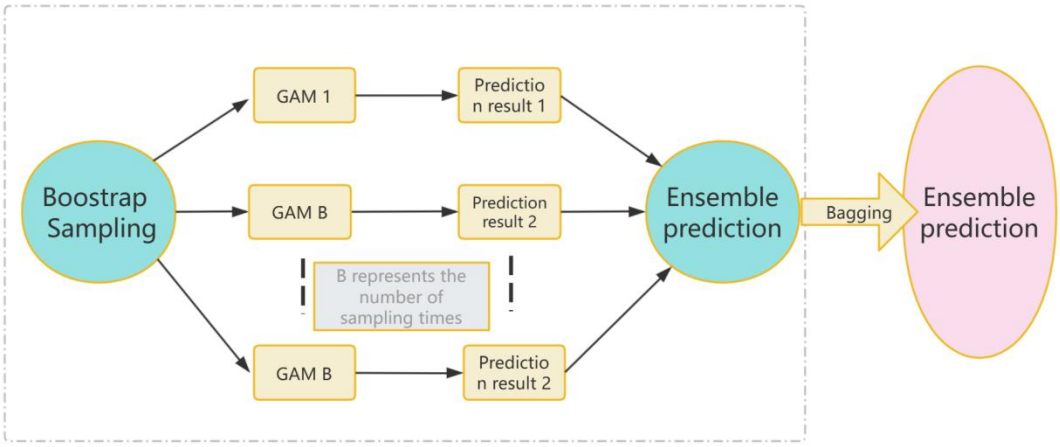

Figure 1. BEGBM structure

Medal counts are predicted nonlinearly based on national strength, project characteristics, historical performance, and other factors using the Gradient Boosting Machine (GBM). It is an iterative decision tree ensemble model that enhances performance by repeatedly correcting the prediction errors of previous models. By incorporating the Bootstrap Sampling Algorithm to perturb the training data [8], the model's robustness is improved, and the uncertainty of predictions is quantified.

Since there are some missing values in the data, the random forest algorithm is used for data imputation [9]. Specifically, samples with non-missing values serve as the training set. By extracting relevant feature values from the existing non-Na values and selecting an optimal feature (assuming there are d features) from the feature set of the current node, for each node of the base decision tree, a subset containing m features is randomly selected from the feature set of that node first. Then, an optimal feature is selected from this subset for partitioning to predict the possible values of NA values.

The reason for introducing a parameter m is to control the degree of randomness introduction: if $m = d$, then the construction of the decision tree is the same as that of the traditional decision tree. If $m = 1$, then a feature is randomly selected for division. Recommended values under normal circumstances is $m = \log_2 d$ balancing randomness and discriminative power. The prediction result of the missing value NA on the non-Na values of the sample is represented as an N-dimensional vector, and each element in the vector represents the prediction result in the corresponding class.

The model begins with a simple initial predictor, which is the mean of all samples. In each iteration, a new decision tree is trained, with its learning goal being the prediction residuals of the previous model. By continually correcting errors, the overall prediction error is gradually minimized. And the negative gradient of the loss function is used as the "pseudo-residual" to guide the learning direction of the new weak learner, ensuring that each iteration is optimized in the direction where the loss function decreases.

Initial model is:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{5}$$

where γ represents the predicted value of the initial model, y_i is the true medal counts of the i-th sample, and $\sum_{i=1}^{n} L(y_i, \gamma)$ is the sum of losses across all samples, n denotes the total number of samples in the training set.

For the prediction of Olympic medal counts, the mean square error (MSE) is adopted as the loss function:

$$L(y, \hat{y}) = (y - \hat{y})^2 \tag{6}$$

where y is the true medal counts, $\hat{y}$ is the model's predicted medal count, and the negative gradient under this loss function equals the true residual $r = y - \hat{y}$, This residual, representing the

gap between the true and predicted values, becomes the fitting target for the new decision tree, which is trained to minimize the sum of squared residuals.

To avoid overfitting, each new weak learner's contribution is weighted by a learning rate (η, typically 0.1 in this study), a value less than 1. This shrinkage mechanism ensures that each iteration corrects only a fraction of the error, allowing the model to converge gradually and generalize better to unseen data.

The final model after M iterations is expressed as the weighted sum of all weak learners:

$$F_M(x) = F_{M-1}(x) + \eta h_m(x) \tag{7}$$

where $F_M(x)$ is the final strong learner,M means the number of iterations, and $h_m(x)$ is the weak learner trained in the m-th iteration.

Bootstrap sampling strengthens the model by generating multiple subsets through repeated sampling with replacement from the original training set. A separate GBM model is trained on each subset, and the final prediction is derived from the average of these models' outputs. This ensemble approach not only reduces overfitting but also quantifies prediction uncertainty.

3. Results

The data are from https://www.olympics.com. Represents the medal counts of the countries participating in the summer Olympics from 1896 to 2024 (shown in Table 1 below).

Table 1. Total Medal Data Chart of Various Countries (Partial display)

Rank	NOC	Gold	Silver	Bronze	Total	Year
1	Unite-States	11	7	2	20	1896
2	Greece	10	18	19	47	1896
3	Germany	6	5	2	13	1896
4	France	5	4	2	11	1896
5	Great Britain	2	3	2	7	1896

3.1. The establishment of log-transformed k-means clustering

To address the skewness and extreme value issues in Olympic medal data, this study employed a log-transformed K-means clustering approach, with the detailed implementation process as follows:

First, rigorous data preprocessing was conducted to ensure data quality. eliminate outliers (data from countries whose results were invalidated for some reasons) and missing values (data from countries with incomplete participation data).

According to the distribution chart of gold medals and other medals of different countries in 2024 (shown in Figure 2 below), most countries have won a relatively small number of medals, while only a few countries have won a large number of medals. That is, the data shows a distinct skewed distribution, which may have an adverse impact on subsequent data analysis and model construction.

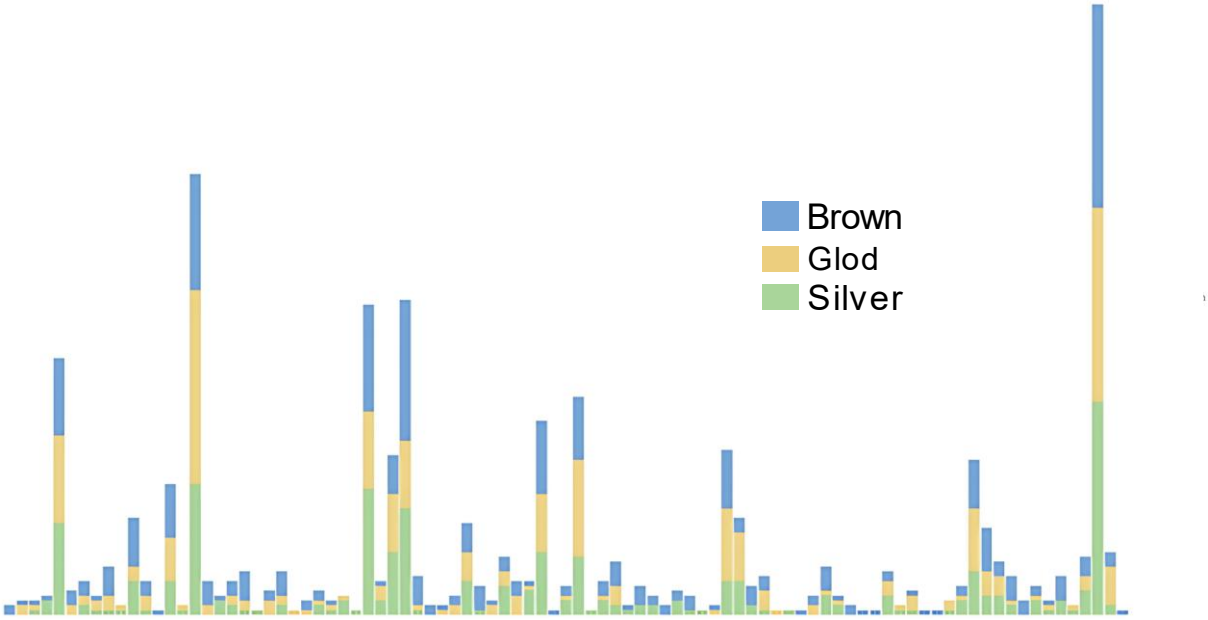

Figure 2. A chart showing the distribution of medals from different countries in 2024

Subsequently, log-transformed was applied to all non-negative feature variables to stabilize the data distribution and mitigate the undue influence of extreme values (Formula 1), the characteristic variables including gold medals (x_1), silver medals (x_2), bronze medals (x_3). yielding the transformed feature matrix $X' = [x_i']_{N\times 3}$. Each element in the matrix was defined as: $x_i' = (\log(x_{i1} + 1),\ \log(x_{i2} + 1),\ \log(x_{i3} + 1))$.

Then, the elbow method is applied to determine the optimal number of clusters, iterate over different k values and calculate the SSE values corresponding to different k values (Formula 4), and then use the ggplot package in R language to generate the elbow plot (Figure 3).

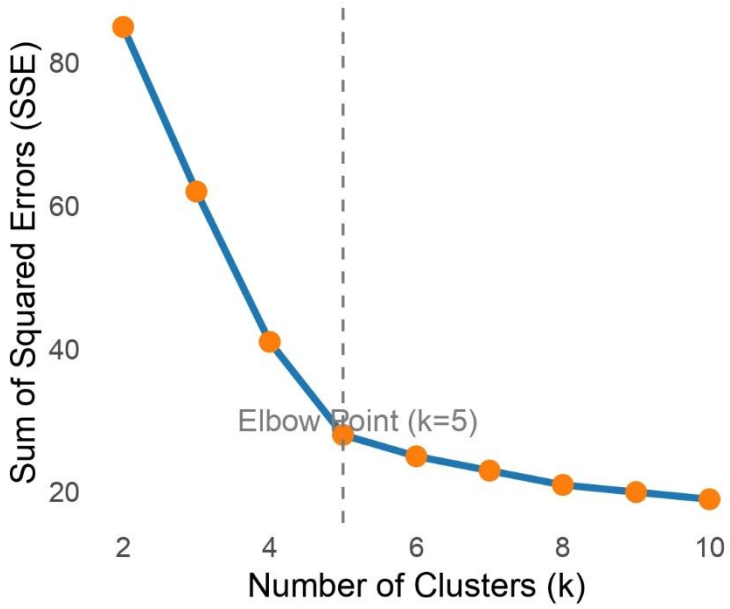

Figure 3. Elbow plot for determining optimal k value in log-transformed K-means clustering

The results show that when $k = 5$, the curve shows a distinct "elbow" [10]. Therefore, the optimal number of clusters is determined to be 5. This can also correspond to five different levels of national strength, such as bottom, upper-middle, medium, lower-middle, and elite (Figure 4).

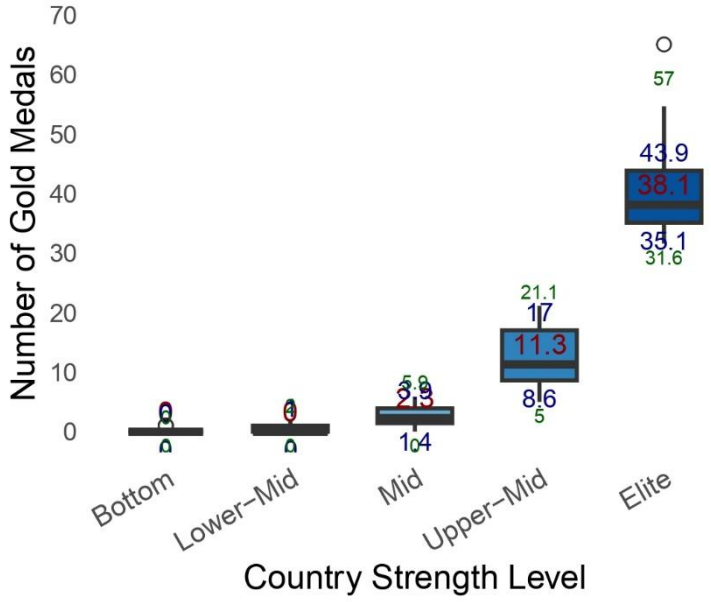

Figure 4. Distribution of Gold Medals Across Clusters

This box plot shows the distribution of the number of gold medals won by countries at different medal strength levels, it indicates that the countries at the bottom of the table, due to the scarcity of sports resources, have hardly won any gold medals, and the difference in the number of gold medals won by countries of different medal strength levels: the higher the level, the stronger the ability to win gold medals and the greater the number.

Randomly select 5 data points from the transformed data matrix X' as the initial centers $c_1, c_2, \ldots, c_5$. Calculate the Euclidean distance between each data point and the five centers according to (Formula 2), and assign the data points to the nearest cluster center. And calculate the new center for each cluster as the mean of its assigned data points using (formula 3). Manually set the threshold to $e = 10^{-4}$, repeat the above process until the maximum displacement of any cluster center between two consecutive iterations is less than the e.

3.2. The establishment of Bootstrap-Enhanced Gradient Boosting Machine

Firstly, converting the country code to the ISO 3166-1 alpha-3 standard encoding to ensure the consistency of the country identification, for special entities without standard codes (such as independent Olympic athletes) and disintegrated countries (such as Czechoslovakia), they are uniformly marked as "uncoded" and handled separately. For the missing features such as the number of historical medals and the number of project participations, the random forest algorithm is adopted for prediction and filling. Take the non-missing samples as the training set, take the missing features as the target variables, and take other relevant features (such as the number of participating events, athletes, and the number of participations, etc.) as the input.

In each iteration of GBM training, the function randomly selects m features for each decision tree according to the feature subset rule of formula 8. In the initial stage, the predictor variables of the model are Year and Cluster. Then during the subsequent iteration process, the model will automatically generate the association formula of “target variable - predictor variable” without the need for manual definition.

The optimal parameters are screened through multiple rounds of training and verification (Initially, a 5-fold cross-validation is adopted. When the features expand, it is adjusted to a 7-fold cross-validation.), Parameter tuning is based on the criteria of mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). Key parameters include tree depth (interaction. depth) and tree count (n. trees). The optimal parameters are selected, and the number of features is expanded to the year, sports event, number of athletes, and cluster.

During the training process, a bootstrap sampling strategy is introduced: multiple subsets are selectively drawn from the training set, and each subset is independently trained with the GBM model. The final prediction result is the average output of all subset models.

The optimal combination of hyperparameters is shown in the following table 2.

Table 2. The optimal combination of hyperparameters in the GBM model

Model Type	Predictors' Number	Cross-Validation Folds	Optimal Parameter Combination	Fixed Parameters
Gold Medals Prediction	4	7	n.trees=50, interaction.depth=2	shrinkage=0.1, n.minobsinnode=10
Total Medals Prediction	4	7	n.trees=150, interaction.depth = 3	shrinkage=0.1, n.minobsinnode=10

The RMSE of the BEGBM model for the number of gold medals is approximately 1.16 and the MAE is approximately 0.33, indicating that the predicted values are close to the true values, with small deviations and high stability. R^2 is approximately 0.50, capturing some of the core influencing factors. The RMSE of the BEGBM model for the total number of medals is approximately 1.91 (greater than that of the gold medal model, with slightly weaker accuracy and dispersion), and the MAE is about 0.81 (with a larger deviation), but the R^2 is approximately 0.61, which can explain 61% of the changes in the total number of medals and capture the influencing factors more fully. The model has been trained and used to predict the year 2028, At the same time, 95% of the prediction results were sampled and calculated using the Bootstrap Sampling method. The prediction results are shown in Table 3 below. The table shows the prediction values and 95% confidence intervals for the top five countries in terms of gold medals and total medals, showing the model's estimates and uncertainty about each country's medal performance.

Table 3. The prediction results for the number of Gold Medals and Total Medals in 2028

NOC	**Gold Medals**		**Total Medals**	
	Predicted	95% CI	Predicted	95% CI
USA	39.27	35.81 - 42.73	110.45	105.12-115.78
CHN	38.15	34.92 - 41.38	89.73	84.56 - 94.90
GBR	21.83	18.47 - 25.19	64.92	59.14 - 70.70
FRA	12.56	9.87 - 15.25	44.38	39.42 - 49.34
GER	10.71	8.24 - 13.18	39.65	35.17 - 44.13

4. Conclusions and outlooks

This study employs log-transformed K-means clustering and a BEGBM model to predict Olympic medal counts. After processing the medal data with skewed distribution through log-transformed, K-means clustering divides countries into different levels based on gold medals counts and the total number of medals, clearly presenting the global sports strength pattern and providing a basis for targeted analysis of countries at different levels. A robust BEGBM model with RMSEs of 1.16 (gold medal) and 1.91 (total medal) was established, it also provided a 95% confidence interval for the number of medals at the 2028 Los Angeles Olympics, enhancing the credibility of the prediction results. The model recognition identified the countries whose medal counts might rise or fall, providing a basis for the allocation of project resources.

However, if the data is small, there will be a problem of algorithm performance degradation, and the randomness is relatively strong. In the future, dynamic variables such as the age distribution of athletes, injury and illness rates, and sports policies can be incorporated to further enhance the prediction accuracy.

References

[1] Millet G P, Hosokawa Y, Sandbakk Ø, et al. Editorial: Tokyo 2020 Olympic and Paralympic games: Specificities, novelties and lessons learned. [J]. Frontiers in sports and active living, 2022, 4: 1026769.

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

[3] Csurilla G, Ferto I. How long does a medal win last? Survival analysis of the duration of Olympic success[J]. Applied Economics, 2022, 54(43): 5006-5020.

[4] Sekitani K, Zhao Y. Performance benchmarking of achievements in the Olympics: An application of Data Envelopment Analysis with restricted multipliers[J]. European Journal of Operational Research, 2021, 294(3): 1202-1212.

[5] Tripepi G, Jager K J, Dekker F W, et al Linear and logistic regression analysis[J]. Kidney International, 2008, 73(7): 806-810.

[6] Bai Y P, Chen Q Q. Spatio-temporal evolution and influencing factors of technological innovation efficiency in the software and information technology service industry of the Yangtze River Economic Belt[J]. Journal of Nanjing University of Posts and Telecommunications (Social Science Edition), 2023, 25(2): 56-67.

[7] 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.

[8] Jiang F, Yu X, Du J, et al. Ensemble learning based on approximate reducts and bootstrap sampling[J]. Information Sciences, 2021, 547: 797-813.

[9] Pantanowitz A, Marwala T. Missing Data Imputation Through the Use of the Random Forest Algorithm[C]//Yu W, Sanchez E N. Advances in Computational Intelligence. Berlin, Heidelberg: Springer, 2009: 53-62.

[10] Sholeh M, Aeni K. Perbandingan Evaluasi Metode Davies Bouldin, Elbow dan Silhouette pada Model Clustering dengan Menggunakan Algoritma K-Means [Performance Comparison of Davies-Bouldin, Elbow, and Silhouette Validation Techniques in K-means Clustering Models] [J]. STRING (Satuan Tulisan Riset dan Inovasi Teknologi), 2023, 8(1): 56-65.