Proceedings of Machine Learning Research vol 278:1–8, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

# Modeling and Analysis of Olympic Medal Table Based on Multiple Features

**HanBang Chen**
*School of Mechanical and Electrical Engineering, Guangdong University of Science and Technology, 523668, China*

**Delin Kong**
*School of Mechanical and Electrical Engineering, Guangdong University of Science and Technology, 523668, China*

**Biao Jin*** 53340450@QQ.COM
*School of Mechanical and Electrical Engineering, Guangdong University of Science and Technology, 523668, China*

**CaiLing Huang**
*School of Mechanical and Electrical Engineering, Guangdong University of Science and Technology, 523668, China*

**YiXiang Ke**
*School of Mechanical and Electrical Engineering, Guangdong University of Science and Technology, 523668, China*

**MingCheng Zhang**
*School of Mechanical and Electrical Engineering, Guangdong University of Science and Technology, 523668, China*

**GuanHua Li**
*School of Mechanical and Electrical Engineering, Guangdong University of Science and Technology, 523668, China*

**Chao Tan**
*School of Mechanical and Electrical Engineering, Guangdong University of Science and Technology, 523668, China*

*Corresponding author*

**Editors:** Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

## Abstract

In the first part, this study first used the winning records and medal data of Olympic competitions. Based on the relevant variation characteristics of medal counts, their impact was assessed by quantifying the fluctuation of medal counts under multiple characteristics. For medal counts, they were incorporated into a medal prediction model under time series through stacked integration. LR, LASSO, SVM, and Catboost were used as base leaners in the first layer ; RF, XGBoost, and LightGBM were used in the second layer of the meta-learner; and the optimal stacked integration learning for medal count prediction under time series was subsequently determined. Subsequently, the medal standings for the 2028 Summer Olympics in Los Angeles, USA were predicted under dynamic simulation as the entire sequential system was varied. Based on the parameter-adjusted feature structure of countries without medal counts, two evaluation models were constructed, one of which was initialized with a fixed medal-associated parameter ratio. According to the model framework, the impact of no medal data is parameterized according to the model parameter distribution law to complete the analysis of countries with no medal counts.

**Keywords:** Olympic competitions ,medal prediction, stacked integration, no medal counts.

## 1. Introduction

Typically, predictions of the final medal counts for each country do not rely exclusively on historical data, but are analyzed based on detailed information about the upcoming Olympic Games. This information includes data on current athletes planning to compete, covering the medal standings of all Summer Olympics, the host country, the number of events at each Olympics, etc., and is also broken down into specific sports.

Schlembach et al. (2022) by utilized the Random Forest model to predict the performance of different teams in the Olympics and the effect of the relationship of different variables on the final result. Sharma et al. (2025) presents a new method for accurate prediction of plant diseases using integrated characterizers combined with vector machines (SVM) in agricultural research. Csurilla and Fertő (2024) examines the determinants of Olympic success and finds that population has a positive impact on Olympic success, emphasizing the role of superpower and sport level effects in the Olympics. Shi et al. analyzed on the Summer Olympics data 1912 to 2021 , using random forest model to predict medal predictability, mining the influence of socio-economic factors on medals based on interpretable machine learning method, found that the predictability of women's events is higher than that of men's events, and the traditional advantages of the representative teams (ping pong in China, track and field in the United States, etc.) have a greater impact on medal prediction. Li et al. (2022) in contrast to previous research methods, this study attempts to investigate the number of medals through the total population and GDP per capita, as well as absolute latitude, which have been less mentioned in the past, and offers new possibilities for medal forecasting.

In addition, there are data on achievements (not awarded medals), and this wealth of data helps to provide a scientific basis for strategy development and resource allocation.

## 2. Assessment

### 2.1. Assessment model construction based on model fusion

In order to satisfy the optimal feature-normalized model analysis (Scelles et al., 2020) as well as the feature-normalized model disambiguation evaluation scheme, we have fully considered the degree of relationship between the original data and the actual measured data, and made the corresponding analysis accordingly.

Based on the above analysis, we chose two feature selection methods, one is the penalty term based feature selection method and the other is the tree model based feature selection method.

Penalty-term-based methods use L1/L2 regularization to select features. Using a base model with L-paradigms (CHEN et al., 2025) reduces the dimensionality of individual features and also reduces the risk of overfitting. Therefore, we chose this method to select features by combining LASSO regression, logistic regression and SVM models (Zhu et al., 0) and analyzed the results, we selected features with non-zero coefficients and the results are shown in figure 1.

The most important feature of tree-model-based feature selection method is that this method can calculate the importance of features so as to exclude irrelevant features with low importance. In this paper, we use tree model feature selection methods such as Random Forest, GBD model, XGBoos model, LightGBM to select features and the results are shown in figure 2.

In order to better determine the selection effect of the model, we merge some of the suppliers here to generate the virtual supply conditions (Yu, 2021) to reach the optimal medal type scenario.

| LASSO regression model | logistic regression model | SVM model |
|---|---|---|
| A1 | C3 | C5 |
| B2 | A2 | B2 |
| D1 | C1 | A3 |
| B1 | A4 | B6 |

Figure 1: Selection according to the characteristics of the penalty program

| Random forest model | GBDT model | XGboost Models | LightGBM Models |
|---|---|---|---|
| A0 | D2 | D3 | D4 |
| A10 | B7 | B0 | A20 |
| A5 | C6 | B11 | A11 |
| C8 | A4 | C21 | B16 |

Figure 2: Feature selection based on tree modeling

Screening for information gain:

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2 (p_i) \tag{1}$$

$$\text{Info}_A(D) = \sum_{J=1}^{V} \frac{D_j}{D} \times \text{Info}\,(D_j) \tag{2}$$

We determine the information gain rate:

$$\text{Gain RATIO} = \frac{\text{GAIN}}{\text{SplitINFO}} \tag{3}$$

With the help of this approach, some of the problems of inductive bias algorithms in the ID3 algorithm can be avoided. After that, we analyze the problem of impact on medals according to the distribution patterns of different suppliers. The corresponding parameter distributions as well as the corresponding symbols are shown in table 1.

Table 1: Symbol information

| mathematical symbol | Representative meaning and explanation |
|---|---|
| $Q_t$ | output value |
| $F_i$ | risk factor |
| $X$ | Functional coefficient |
| $C_0$ | thresholds |
| $E_{(w)}$ | loss function |
| $x$ | random sample |
| $\beta$ | Regression coefficient value |
| $\varepsilon_i$ | Randomized perturbation value (RV) |

### 2.2. Feature-normalized model-loss assessment model fusion process

In order to effectively assess the distribution of the feature normalization model, here, in order to better illustrate the implementation effect situation of the identified scheme (Powell et al., 2025), we choose to use the fusion classification model using stack integration learning, selecting Xgboost, Catboost, LR and Lightgbm algorithms with higher AUCs as the base classifiers at the first layer, and selecting Lightgbm algorithm at the second layer,. This treatment will better enable the evaluation of the implementation effectiveness situation and the AUC of the fusion machine algorithm will be more desirable,the parameter distribution is shown in figure 3.

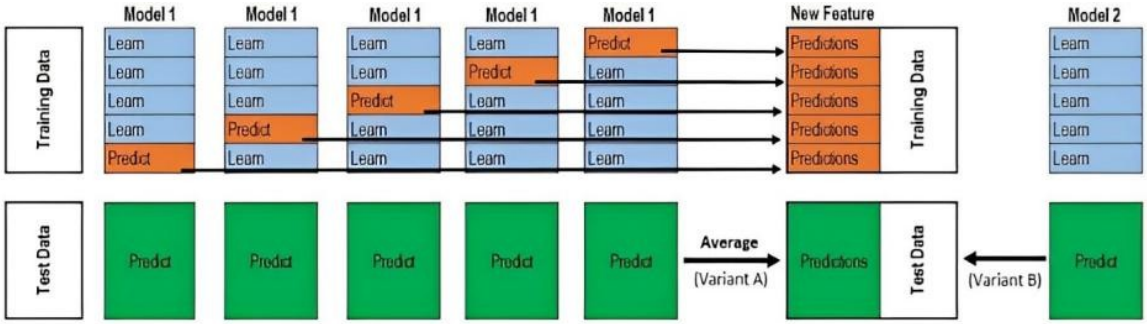


Figure 3: Illustration of Stacking fusion modeling process

After that, this paper establishes a fusion model for machine algorithms based on the fusion classification model of stack integration learning, and improves the fusion model to some extent: instead of using one machine model for each step, four different machine learning algorithms process the evaluations together and use the evaluation results in the outputs of the first-layer base classifiers as the inputs of the second-layer meta-learner.

We choose Xgboost, Catboost, LR and Lightgbm algorithms as the base learners for the first layer of our classification model based on stacked integrated learning, and Lightgbm algorithm as the second layer of the meta-learner.

The final output AUC is shown below: on the test data, the stacked integrated model scores higher than all base classifiers. The AUC of the integrated model is 86.37, which also indicates that the model is relatively stable and does not have an overfitting state.

From the results of figure 4, the classification model based on stacked integrated learning is able to fully integrate the performance advantages of all the base classifiers so as to take advantage of their respective strengths, resulting in a constructed model with better results and stronger generalization ability.

## 3. Results and analysis

### 3.1. Modeling Differentials in Zero-Medal Countries

We are based on the method of changing the proportion of medals, through the modeling algorithm framework in the previous section, retesting the medals in it, and introducing the scarcity data, and effectively filtering the medal points through the idea of differentiation, and the expression of the

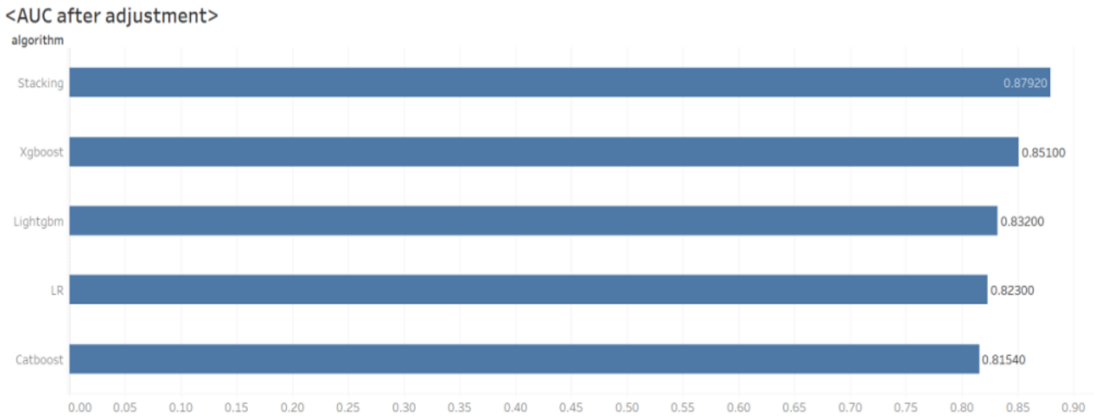


Figure 4: AUC after adjusting the parameters of different algorithms

basic model is as follows:

$$\begin{cases} \frac{\mathrm{dS}}{\mathrm{d}} = \lambda - \frac{(\beta-\beta_1)S(\mathrm{I}+\mathrm{I}_1)}{\mathrm{N}} - \mu\mathrm{S} \\ \frac{\mathrm{dI}}{\mathrm{dt}} = \frac{(\beta S(\mathrm{I}+\mathrm{I}_1)}{\mathrm{N}} - (\alpha+\mathrm{k}+\mu)\mathrm{I} \\ \frac{\mathrm{dI}_1}{\mathrm{dt}} = \frac{(\beta S(\mathrm{I}+\mathrm{I}_1)}{\mathrm{N}} - \mathrm{k}_1\mathrm{I}_1 - \mu\mathrm{I}_1 \\ \frac{\mathrm{dR}}{\mathrm{dt}} = \lambda_1 + \mathrm{kL} + \mathrm{k}_1\mathrm{I}_1 - \mu\mathrm{R} \end{cases} \tag{4}$$

Remember that the total distribution rate $N$ is

$$\mathrm{N} = \mathrm{S} + \mathrm{I} + \mathrm{I}_1 + \mathrm{R} \tag{5}$$

The Jacobian matrix of the equilibrium point equation is

$$J = \begin{cases} -\mu - \frac{\lambda(\beta+\beta_1)}{\mu\left(\frac{\lambda}{\mu}+\frac{\lambda_1}{\mu}\right)} - \frac{\lambda(\beta+\beta_1)}{\mu\left(\frac{\lambda}{\mu}+\frac{\lambda_1}{\mu}\right)}0 \\ 0\frac{\beta\lambda}{\mu\left(\frac{\lambda}{\mu}+\frac{\lambda_1}{\mu}\right)} - C\frac{\beta\lambda}{\mu\left(\frac{\lambda}{\mu}+\frac{\lambda_1}{\mu}\right)}0 \\ 0\mathrm{k}\mathrm{k}_1 - \mu \end{cases} \tag{6}$$

Solving the corresponding characteristic equations yields four characteristic roots:

$$\begin{gathered} x_1 = -\mu; x_2 = -\mu \\ x_{3,4} = \frac{1}{2\left(\lambda+\lambda_1\right)}(D \pm \sqrt{D^2 - 4(\lambda+\lambda_1)(C(\lambda+\lambda_1)(\mu+k_1) - c\beta_1\lambda - \beta\lambda(\mu+k_2))}) \end{gathered} \tag{7}$$

### 3.2. Optimal hyperparameter selection

We chose a suitable algorithm to classify the medal parameterization more efficient-ly and operate for different situations, starting with an optimal selection of parameters.

1. Random Forest Model

We first determine the number of decision trees and the maximum depth using grid search, and then determine the minimum number of samples required for splitting and the minimum number of samples carried by leaf nodes, and Based on table 2, it can be seen that the linear model performs better than the random forest model (Shi et al.). Moreover, the code of the random forest model

Table 2: Random forest model parameters

| Parameter name | Parameter range | Parameter tuning results | After adjusting the reference to F1 | AUC after parameter adjustment |
|---|---|---|---|---|
| max_depth | [1,10] | 211 | 0.85 | 0.80 |
| n_estimators | [10,200] | 60 | | |
| min_samples_split | [10,150] | 150 | | |
| min_samples_leaf | [10,100] | 250 | | |

will run much faster than the linear model because the random forest is easy to make a parallelized method.

2.GBDT model

The first step of our GBDT model tuning is to determine its learning rate first, after determining the learning rate of 0.15, we use the grid to search out the maximum number of optimal values, and then get the optimal parameters. Then we made some adjustments to the minimum number of samples required for classification and the minimum number of samples carried by the leaf nodes, and finally we searched for the optimum according to the AUC for the proportion of subsamples to the total samples to get the final result are obtained as shown in table 3. In the calculation process, in order to prevent overfitting of the model, we adjusted the minimum number of samples required for classification so as to reduce the learning rate to 0.01 and improve the fitting accuracy.

Table 3: GBDT model parameters

| Parameter name | Parameter range | Parameter tuning results | After adjusting the reference to F1 | AUC after parameter adjustment |
|---|---|---|---|---|
| max_depth | [1,15] | 12 | 0.86 | 0.78 |
| n_estimators | [10,200] | 60 | | |
| subsample | [0.5,1] | 0.6 | | |
| learning_rate | [0.01,1] | 0.01 | | |
| min_samples_split | [10,150] | 150 | | |
| min_samples_leaf | [10,100] | 200 | | |

3. XGBoost (Tian et al., 2025) Model

Similar to the RF model and GBDT model, we first determine the learning rate, which is 0.15, then we adjust the number of decision trees and the maximum depth of the decision trees by grid search, and then we compare the optimal proportion of sub-samples to the total samples, and finally we adjust the learning rate of the model according to the results of the AUC to get the best fitting effect in table 4.

Table 4: XGBoost model tuning parameters

| Parameter name | Parameter range | Parameter tuning results | After adjusting the reference to F1 | AUC after parameter adjustment |
|---|---|---|---|---|
| max_depth | [1,15] | 8 | 0.97 | 0.76 |
| n_estmators | [10,200] | 150 | | |
| learning_rate | [0.01,1] | 0.1 | | |
| subsample | [0,2] | 0.9 | | |

## 4. Conclusion

Through the above modeling rules, we have established the judgment model as well as the specific rules, and by analyzing the influence of different laws, we have derived the preliminary laws and given the specific parameterization scheme under the influence of multi-parameterization. The program is shown in Figure 5 and Figure 6.

| categ orization | Indicator parameter series index q | Elimination of the series index T | show |
|---|---|---|---|
| Class 1 Medal | q<87.5 | 0 | Very favorable diffusion of indicator parameters |
| Level 2 Medal | 87.5<q<147 | 0 | Facilitate dissemination of indicator parameters |
| Level 3 Medal | 147<q<300 | 0 | Disadvantageous dissemination of indicator parameters |
| Class 4 Medal | q<87.5 | 1 | Disadvantageous dissemination of indicator parameters |
| Meda l level 5 | 87.5<q<147 | 1 | Very unfavorable dissemination of indicator parameters |
| Meda l level 6 | 147<q<300 | 1 | Very unfavorable dissemination of indicator parameters |

Figure 5: Rules for categorizing project relevance levels

| Classification | Indicator Time series index a | Elimination series index T | Description |
|---|---|---|---|
| Time factor level 1 | a<14.2 | 0 | Very favorable for indicator parameter diffusion |
| Time factor level 2 | 14.2<a<19 | 0 | Favorable diffusion of indicator parameters |
| Time factor level 3 | 19<a<78.9 | 0 | Less favorable to the diffusion of indicator parameters |
| Time factor level 4 | 78.9<a<87.5 | 1 | Unfavorable to diffusion of indicator parameters |
| Time factor level 5 | 87.5<a<100.41 | 1 | Very unfavorable for diffusion of indicator parameters |

Figure 6: Medal classification rules

We used the algorithmic modeling described above, combined with validated data as well as fused dismembered data, to ultimately solve for the distribution of the medal influencing factor components of the 2028 Summer Olympics in Los Angeles, California,as shown in figure 7.

## References

Jian-yu CHEN, Shi-jing FENG, Shou-zheng YAN, Xiao-huan WANG, and Min-min ZHANG. Optimization and application of ozone concentration prediction model based on lightgbm machine learning. *Arid Environmental Monitoring*, 39(01):43–48, 2025.

Gergely Csurilla and Imre Fertő. How to win the first olympic medal? and the second? *Social Science Quarterly*, 105(5):1544–1564, 2024. doi: https://doi.org/10.1111/ssqu.13436.

F. Li, W. G. Hopkins, and P. Lipinska. Population, economic and geographic predictors of nations' medal tallies at the pyeongchang and tokyo olympics and paralympics. *Frontiers in sports and active living*, 2022. doi: https://doi.org/10.3389/fspor.2022.931817.

| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 | F11 | F12 | F13 | F14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1t | 59.81 | | 7.68 | 1.41 | 1.73 | 13.05 | 6.04 | 2.18 | 0.35 | 0.97 | 4.5 | 0.12 | | |
| 2t | 67.71 | | 7.37 | | 1.35 | 11.15 | 2.39 | 2.51 | 0.2 | 1.38 | 4.18 | 0.11 | | |
| 3t | 61.58 | 3.04 | 10.95 | 3.12 | 0.11 | 4.16 | | 0.7 | 30.61 | 6.22 | 6.34 | 0.23 | | |
| 4t | 65.23 | | 9.67 | 7.12 | 1.65 | 6.44 | 2.06 | 8.18 | | | 0.79 | | | 0.36 |
| 5t | 61.71 | | 12.37 | 5.87 | 1.11 | 5.5 | 2.16 | 5.09 | 1.41 | 2.86 | 0.7 | 0.1 | | |
| 6t | 87.05 | | 5.19 | 2.01 | | 4.06 | | 0.78 | 0.25 | | 0.66 | | | |
| 7t | 69.61 | | 9.42 | 6.11 | 0.82 | 3.93 | 1.74 | 3.87 | | | 1.17 | | | 0.39 |

Figure 7: Selected data on the components of the indicator for factor A

Cormac Powell, David B. Pyne, Emmet Crowley, and Iñigo Mujika. What it takes to win: Examining predicted versus actual swimming performances at the paris 2024 olympic games, and what comes next. *International Journal of Sports Physiology and Performance*, 20(4):504 – 514, 2025. doi: 10.1123/ijspp.2024-0409.

Nicolas Scelles, Wladimir Andreff, Liliane Bonnal, Madeleine Andreff, and Pascal Favard. Forecasting national medal totals at the summer olympic games reconsidered. *Social Science Quarterly*, 101(2):697–711, 2020. doi: https://doi.org/10.1111/ssqu.12782.

Christoph Schlembach, Sascha L. Schmidt, Dominik Schreyer, and Linus Wunderlich. Forecasting the olympic medal distribution – a socioeconomic machine learning model. *Technological Forecasting and Social Change*, 175:121314, 2022. doi: https://doi.org/10.1016/j.techfore.2021.121314.

Piyush Sharma, Devi Sharma, and Sulabh Bansal. Optimum rbm encoded svm model with ensemble feature extractor-based plant disease prediction. *Chemometrics and Intelligent Laboratory Systems*, 258:105319, 03 2025. doi: 10.1016/j.chemolab.2025.105319.

Huimin Shi, Dongying Zhang, and Yonghui Zhang. Can olympic medals be predicted?

Y. Tian, F. Yan, W.and Gao, and et al. Prediction model for defects in small billets based on pso-dexgboost. *Energy for Metallurgical Industry*, 44(2):75–80, 3 2025.

Jun Yu. Sports intelligence in olympic events. *Competitive Intelligence*, 17(05):2–9, 2021. doi: 10.19442/j.cnki.ci.2021.05.003.

Lin Zhu, Yingxi Zhu, Xiaoyu Gu, and Junchao Lin. Research on financial risk prediction model based on news text and ls-svm. *International Journal of High Speed Electronics and Systems*, 0 (0):2540382, 0. doi: 10.1142/S0129156425403821.