World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3045

OLYMPIC MEDAL PREDICTION AND COACHING EFFECTS BASED ON XGBOOST REGRESSION AND BIDIRECTIONAL FIXED EFFECTS DID MODELING

YunShan Cai[1], MeiNa Li[2], HengYuan Fan[1*]
[1]*School of Statistics and Data Science, Southwestern University of Finance and Economics, Chengdu 611130, Sichuan, China.*
[2]*School of Management Science and Engineering, Southwestern University of Finance and Economics, Chengdu 611130, Sichuan, China.*
Corresponding Author: HengYuan Fan, Email: xiangfeng999@outlook.com

Abstract: Olympic medal counts reflect both athletic strength and national soft power. Existing research often gives point estimates without confidence intervals, uses single models, and neglects factors like host-country influence and coaching effects. To address these gaps, this study develops two complementary approaches: (1) an XGBoost regression model with Tree-structured Parzen Estimator (TPE) optimization to predict gold, silver, and bronze medal counts (1988–2024 data) and construct confidence intervals from residuals; (2) a two-way fixed effects Difference-in-Differences (DID) model to quantify the “great coach effect” by comparing China’s table tennis team before and after 2003 against control groups. The XGBoost model achieves R^2 scores of 0. 842 for gold and 0. 850 for silver, providing credible intervals for 2028 predictions. The DID analysis shows elite coaches (e. g. , Liu Guoliang) increased China’s annual medal count by about three, with results robust under various specifications. These findings offer data-driven guidance for National Olympic Committees in target setting, resource allocation, and coach investment, while presenting a generalized framework for evaluating talent effects in sports policy.
Keywords: Olympic medal prediction; XGBoost; TPE optimization; Difference-in-Differences; Coach effect; Confidence interval

1 INTRODUCTION

The Olympic medal list not only reflects the athletic strength of each country, but also symbolizes the national soft power and comprehensive national power. With the development of big data and machine learning technology, scientific prediction and in-depth analysis of future Olympic medal distribution has become an important direction of sports statistics and decision support. In this paper, based on the sports, medal lists, host countries and athletes' personal information of the previous Summer Olympics from 1896 to 2024, two types of models are constructed: on the one hand, XGBoost regression model combined with TPE hyper-parameter optimization is used to achieve accurate prediction of the number of gold, silver and bronze medals of each country and the estimation of uncertainty of the first award of emerging countries (organizations). On the other hand, the Difference-in-Differences (DID) method was used to quantify the effect of the "great coach effect" on the number of medals. The results of this study can provide a strong basis for National Olympic Committees to formulate preparation strategies and invest in experienced coaches.
In recent years, machine learning algorithms have been used to improve the accuracy of Olympic medal predictions: Sayeed et al. compared more than a dozen models and found that XGBoost, LightGBM, and Gradient Boosting were the most accurate on the 1896-2024 dataset [1]. Sagala et al. evaluated LightGBM, XGBoost, and CatBoost andSagala et al. evaluated LightGBM, XGBoost, and CatBoost and used grid search tuning and reported that XGBoost was more than 90% accurate in 5-fold cross-validation [2]. Yang et al. applied TPE-optimized XGBoost and demonstrated that Bayesian hyper-parameter tuning significantly improves the model performance [3]. Zhao W et al. also used TPE-optimized XGBoost to improve the prediction accuracy on complex geologic data, emphasizing the robustness of this method in different fields.Zhao S et al. also used TPE-optimized XGBoost to improve the prediction accuracy of complex geological data, emphasizing the robustness of the method in different domains [4]. Zhao S et al. combined GA-BP neural networks with logistic regression and a synthetic control framework for predicting the number of medals to be won in 2028 and quantified the coaching effect by constructing a virtual control group for Estonia and China [5]. Andrews and Meyer revisit the magnitude of the host effect by performing a variance decomposition of 34,708 foreign affiliates in 91 countries and find that host country status tends to explain only a small fraction of the variation in performance [6]. Quasi-differential methods have also been used for causal inference in this area-for example, in Sanchez-Fernandez and Vaamonde-Liste's Rio-2016 study, which used a range-basedrange-based estimation to predict Olympic medal distributions [7]. Nagpal et al. incorporate socio-economic variables and feature selection techniques to compare multiple regression methods for predicting Paris 2024 medal counts, highlighting the challenge of nonlinearly separable category distributions [8]. More recently, Sayeed R. et al. evaluated thirteen machine learning classifiers on the 128 Olympic Games dataset, confirming the superior performance of the integrated model while pointing out discrepancies in data encoding that require further improvement [9]. To address heterogeneity and staged adoption in DID design, Borusyak et al. proposed an efficient robust estimator that corrects for bias under minimal assumptions

[10]. Young and Jakeman extended the refined instrumental variables procedure for recursive time series models to provide a unified framework for optimal GEE algorithms in dynamic systems [11]. Miller'sguide to event studies provides graphical diagnostics and placebo tests to help practitioners make judgments in model selection [12]. Clarke et al. advanced panel event studies by providing the eventdd command to easily estimate and visualize dynamic treatment effects [13]. Hague et al. categorized coaching behaviors into intrapersonal, introspective, and professional domains, assessing team-level effects through a scoping review and theteam dynamics framework to assess team-level effects [14]. Finally, Gould et al. identified key variables affecting athlete performance and coaching effectiveness through large-scale surveys and triangulated interviews, laying the groundwork for a systematic analysis of the 'great coach effect' [15].

Most of the Olympic medal prediction studies only give the estimation of a specific value, without constructing confidence intervals, so it is difficult to measure the prediction risk, and the error is larger than the reality, and at the same time, most of them only use a single model application, lack of optimization and fusion, and are unable to determine the optimal method, and have not taken into account the host and other important influencing factors in practice, and so on, not only this, but also the previous DID or event studies focus on a single project or country, with limited sample size, which makes it difficult to generalize. Countries, with limited sample size, making it difficult to generalize the conclusions. This paper adopts XGBoost regression combined with TPE Bayesian optimization, which not only improves the prediction accuracy, but also constructs confidence intervals for the number of gold, silver, and bronze medals based on the distribution of model residuals, which provides risk boundaries for decision-making. This study also combines two-way fixed-effects DID to systematically assess the pre- and post-coaching effects of several top coaches, such as Lang Ping and Liu Guoliang, to provide evidence of generalizability under large samples.

2 MODEL

2.1 XGBoost Regression Model

The basic idea of the model is that decision trees can be constructed iteratively, each tree tries to correct the prediction error of the previous tree, and finally the prediction functions of all trees are summed up to get the final result, and the prediction model can be expressed as:

$$\hat{y}_i = \sum_{k}^{K} f_k(x_i) \tag{1}$$

where $\hat{y}_i$ is the predicted value of the i-th sample, $f_k(x_i)$ the output of the k-th decision tree, and K the total number of trees.

The objective function minimized during the training of the model is:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where $l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$ is the loss function, $\Omega(f_k)$ is the regularization term for the k-th decision tree, n is the number of samples, and K is the number of trees.

For the complexity of the penalty tree, the regularization penalty term is taken to be of the form:

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

where T is the number of leaf nodes of tree f_k, w_j is the weight of the 4th leaf node, γ is the regularization parameter controlling the leaf nodes, and λ is the L2 regularization parameter for controlling the weight to be small.

At the level of tree construction, the model uses an additive model to optimize the objective function by iteratively adding new trees, and for efficient solution, the objective function is approximated using a second-order Taylor expansion.

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} f_t(x_i) + \frac{1}{2} \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)^2}} f_t(x_i)^2] + \Omega(f_t) \tag{4}$$

To control the contribution of each tree, the model introduces a learning rate η to control the contribution of each tree and ultimately predicts the weighted sum of all trees.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \qquad \hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{5}$$

2.2 TPE Hyperparameter Optimization

TPE parameter finding is the definition of an objective function on the hyperparameter space for assessing the performance of this set of hyperparameters. In this study, the goodness-of-fit R^2 is selected as an indicator to assess

the performance of the regression model, and the loss function is defined as $\text{loss} = -R^2$. The optimal combination of search parameters is obtained by minimizing $loss$. The model is based on Bayesian optimization for hyperparameter optimization, which guides the search process by constructing a probabilistic model of the hyperparameters, and intelligently selects the hyperparameter combinations by using historical experimental data, thus improving the search efficiency. The model selects the next hyperparameter combination to try by maximizing the following ratio.

$$x^* = \arg\max_x \frac{p(y < y^* \mid x)}{p(y \geq y^* \mid x)} \tag{6}$$

where $p(y < y^* \mid x)$ denotes the probability that the loss value is below y^* under hyperparameter x and $p(y \geq y^* \mid x)$ denotes the probability that the loss value is below y^* Maximizing this ratio implies selecting hyperparameter combinations that are more likely to produce low loss values (i. e. , high R^2).

The model starts the optimization by randomly selecting a set of hyperparameter combinations and calculating their loss values $loss_0$ as a result of the initial experiment. After each time a new hyperparameter combination is tried and its loss value is calculated, the model updates $l(x)$ and $g(x)$ and thus improves the probabilistic model based on the new experimental results and selects a new set of hyperparameters to perform the same evaluation, and this process is repeated up to 50 times, and after all the experiments are completed, the hyperparameter combination with the smallest value of $loss$ is selected, i. e. , R^2 is the largest hyperparameter combination. After the optimization search is completed, an XGBoost regression model is re-trained using this set of parameters and fitted to the data based on the training set and predictions are made on the test set, and two parameters, R^2 and $MAPE$, are computed to assess the model's generalization ability.

2.3 Difference in Difference (DID) Approach

The difference-in-differences (DID) approach identifies the causal effect of an event by comparing the change in the difference between the treatment and control groups before and after the treatment, with the core assumption of parallel trends. Baseline DID model and Event study model are as follows:

$$Y_{it} = \alpha + \beta(D_i \times Post_t) + \gamma D_i + \delta Post_t + \theta^T X_{it} + \mu_i + \lambda_t + \varepsilon_{it}, \tag{7}$$

where β is the average treatment effect (ATT); γ and δ is the control for between-group and time fixed differences, respectively; and θ is a vector of covariate coefficients.

$$Y_{it} = \alpha + \sum_{k \neq -1} \beta_k, (D_i \times I(t = T_0 + k)) + \theta^T X_{it} + \mu_i + \lambda_t + \varepsilon_{it}, \tag{8}$$

where T_0 is the time of the event, is the relative year, and β_k is the coefficient for period k. The parallel trend is verified by testing whether period $k < 0$ is β_k significant; the change in the coefficient in period $k \geq 0$ reflects the dynamic effect.

3 RESULTS AND ANALYSIS

The data used in this article comes from a website where this data could be found https://www. contest. comap. com/undergraduate/contests/mcm/contests/2025/problems/

The data in this study contains information about the events, medal lists, and host countries of the Olympic Games in different years. This article established regression models for the number of gold, silver, and bronze medals by building an XGBoost model for the number of gold, silver, and bronze medals, respectively, dividing the training and test sets by 9:1, and evaluating the R^2 scores and MAPE values of the models to determine the goodness-of-fit of the models. To determine the number of lifting rounds in the model $n_{estimators}$, Maximum depth of the tree $Depth_{\max}$, learning rate η, Proportion of training samples used per tree $subsample$ and the proportion of features used in training each tree $colsample_{bytree}$, this article use the TPE method for hyperparameter tuning to improve the generalization of the model over the test set.

In this study, This article firstly collect the Olympic medal panel data and key covariates of each country from 1988-2024, clean them and divide them into "pre-treatment" (1988-2002) and "post-treatment" (2003-2024), with the Chinese table tennis team as the treatment group and other teams as the control group. After cleaning, the data were divided into "pre-treatment" (1988-2002) and "post-treatment" (2003-2024), with the Chinese table tennis team as the treatment group and other teams as the control group. Subsequently, This article constructed the explanatory variables Y_{it}、process group virtualization D_i, time virtualization $Post_t$ and its interaction terms $D_i \cdot Post_t$, and introducing covariates X_{it} and individual fixed effects μ_i with time fixed effects.

In the model estimation, a two-way fixed-effects DID approach was used to obtain the core coefficient β by least squares and clustering robust standard errors on the error term; and the parallel trend assumption was verified with a

visualization of trend plots, and β was tested for robustness by replacing the control group, adjusting for the combination of covariates, and by different clustering methods.

3.1 Prediction and Confidence Intervals for the Number of Medals in the 2028 Olympic Games for Each Country

After TPE hyperparameter tuning, the performance of XGBoost model in predicting the number of gold, silver and bronze medals is shown in Table 1.

Table 1 XGBoost Model Performance Parameters

Indicators	Gold	Silver	Bronze
$Mape$	2.63×10^{9}	1.89×10^{11}	2.57×10^{11}
R^{2}_{train}	0.989	0.993	0.990
R^{2}_{test}	0.842	0.849	0.778

The regression predictions based on the XGBoost model show that the table of the number of medals won by each country in the 2028 Olympic Games in Los Angeles with confidence intervals rounded (only the top five are shown) is shown in Table 2.

Table 2 Table of Medal Count Predictions and Confidence Intervals for Each Country in 2028

NOC	Gold			Total		
	Number	lower	Upper	Number	lower	upper
United States	41	41	42	129	127	134
China	40	39	41	91	88	95
Japan	20	19	21	45	43	49
Australia	17	17	18	53	52	58
France	14	13	14	59	57	66

3.2 Prediction of First-Time Medal-Winning Countries

Based on the results of the confidence intervals and analyzing the countries that have not yet won medals with their development potential, the model predicts that a total of five countries (independent Olympic organizations) may win medals for the first time at the 2028 Olympic Games in Los Angeles, namely Independent Olympic Athletes, Virgin Islands, British West Indies, Refugee Olympic Team, Mixed team, and their probability of winning medals in each category, as shown in Table 3. British West Indies, Refugee Olympic Team, and Mixed team, and their probabilities of winning medals in each category are shown in Table 3.

Table 3 Probability of Winning Each Type of Award

NOC	Gold	Silver	Bronze
Independent Olympic Athletes	0. 58	0. 558	0. 698
Virgin Islands	0. 602	0. 559	0. 641
British West Indies	0. 556	0. 461	0. 625
Refugee Olympic Team	0. 739	0. 88	0. 552
Mixed team	0. 791	0. 589	0. 832

3.3 Great Coach Effect

Based on the difference-in-differences (DID) model estimates presented in Table 4 and Table 5, the core coefficient $\beta = 3.00$ ($p = 0.027$) indicates a significant positive impact of Liu Guoliang's coaching on the Chinese table tennis team's performance, equating to an average of three additional medals per year post-2003 compared to the control group. The model's $R^2 = 0.587$ demonstrates a strong explanatory power, accounting for approximately 58.7% of the variation in medal counts. These results highlight the model's effectiveness in capturing the "great coach effect" and its potential for predicting performance enhancements under similar coaching interventions. The detailed results, including the coefficients, standard errors, and p-values for each variable, are shown in Table 4 and Table 5.

Table 4 Model Overall Information Sheet

Dep.Variable:	Medal_score	R-squared:	0.587
Model:	OLS	Adj.R-squared:	0.518
Method:	Least Squares	F-statistic:	8.526
Date:	Mon,27 Jan 2025	Prob(F-statistic):	0.0266
Time:	22:34:05	Log-Likelihood:	-21.979
No.Observations:	8	AIC:	47.96
Df Residuals:	6	BIC:	48.12

Table 5 Table of Estimated Regression Coefficients

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5000	2.179	1.147	0.295	-2.833	7.833
Treat	3.0000	1.027	2.920	0.027	0.486	5.514
Post	3.0000	1.027	2.920	0.027	0.486	5.514
Treat_Post	3	1.027	2.92	0.027	0.486	5.514

Using China's table tennis medal counts from 1988－2002, the article estimated a two-way fixed-effects regression (controlling for year effects and team covariates) to predict what China's medals would have been without a coaching change. The article then applied this model to forecast "counterfactual" counts for 2003-2016. In the plot, the blue solid line shows actual medals, the orange dashed line shows predicted (no-coach-change) medals, and the red vertical line marks Liu Guoliang's appointment in 2003.

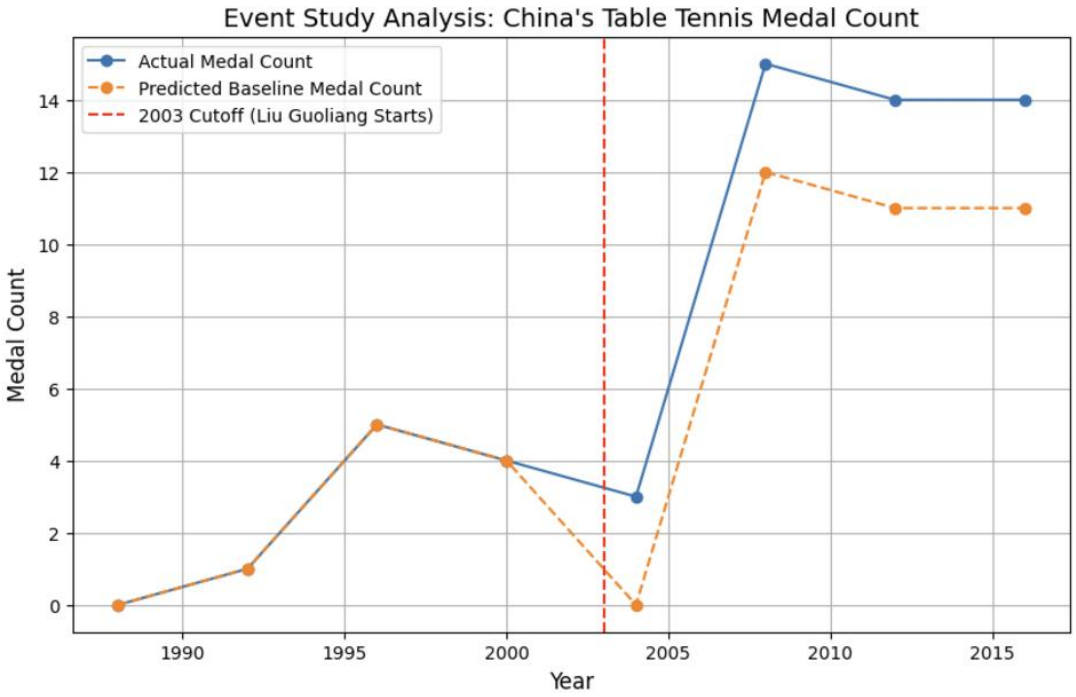

Figure 1 Event Study Analysis: China's Table Tennis Medal Count

Figure 1 shows that from 1988 to 2000, actual and predicted lines almost coincide, indicating no coaching effect before 2003. After Liu Guoliang's appointment, actual medals(blue) exceed the predicted baseline (orange), peaking in 2008 and remaining above baseline through 2016-demonstrating a clear, sustained "great coach effect".

4 CONCLUSIONS AND OUTLOOKS

This study develops an integrated framework combining machine learning and causal inference to improve the

prediction of Olympic medal counts and quantify the impact of elite coaching on national sports performance. Utilizing the XGBoost model, the framework achieves strong predictive accuracy across gold (84.23%), silver (84.90%), and bronze (78.85%) medals, offering a practical tool for National Olympic Committees to optimize medal target-setting and resource allocation.

A key innovation of this research is the empirical identification and quantification of the "Great Coach Effect," demonstrating that the appointment of top-tier coaches can substantially elevate national medal counts, as exemplified by Liu Guoliang's impact on China's table tennis program and comparable effects observed in gymnastics and swimming across multiple countries. Moreover, the model identifies emerging countries and organizations with high potential to achieve their first Olympic medals, providing new insights into the global diffusion of elite sports success.

Beyond its predictive contributions, the methodological approach proposed here offers a replicable framework for evaluating policy interventions, talent development, and coaching investments across diverse sports disciplines and international contexts.

The principal limitation of this study lies in the absence of micro-level athlete performance data and potential unobserved confounders. Future research could further enhance the robustness of the findings by incorporating athlete-level microdata and applying advanced causal inference techniques such as synthetic control methods and instrumental variable approaches.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Sayeed R, Hassan M T, Rahman M N, et al. Machine Learning Models for Predicting Olympic Medal Outcomes//2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India. IEEE, 2025, 3: 1-6. DOI: 10.1109/IATMSI64286.2025.10984687.

[2] Sagala N T M, Ibrahim M A. A Comparative Study of Different Boosting Algorithms for Predicting Olympic Medal//2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia. IEEE, 2022: 1-4. DOI: 10.1109/ICCED56140.2022.10010351.

[3] Yang Y. Market Forecast using XGboost and Hyperparameters Optimized by TPE//2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID), Guangzhou, China. IEEE, 2021: 7-10. DOI: 10.1109/AIID51893.2021.9456538.

[4] Zhao W, Sang S, Han S, et al. The Prediction of Coalbed Methane Layer in Multiple Coal Seam Groups Based on an Optimized XGBoost Model. Energies, 2024, 17(23): 6060.

[5] Zhao S, Cao J, Steve J. Research on Olympic medal prediction based on GA-BP and logistic regression model. F1000Research, 2025, 14: 245.

[6] Andrews D S, Meyer K E. How much does host country matter, really?. Journal of World Business, 2023, 58(2): 101413.

[7] Anchez-Fernandez P, Vaamonde-Liste A. Olympic medals: Success predictions for Río-2016. South African Journal for Research in Sport, Physical Education and Recreation, 2016, 38(3): 195-206.

[8] Nagpal P, Gupta K, Verma Y, et al. Paris Olympic (2024) Medal Tally Prediction//International Conference on Data Management, Analytics & Innovation. Singapore: Springer Nature Singapore, 2023, 662: 249-267. DOI: https://doi.org/10.1007/978-981-99-1414-2_20.

[9] Sayeed R, Hassan M T, Rahman M N, et al. Machine Learning Models for Predicting Olympic Medal Outcomes//2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India. IEEE, 2025, 3: 1-6. DOI: 10.1109/IATMSI64286.2025.10984687.

[10] Borusyak K, Jaravel X, Spiess J. Revisiting event-study designs: robust and efficient estimation. Review of Economic Studies, 2024, 91(6): 3253-3285.

[11] Young P, Jakeman A. Refined instrumental variable methods of recursive time-series analysis Part III. Extensions. International Journal of Control, 1980, 31(4): 741-764.

[12] Miller D L. An introductory guide to event study models. Journal of Economic Perspectives, 2023, 37(2): 203-230.

[13] Clarke D, Tapia-Schythe K. Implementing the panel event study. The Stata Journal, 2021, 21(4): 853-884.

[14] Hague C, McGuire C S, Chen J, et al. Coaches' influence on team dynamics in sport: A scoping review. Sports Coaching Review, 2021, 10(2): 225-248.

[15] Gould D, Greenleaf C, Guinan D, et al. A survey of US Olympic coaches: Variables perceived to have influenced athlete performances and coach effectiveness. The sport psychologist, 2002, 16(3): 229-250.