Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/106/2025.22916

A Data-Driven Approach to Predicting Olympic Medal Distribution: Integrating Machine Learning and Graph Theory

Yunhan Xu

Southwestern University of Finance and Economics, Chengdu, China
xuyunhan2005@hotmail.com

Abstract: This study proposes a comprehensive medal prediction model for the 2028 Summer Olympics, providing valuable insights for the Olympic Committee. To achieve this, we conduct rigorous data preprocessing and analysis, employing K-means clustering to classify countries into distinct groups based on key attributes. We introduce an innovative evaluation framework that quantifies national competitiveness using weighted scores, forming the foundation for our medal prediction model, which integrates regression analysis and the Informer time-series model. A key focus of our research is to explore the relationship between sporting events and national medal counts by comparing Spearman correlation coefficients, while also empirically validating the host nation advantage. For medal table prediction, we implement a stacked ensemble model, combining linear regression, random forest, support vector regression (SVR), K-nearest neighbors (KNN), and XGBoost, ensuring robustness and accuracy. To address the first-time winning country problem, we reformulate it as a binary classification task using logistic regression, evaluating performance through accuracy metrics and confusion matrix analysis. Additionally, we investigate the "great coach" effect by modeling it as a maximum flow problem in graph theory, proving its existence via bottleneck capacity constraints. Furthermore, we conduct uncertainty quantification and hyperparameter tuning to enhance the model's reliability and predictive performance. Our findings contribute to a data-driven understanding of Olympic medal distributions, offering a novel perspective on factors influencing national athletic success.

Keywords: Machine Learning, Time-series Analysis, National Competitiveness, Regression Modele

1. Introduction

The Olympic Games, as the world's largest and most influential sporting event, bring together top athletes from across the globe to compete at the highest level. The distribution of Olympic medals not only reflects a nation's investment in sports development, resource allocation strategies, and athletic capabilities, but also mirrors the evolving dynamics of global competitiveness in sports [1–2]. As highlighted by the results of the 2024 Paris Olympics, the United States topped the medal tally with 126 medals, while both the United States and China shared the lead in gold medals, each winning 40. The host nation, France, ranked fifth with 16 gold medals. Moreover, countries such as Albania, Cape Verde, Dominica, and Saint Lucia earned their first-ever Olympic medals—an achievement that

DOI: 10.54254/2753-8818/106/2025.22916

underscores the rising prominence of smaller nations in the global sports arena. These developments offer valuable data and motivation for analyzing and forecasting future Olympic medal distributions.

Forecasting based on socio-economic indicators has a long-standing tradition in academic research, particularly within the social sciences. Historically, such predictions have informed public policy and program planning by helping avoid adverse outcomes that could limit national development. This approach has been widely applied in fields such as economics, public health, civil engineering, ecology, and urban planning. Within the realm of sports economics, predicting Olympic performance—typically through medal forecasts—has garnered significant scholarly attention in recent decades. For governments that invest heavily in elite athlete training programs, such forecasts serve as benchmarks to evaluate the return on investment and guide future funding decisions. Accurate medal predictions can also shape national pride and promote public engagement in sports, which may lead to broader social benefits such as reduced healthcare costs.

Despite the relevance of traditional forecasting methods—which rely on historical trends, economic indicators, and the host country advantage—these approaches often fall short in capturing the complex, data-driven factors underpinning Olympic success. In recent years, advancements in machine learning and time-series forecasting have enabled more precise, scalable, and interpretable models for predicting medal outcomes [3-5].

In this context, our study aims to develop a comprehensive data-driven framework for predicting medal distributions at the 2028 Los Angeles Olympics. We leverage historical medal data, event counts, and the host country effect while systematically analyzing the key variables influencing medal shifts. Specifically, we propose an innovative national competitiveness evaluation framework that quantifies athletic strength through a weighted scoring mechanism, forming the foundation of our prediction model. By applying K-means clustering, we classify countries based on core attributes, thereby enhancing model interpretability. We further integrate regression analysis and the Informer time-series model to improve predictive accuracy and stability. Additionally, Spearman correlation analysis is employed to examine the relationship between specific events and medal counts, and we empirically validate the host advantage hypothesis.

To ensure model robustness, we implement uncertainty quantification and hyperparameter optimization, refining both performance and reliability. This study not only improves the accuracy of Olympic medal predictions but also provides actionable insights for sports development strategies and policy decisions. The key contributions of this research are as follows:

1. Proposing a comprehensive Olympic medal prediction framework that integrates machine learning, time-series forecasting, and clustering techniques.
2. Empirically analyzing key factors influencing Olympic medal distribution [6], including host country effects, event-specific advantages, and national competitiveness metrics.
3. Innovatively modeling the “great coach” effect using graph theory’s maximum flow to provide theoretical support for the influence of coaches in improving a nation’s athletic performance.
4. Conducting model evaluation and uncertainty quantification to ensure the robustness and interpretability of the predictions.

By combining data-driven methods with domain knowledge, this research advances the modeling of Olympic medal prediction and offers practical decision-making support for sports managers, policymakers, and sports science researchers.

2. Problem statement

2.1. Problem assumption

To ensure the validity and applicability of the proposed model for predicting Olympic medal distributions, the following assumptions are made:

DOI: 10.54254/2753-8818/106/2025.22916

Assumption 1: The historical data used in this study, including but not limited to medal counts, athlete information, and event settings, are assumed to be accurate, complete, and representative of future trends in Olympic medal distribution.

Assumption 2: It is assumed that the medal counts of different countries are independent of each other. Features extracted from historical data, such as individual athlete performance and overall national competitiveness, can be effectively quantified and weighted to predict future Olympic performance.

Assumption 3: The variation in a country's total medal count is primarily influenced by historical performance, the structure of Olympic events, athlete capabilities, and home-field advantage.

Assumption 4: It is assumed that national sports policies, resource allocation for athlete development, and the rules governing Olympic events will remain relatively stable over time and will not be significantly disrupted by unforeseen external events.

Assumption 5: Athlete ability is assumed to be evenly distributed within each country. Consequently, a nation's overall competitiveness can be effectively measured by aggregating the performance scores across individual events.

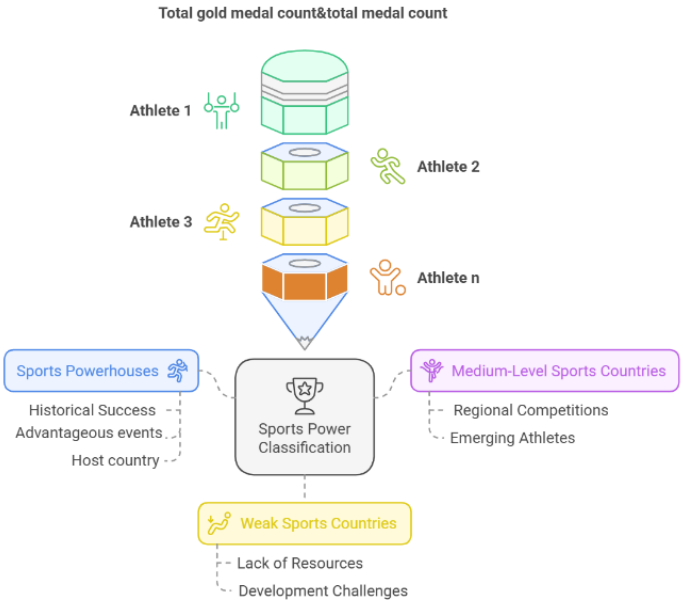

Figure 1: Country classification

2.2. Data preparation

For each Athlete A_i, the calculation of his/her score is based on his/her performance in the last five Olympic Games. Each Athlete is allocated a different number of points for each Olympic Games, depending on the medals he or she has won (or participated in).

Points allocation rules:

Gold Medal: 10 points,Silver medal: 6 points, Bronze: 3 points, No award: 1 point

Weighting: In order to assign decreasing weights to the historical performances of the athletes, we set decreasing weight factors for each Olympic Games. Specifically, the weight of the kth Olympics is w_k , where $w_k = 1 - 0.2(\mathrm{k} - 1)$ and $\mathrm{k} \in \{1,2,3,4,5\}$, which corresponds to the last five Olympics. Namely:

$$\mathrm{w1} = 1, \quad \mathrm{w2} = 0.8, \quad \mathrm{w3} = 0.6, \quad \mathrm{w4} = 0.4, \quad \mathrm{w5} = 0.2 \tag{1}$$

DOI: 10.54254/2753-8818/106/2025.22916

Therefore, the score S_{t_k} , A_i of athlete A_i in the kth Olympic Games is determined by his/her performance in previous Olympic Games, and his/her weighted total score P_{A_i} is calculated using the following formula:

$$P_{A_i} = \sum_{k=1}^{5} S_{t_k}$$

This scoring mechanism, as shown in Fig. 1, allows recent achievements to carry more significance while still considering historical performance, thereby providing a balanced and temporally aware evaluation of each athlete's overall Olympic record.

3. Method

To construct a robust and interpretable medal prediction framework for the 2028 Summer Olympics, we employ a diverse set of machine learning models that collectively capture both linear dependencies and complex nonlinear interactions among the influencing factors. The selected methods—ranging from classical linear regression to advanced ensemble techniques such as Random Forest, Gradient Boosting Decision Trees (GBDT), and XGBoost—are well-suited to model the multifaceted nature of Olympic performance indicators. Each model is rigorously evaluated based on its capacity to generalize across nations with heterogeneous profiles, account for temporal trends, and handle feature sparsity. In the following subsections, we briefly describe the mathematical formulation and objective functions of each method employed in this study.

(1) Linear Regression

Linear regression predicts the target value by fitting a straight line and optimises the number of models by minimising the mean square error. Target formula:

$$\hat{y} = w1x1 + w2x2 + \ldots + wnxn + b \tag{2}$$

where w1, w2, ... , wn are the weights of the model, x1, x2, ... , xn are the input features and b is the bias term. Linear regression optimises the model parameters by minimising the loss function (usually the mean square error) with the objective function:

$$\text{Obj} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

(2) Random Forest

Random forests reduce the risk of overfitting a single tree by training multiple decision trees with randomly selected characteristics and samples. Each tree is trained independently of the others, and the final prediction is the average or majority vote of all the trees. The objective function of a random forest can be expressed as follows:

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) \tag{4}$$

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

(3) GBDT

GBDT improves the model performance by progressively fitting the residuals of the data, optimising the loss function and adding a regularisation term. The objective function of the algorithm is:

$$\text{Obj} = \sum_{i=1}^{n} (y_i - \hat{y}_i^{(t-1)} - f_t(x_i))^2 \tag{5}$$

DOI: 10.54254/2753-8818/106/2025.22916

(4) XGBoost

XGBoost further improves the generalisation and overfitting resistance of the model by optimising the objective function with a regularisation term.The objective function of XGboost can be expressed as:

$$\text{Obj} == \sum_{i=1}^{n}(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) = \sum_{i=1}^{n}(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \quad (6)$$

(5) CatBoost

Similar to GBDT, describing CatBoost adds a regularisation term to the objective function to reduce overfitting and improve model robustness. The objective function of CatBoost can be expressed as:

$$\text{Obj}= Obj^t = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{j=1}^{T} \Omega(f_j) \quad (7)$$

(6) Extra Trees

Show Extra Trees constructs decision trees by randomly selecting features and thresholds to reduce the risk of overfitting and improve model performance. The objective function of the algorithm can be expressed as:

$$\text{Obj}= \sum_{i=1}^{n} L(y_i, \hat{y}_i) + + \sum_{j=1}^{m} \Omega(T_j) \quad (8)$$

4. Environment and results

To identify the optimal combination of hyperparameters for our medal prediction model, we perform an extensive grid search process augmented with cross-validation and early stopping. This systematic approach ensures model robustness by preventing overfitting and enhancing generalization performance across varying data distributions. Figure 2 illustrates the final selected hyperparameter configuration, highlighting the combination that achieves the best predictive performance across multiple evaluation metrics.

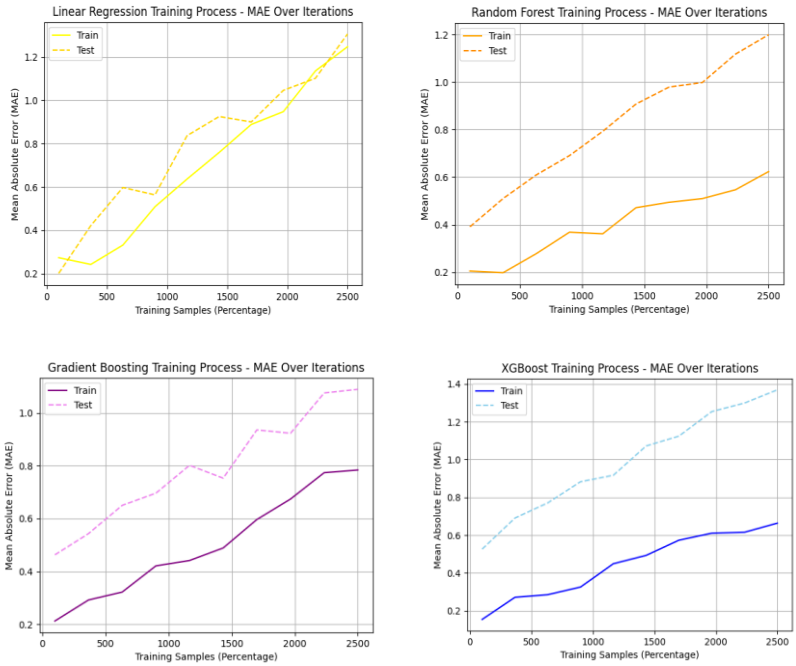

Figure 2: Optimal model parameter combinations

DOI: 10.54254/2753-8818/106/2025.22916

In anticipation of the 2028 Summer Olympics in Los Angeles, significant changes in the competition landscape have been officially announced. Notably, traditional events such as weightlifting and boxing will be excluded, while five new sports—baseball/softball, lacrosse, cricket, squash, and rugby—are to be introduced. These adjustments are expected to influence the medal distribution among participating nations, particularly those with historical strengths in the removed or added disciplines.

Based on our forecasting framework, and assuming a similar set of competing countries as in the 2024 Olympics, we observe a notable reshaping of national medal tallies. The United States, serving as the host nation in 2028, is projected to experience a substantial boost in both gold and total medal counts. In contrast, countries like China and Japan are expected to face a decline in medal performance, primarily due to the exclusion of sports in which they have historically excelled, and the absence of the home field advantage enjoyed by Japan in 2024.

Table 1: National medal ranking prediction table

Country	Gold Medals	Total Medals
United States	51	151
China	31	86
Japan	19	45
Australia	18	52
France	17	62
Netherlands	16	38
Great Britain	14	64
South Korea	11	35

The U.S. is forecasted to dominate the medal table, significantly outperforming its competitors. This outcome can be attributed to both home advantage effects, such as familiar environments and strong crowd support, and the alignment of new events with American sporting culture and infrastructure. For example, sports like baseball/softball and lacrosse are deeply rooted in U.S. athletic traditions, providing a strategic edge in newly introduced categories.

On the other hand, China's predicted decline is primarily associated with the removal of weightlifting and boxing, where Chinese athletes have historically secured multiple podium finishes. Japan, despite strong performance as the host in 2024, is expected to regress due to the loss of home advantage and the realignment of events.

These findings underscore the critical role of environmental and structural factors—including competition restructuring, host nation dynamics, and the inclusion of culturally relevant sports—in shaping Olympic outcomes. Moreover, they highlight the uncertainty and variability inherent in multi-nation sporting events, where external influences can significantly shift competitive balance.

In summary, the 2028 Los Angeles Olympics are anticipated to present a complex interplay of opportunity and challenge for competing nations. Our data-driven forecast emphasizes the strategic importance of understanding policy-level changes and leveraging national strengths to adapt to evolving Olympic structures.

5. Conclusion

This study proposes an integrated, data-driven framework for forecasting Olympic medal distributions, combining machine learning, time-series analysis, and clustering techniques. By incorporating factors such as host country advantage, national competitiveness, and event-specific trends, the model demonstrates strong predictive performance and interpretability. The findings offer

DOI: 10.54254/2753-8818/106/2025.22916

valuable insights into the determinants of Olympic success and provide practical guidance for policymakers and sports development strategists. This work contributes to the growing field of sports analytics by advancing robust, scalable methods for international performance prediction.

References

[1] *Schlembach C, Schmidt S L, Schreyer D, et al. Forecasting the Olympic medal distribution during a pandemic: a socio-economic machine learning model[J]. arXiv preprint arXiv:2012.04378, 2020.*

[2] *Wang Y, Wang J, Huang T Y, et al. STGCN-LSTM for Olympic Medal Prediction: Dynamic Power Modeling and Causal Policy Optimization[J]. arXiv preprint arXiv:2501.17711, 2025.*

[3] *Behrang M A, Assareh E, Assari M R, et al. Using bees algorithm and artificial neural network to forecast world carbon dioxide emission[J]. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2011, 33(19): 1747-1759.*

[4] *Dwivedi Y K, Sharma A, Rana N P, et al. Evolution of artificial intelligence research in Technological Forecasting and Social Change: Research topics, trends, and future directions[J]. Technological Forecasting and Social Change, 2023, 192: 122579.*

[5] *Nagpal P, Gupta K, Verma Y, et al. Paris Olympic (2024) Medal Tally Prediction[C]//International Conference on Data Management, Analytics & Innovation. Singapore: Springer Nature Singapore, 2023: 249-267.*

[6] *He Z, Wang Z. Prediction of olympic medal count for USA based on robust time series model and computer implementation[C]//Third International Conference on Electronic Information Engineering and Data Processing (EIEDP 2024). SPIE, 2024, 13184: 1361-1369.*