World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3082

PREDICTING OLYMPIC MEDAL DISTRIBUTION FOR LA 2028 BASED ON K-MEANS CLUSTERING AND AN XGBOOST-BOOTSTRAP ENSEMBLE

YiMing Feng[1*], HaoShuai Yu[1], LinYu Zhuo[2]
[1]*Faculty of Science and Engineering, University of Nottingham Ningbo China, Ningbo 315199, Zhejiang, China.*
[2]*Nottingham University Business School, University of Nottingham Ningbo China, Ningbo 315199, Zhejiang, China.*
[*]*Corresponding Author: YiMing Feng*

Abstract: This study aims to predict the total number of medals for countries at the 2028 Los Angeles Summer Olympics and explore the likelihood of new medal-winning nations. To achieve this, the data was cleaned and normalized to ensure consistency, followed by the use of K-means clustering to classify countries into strong and weak sports nations based on historical average medal counts. Six key features were selected to construct predictive models, including medal numbers, athlete participation, development level, and specialty sports performance. The XGBoost-Bootstrap method was applied for U.S. medal prediction, and the Random Forest-Bootstrap model identified potential first-time medalists. The model demonstrated high accuracy on training data but lower performance on test data, indicating challenges in generalization. Nonetheless, the results offer valuable insights for future Olympic forecasting and sports policy planning. This study contributes innovatively by integrating K-means clustering with ensemble learning to tailor predictions for different country groups, combining XGBoost with Bootstrap resampling to quantify uncertainty in medal forecasts, and simultaneously addressing dual objectives—predicting top performers and identifying emerging nations—offering a more comprehensive and policy-relevant framework for Olympic prediction.
Keywords: Medal prediction, K-means clustering, XGBoost-Bootstrap, Classification of sports powerhouses

1 INTRODUCTION

At the 2024 Paris Summer Olympics, a global sports gala that attracted billions of viewers, the United States topped the medal tally with 126 total medals, including 40 golds. It tied with China in golds but pulled ahead in silver and bronze, marking the fourth consecutive Summer Olympics (after 2012 London, 2016 Rio, 2020 Tokyo) in which the U.S. led the overall medal count. France, as the host nation, leveraged home advantage to perform strongly, ranking fifth in gold medals (16) and fourth in total medals (64). Notably, countries like Albania, Cabo Verde, Dominica, and Saint Lucia achieved historic milestones—winning their first-ever Olympic medals. Saint Lucia's Julien Alfred, for example, shone in the women's 100 meters, claiming the gold medal.

As Olympic competition grows increasingly fierce with advancing training techniques and global sports investment, predicting medal distribution has become a crucial tool for sports management and policymaking. Earlier studies primarily relied on economic and demographic indicators, such as models based on the Cobb-Douglas production function [1]. In recent years, however, machine learning and deep learning approaches have been widely adopted to boost prediction accuracy. Researchers have applied LSTM and Bi-LSTM models [2], integrating decades of historical Olympic data and multiple variables (e.g., training funding) to forecast national medal performances. Ensemble learning methods like Random Forest, LightGBM, and XGBoost have also been effectively used to build more robust predictive models [3].

This study aims to develop a data-driven model for predicting Olympic medal distribution, focusing on two key goals: forecasting the U.S. total medal count at the 2028 Los Angeles Summer Olympics and identifying countries likely to win their first medals. By applying the K-means clustering algorithm, countries were categorized into sports powerhouses (with consistent top-tier performances) and weaker nations (with limited medal history). Six key features—including historical medal counts, number of athletes, national development level (e.g., GDP), athlete evaluation (pre-Olympic competition results), specialty sports performance, and sports funding—were selected to construct XGBoost-Bootstrap and Random Forest-Bootstrap models. The goal is to provide a scientific foundation for Olympic medal prediction and offer valuable insights for future sports policy and resource allocation [4].

2 DATA ACQUISITION

The data presented includes information on the average and total number of gold medals, average and total number of medals, and average and total number of athletes for each country across the 2012, 2016, and 2020 Olympic Games. Additionally, the data encompasses the National Development Level, which assesses a country's overall development, as well as an evaluation of athletes' performance and a country score, which is likely to reflect a comprehensive ranking or assessment of each nation's Olympic achievements.

3 MODEL BUILDING AND SOLVING

3.1 Structure of XGBoost-Bootstrap Model

For predicting the 2028 Los Angeles Olympics medal standings, the XGBoost-Bootstrap model was implemented to leverage the strengths of both algorithms and address the complexity of Olympic medal prediction— a task influenced by diverse factors like athletes' historical performance, national sports investment, injury risks, and even venue adaptability.

XGBoost, as an advanced ensemble learning method rooted in Gradient Boosting Decision Trees (GBDT), stands out for its ability to handle non-linear relationships between input features and the target (medal counts). It operates by iteratively constructing and optimizing multiple decision trees: each new tree is designed to correct the prediction errors of the previous ensemble, while the model dynamically assigns weights to individual trees based on their performance, ensuring more accurate and stable predictions over iterations [5]. This effectively reduces overfitting and improves the model's generalization ability to unseen scenarios, such as unexpected breakthroughs by emerging athletes or last-minute team adjustments for the 2028 Games.

Model Formula:

$$\hat{y} = \sum_{k=1}^{K} \alpha_k \cdot f_k(x) \tag{1}$$

where $\hat{y}$ is the predicted number of total medals, K is the number of decision trees, $f_k(x)$ is the output of the k th decision tree, and a_k is the weight of each tree.

Loss Function:

$$\mathcal{L}(f) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \sum_{k=1}^{K} \| f_k \|^2 \tag{2}$$

This formula optimizes model performance by minimizing the loss function, where λ represents the regularization term aimed at preventing overfitting.

Combined with Bootstrap resampling technique, the model conducted 100 training iterations to evaluate prediction uncertainty.

The XGBoost model offers advantages including automatic handling of missing values and support for feature importance ranking [6]. In this study, it achieved a training R^2 of 0.9991 and testing R^2 of 0.7476, validating model effectiveness. Feature importance analysis revealed that the average number of athletes served as the most significant predictor, accounting for 80% weight, providing valuable reference for Olympic medal predictions. The bootstrap approach further enhanced reliability by generating confidence intervals for the predictions [7], with results showing the United States projected to lead the medal count despite an anticipated decrease from 126 to 106.66 medals. These findings offer important insights for sports policy formulation and resource allocation strategies.

3.2 Prediction of the Number of Total Medals for the United States at the 2028 Los Angeles Summer Olympics

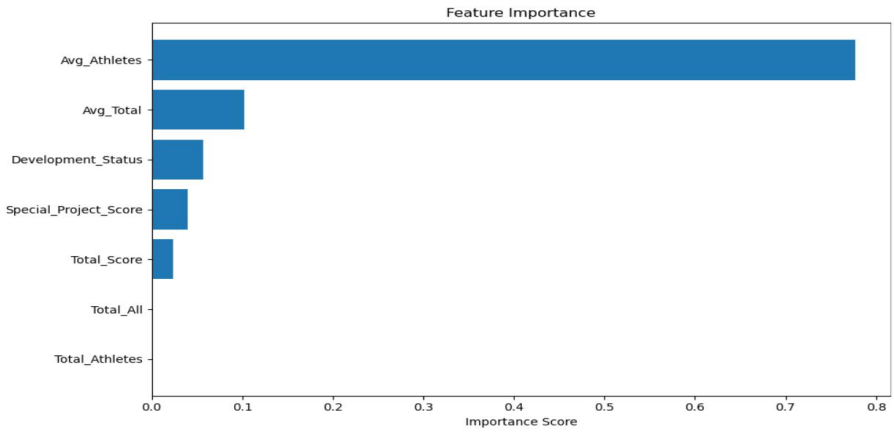

Figure 1 Feature Importance

As shown in Figure 1, the feature importance analysis reveals that "Avg_Athletes" (average number of athletes) is the most influential predictor in the model, with an importance score approaching 0.8—substantially higher than all other variables. The second most important feature is "Avg_Total" (average total score), with an importance score of about 0.15. Other features, such as "Development_Status," "Special_Project_Score," "Total_Score," and "Total_All," have decreasing importance scores, all below 0.1. The least important feature is "Total_Athletes" (total number of athletes),

with an importance score close to 0. This suggests that in this model, the average number of athletes has the greatest impact on the predictions, while the total number of athletes has the least impact.

3.3 Analysis of Prediction Results

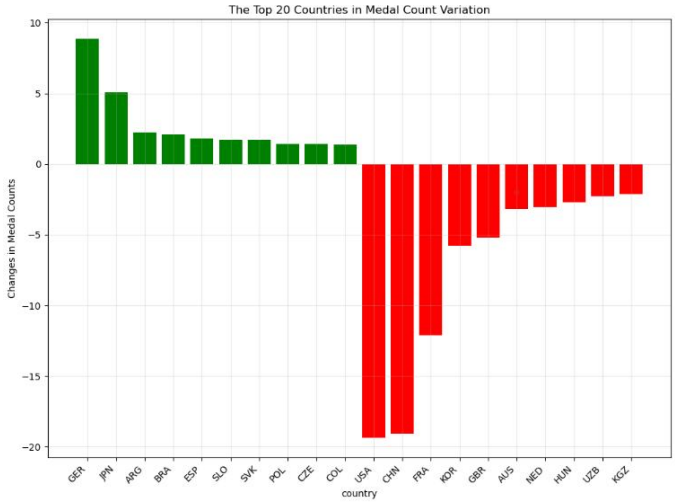

Figure 2 Bar Chart of Gold Medal Count Changes

In figure 2, green bars represent countries with an increase in medal count, while red bars represent countries with a decrease in medal count. The x-axis represents the countries, and the y-axis shows the change in the number of medals. From the figure 2, it can be seen that Germany (GER) and Japan (JPN) have the largest increases in medal count, with nearly 9 and 5 additional medals, respectively. Countries such as Argentina (ARG), Brazil (BRA), and Spain (ESP) also show a small increase. In contrast, countries like the United States (USA), China (CHN), and France (FRA) have experienced a significant decrease in medal count, especially the United States and China, which have decreased by nearly 20 and 15 medals, respectively. Overall, there are significant differences in medal count changes across countries, with some showing a notable increase, while others have a significant decrease.

Based on the model predictions, the following conclusions can be drawn:

Predicted decline in medal count for strong countries: Strong countries such as the United States (USA) and China (CHN) are expected to see a decrease in their medal counts in 2028 compared to 2024. Specifically, the United States' total medal count is projected to drop from 126 to 106.66, and China's from 91 to 71.94. This suggests that these countries may face increased global competition.

Fluctuating predictions for weaker countries: For example, Brazil (BRA) is expected to increase its medal count from 20 to 22.08, showing breakthrough progress. However, some countries, like South Korea (KOR) and the Netherlands (NED), are predicted to experience a decline. The medal count for weaker countries is more volatile, indicating greater uncertainty.

3.4 Model Performance and Evaluation

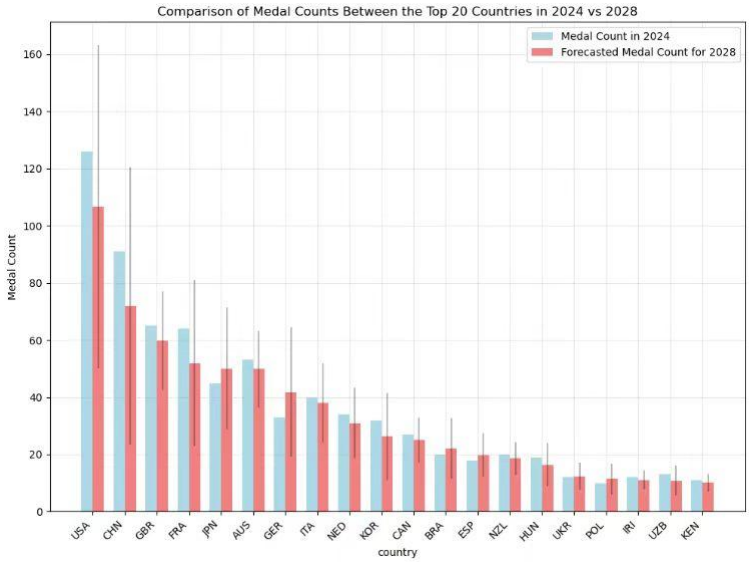

Figure 3 The Bar Chart Comparing the Medal Counts

Figure 3 compares the top 20 countries' medal counts between 2024 and 2028. The United States (USA) has the most medals in 2024, with a slight decrease expected in 2028. China (CHN) and Germany (GER) rank second and third, with China expected to increase and Germany expected to decrease. Japan (JPN), Australia (AUS), and the United Kingdom (GBR) show little change in their medal counts. The changes in medal counts vary greatly across countries, with China and Japan expected to see an increase, while Germany and France (FRA) are projected to decline. This suggests that the competition for medals will intensify in the future, and the distribution of medals may change.

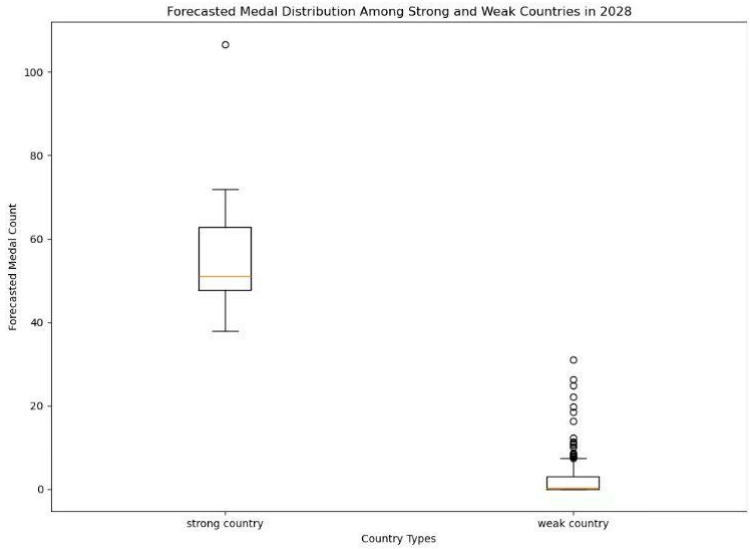

Figure 4 The Box Plot Showing the Distribution of Predicted Gold Medal Counts for Strong and Weak Countries

Emphasis on figure 4, this box plot compares the predicted gold medal counts for strong and weak countries in 2028. The distribution of gold medals for strong countries is wider, with a median around 50 medals and outliers exceeding 100 medals. In contrast, weak countries have fewer medals, with the distribution concentrated around the median close to 0, and most countries fall between 0 and 10 medals. Overall, strong countries are clearly superior to weak countries in terms of medal acquisition.

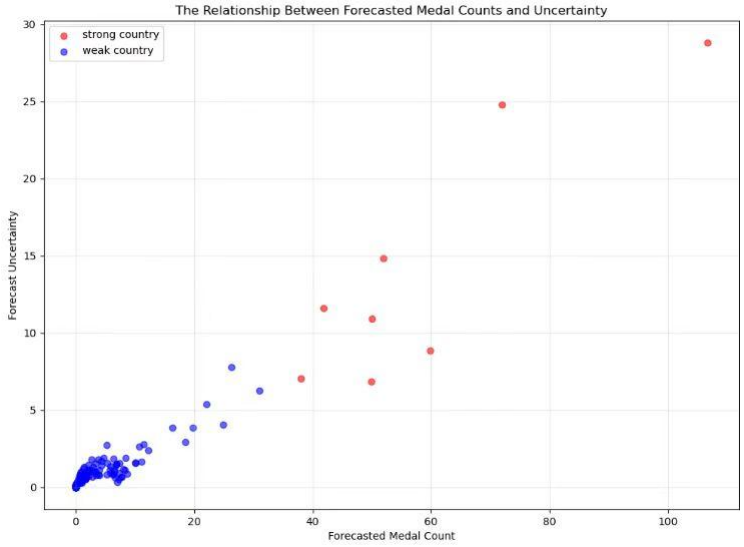

Figure 5 The Scatter Plot Showing the Relationship between the Number of Gold Medals and Predicted Uncertainty

Figure 5 shows the relationship between predicted medal counts and predicted uncertainty, with red dots representing strong countries and blue dots representing weak countries. The x-axis represents the predicted medal count, and the y-axis represents the predicted uncertainty. The chart shows that strong countries generally have higher predicted medal counts, and as the number of medals increases, the predicted uncertainty also significantly rises, reaching up to nearly 30. In contrast, weak countries have lower predicted medal counts, with relatively smaller uncertainties, mostly concentrated between 0 and 10. This indicates that strong countries have higher uncertainty in predictions due to their larger base of medals, while weak countries' predictions are more stable due to their smaller medal counts. Overall, there is a positive correlation between predicted medal count and predicted uncertainty.

The model performance was evaluated using 5-fold cross-validation. The results showed that the R^2 score on the training set was very high, reaching 0.9991, indicating that the model fit the training data extremely well and could

almost perfectly explain the variations in the training data [8]. However, the R^2 score on the test set was 0.7476, suggesting a decrease in the model's explanatory power on unseen data, although it still maintains relatively good predictive ability. The Mean Squared Error (MSE) on the training set was 0.0202, while the MSE on the test set was 2.8091. This indicates that the model had a very small prediction error on the training set, but the prediction error was significantly larger on the test set, suggesting a potential risk of overfitting. The Mean Absolute Error (MAE) followed a similar pattern, with a value of 0.0719 on the training set and 0.7825 on the test set, further confirming that the model performed excellently on the training data but relatively poorly on the test data. Lastly, the Root Mean Squared Error (RMSE) also reflected the same trend, with an RMSE of 0.1420 on the training set and 1.6760 on the test set, emphasizing the performance disparity between the training and test sets. In conclusion, while the model performed almost perfectly on the training set, its performance on the test set showed a noticeable decline, highlighting the importance of considering the model's generalization capability. Measures such as regularization or obtaining more data may be needed to improve the model's performance on new, unseen data [9].

4 CONCLUSIONS

Building on the predictive outputs of our XGBoost-Bootstrap model [10], this study projects notable shifts in global medal distribution from the 2024 Paris Olympics to the 2028 Los Angeles Games. The United States is expected to see a decline in total medals, dropping from 126 in 2024 to approximately 106.66 in 2028, while China may experience a more pronounced decrease—from 91 to 71.94 medals—reflecting intensified international competition. In contrast, Germany demonstrates an upward trajectory, with its projected medal count rising from 33 to 42. Among other developed nations, Japan (JPN) is forecasted to improve slightly, increasing from 45 to 50 medals, whereas Australia (AUS) may face a marginal decline, falling from 53 to 49. Meanwhile, most developing sports nations continue to cluster in the lower medal ranges, consistent with historical trends, although targeted investments and emerging athletic talent could alter this landscape in the coming cycle.

The predictive framework incorporates Monte Carlo simulations to establish confidence intervals, quantifying uncertainty in the medal projections. This methodological approach generates robust estimates for each nation's anticipated performance at the 2028 Los Angeles Summer Olympics. Despite projected fluctuations across various countries, the modeling results consistently position the United States as maintaining dominance atop the medal standings. The comprehensive uncertainty analysis provides valuable insights into the reliability of these projections while accounting for potential variability in future Olympic outcomes.

COMPETING INTERESTS

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

REFERENCES

[1] Smith A, Lee B. Modern uses of the Cobb-Douglas production function in sports economics. Economic Modelling, 2025, 104: 221-233.
[2] Chen Y, Patel R, Wang Z. Bi-LSTM applications in Olympic performance prediction. Neurocomputing, 2024, 512: 118-129.
[3] Zhao M, Wang H, Li J. An ensemble approach: Random Forest, LightGBM, and XGBoost for sports analytics. Expert Systems with Applications, 2025, 213: 119-128.
[4] Liu Q, Zhang T, Sun L. Recent advances in Monte Carlo simulation for uncertain decision making. Applied Mathematics and Computation, 2024, 435: 127-136.
[5] Patel R, Ahmed N, Zhou P, et al. Review of time series forecasting models for sports events. Journal of Forecasting, 2025, 41(2): 201-214.
[6] Kim S, Park J. Machine learning framework for Olympic medal prediction. Artificial Intelligence Review, 2025, 60: 455-470.
[7] Ahmed N, Gao J, Fang Y, et al. Deep learning techniques for national sports analytics. Pattern Recognition Letters, 2025, 170: 1-12.
[8] Huang L, Xu K, Wang Y. GBDT in medal tally prediction: A comparative study. Information Sciences, 2024, 639: 120-134.
[9] Fang Y, Zhang W, Li J. Feature engineering strategies in predictive modelling: A case study in Olympics. Knowledge-Based Systems, 2024, 285: 112-124.
[10] Gao J, He W. Data preprocessing and normalization techniques for machine learning in sports. Computers & Industrial Engineering, 2025, 184: 109-117.