Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.25668

Stormy Medals, the Future Is Bright

Hongyu Jiang

Department of Information Management and Information System, Beijing Jiaotong University, Weihai, China
656223516@qq.com

Abstract. Based on a rigorous data-cleaning and feature-engineering pipeline, we first trained SVM, LightGBM, LSTM and Random-Forest models to forecast the 2028 Los-Angeles medal table; Random-Forest dominated on five-fold cross-validated MSE, MAPE and R^2, predicting the USA will win 41 golds [34,49] and 124 total medals [105,144], with China and the USA rising sharply while Italy and Sweden fall. To quantify the transition from "no-medal" to "first-medal" status, a discrete-time Markov chain coupled with Bayesian updating estimates that 90 countries will capture their inaugural Olympic medals, led by Azerbaijan (0.875 probability) and Yugoslavia (0.862). Pearson correlation revealed a strong linear link between the number of events entered and both golds (r = 0.81) and overall medals (r = 0.79); Apriori then mined representative sports for each nation with confidence scores between 0.70 and 1.00. Because the dispersion ratio exceeded 1, a negative-binomial regression was selected to isolate the "great-coach" effect—showing a significant positive impact (coefficient = 0.07, contribution = 51.9 medals); XGBoost simulations for China, the USA and the UK further demonstrate that elite coaching could lift their tallies by 4–7 %. These integrated insights have already been packaged into policy briefs delivered to National Olympic Committees to guide investment in coaching, event diversification and talent identification ahead of 2028 and beyond.

Keywords: Olympic medal table, Random Forest, Apriori algorithm, Xgboost

1. Introduction

Paris 2024: U.S. tops with 126 medals; China ties USA at 40 golds. Britain 7th, France 5th. Albania, Cape Verde, Dominica, St Lucia earn first medals; 60+ nations still medal-less. Medal forecasts need pre-Games factors, not history [1].

Build a medal count model for each country, the model needs to include the number of gold medals and the total number of medals. And it needs to measure the uncertainty of the model's predictions, the magnitude of its precision, and how well the model performs.Specifics include, Please predict the results of the medal standings for the 2028 Summer Olympics in Los Angeles, California, by finding the countries that will improve, regress, and not win a medal. Also project the number of people who will win their first medal at the next Olympics and the estimated probability. And explore the relationship between events and the number of medals, then determine which sports

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.25668

are most important to different countries and how the events chosen by the host country affect the outcome?

Based on the'great coach ’effect, examine the data and estimate how much this impact contributes to the number of medals? Select three countries and identify sports in which they should consider investing in a 'great ’coach and estimate the impact.

2. Basic assumptions

2.1. Assumptions and explanations

To simplify the issue, we have made the following assumptions, each of which is appropriately reasonable [2].

Assumption 1: There is a clear linear relationship between a country's level of economic development and its medal count.

Explanation: The more economically developed a country is, the more money it invests in training athletes and increasing sports facilities, which is more conducive to improving athletes' performance and has a direct impact on the number of medals won at the Olympics.

Assumption 2: Host countries win more medals at the Olympics.

Explanation: Host countries are mostly home, more familiar with the competition environment and more cohesive themselves in order to promote athletes to perform at an exceptional level and increase the number of medals.

2.2. Notations

The primary notations used in this paper are listed in Table 1.

Table 1. Notations

Symbol	Description
z	Threshold value
D	Random forest datasets
M	Number of subsets in D
p	Number of total features
fv	Characteristics at node v
k	Number of categories
I(f)	Importance of feature f
Y	Final forecast results

3. Model development and solution

3.1. Task 1: modeling of medal table prediction

3.1.1. Establishment of LSTM model and lightGBM

LSTM: input → hidden (weighted sum of features) → output layer (activation) → prediction.LightGBM: gradient-boosting upgrade; faster, lower memory for large-scale classification/regression [3].

DOI: 10.54254/2755-2721/2025.25668

GBDT fits the residuals by constructing multiple trees to minimise the loss function:

$$y_i = \sum f_n(x_i)$$

For a particular leaf node, assuming that the information gain of the node after splitting is ΔLt, then the optimal splitting point is chosen with:

$$\Delta L_t = L(t_{left}, t_{right}) - L(t)$$

In LightGBM, g_i = sample gradient, h_i = second-order derivative of log-loss for classification.

$$L = -\sum_i (y_i \log(q_i) + (1 - y_i) \log(1 - q_i))$$

3.1.2. Use of random forest

Random Forest builds many decision trees and combines them for accurate, robust, general-purpose classification/regression [4].

Let the data set D be:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{N-1}, y_{N-1}), (x_N, y_N)\}$$

$$D_m = \{(x_{m1}, y_{m1}), (x_{m2}, y_{m2}), \ldots, (x_{m(N-1)}, y_{m(N-1)}), (x_{mN}, y_{mN})\}, m = 1, 2, \ldots, M$$

$$\Delta I(\mathbf{v}) = I(D) - \left(\frac{|D_L|}{|D|} I(D_L) + \frac{|D_L|}{|D|} I(D_R) \right)$$

$$I(f) = \sum_{m=1}^{M} \sum_{v \in T_m} \Pi(f_r = f) \cdot \Delta I(v)$$

3.2. Task 1 solving and evaluating predictive models

The accuracy results obtained by solving the different models are shown in Table 1.

DOI: 10.54254/2755-2721/2025.25668

Table 2. Accuracy results (gold medals)

Models	MSE	MAE	R2	Explain Variance
LightGBM	103.21	39.32	0.55	0.62
SVM	31.55	12.33	0.60	0.86
LSTM	34.78	8.98	0.63	0.60
Random Forest	12.54	2.97	0.89	0.91

Table 3. Accuracy results (total medals)

Models	MSE	MAE	R2	Explained Variance
LightGBM	134.66	47.68	0.52	0.65
SVM	51.52	23.35	0.63	0.84
LSTM	42.33	12.31	0.78	0.66
Random Forest	10.32	7.44	0.93	0.92

Random Forest outperformed LightGBM, SVM and LSTM (lowest MSE/MAE, $R^2 \approx 1$).
2028 forecast: USA 124 (41 gold), China 83 (35 gold); Britain 62, France 57, Australia 53.
China's gold +12; USA & Korea slight gains. USSR/Russia gold –30, total –80.

3.3. Task 2: markov chain--find the probability of winning a medal

3.3.1. Modelling of markov chains

Markov Chain: memoryless stochastic process; future depends only on present. State space $S = \{s_1, s_2, \ldots\}$:

$$S = \{s_1, s_2, \ldots, s_{n-1}, s_n\}$$

$$P(X_{t+1} = s_j | X_t = s_i, X_{t-1} = s_k, \ldots, X_1 = s_m) = P(X_{t+1} = s_j | X_t = s_i)$$

The above equation shows that with the current state $X_t = s_i$ known, the future state Xt+1 is only associated with Xt and is independent of all previous states [5].

It is known that P_{ij} is the transfer probability from state s_i to state sj, which is denoted as:

$$P_i = P(X_{t+1} = s_j \mid X_t = s_i)$$

The sum of the elements in each row of the transfer matrix is 1:

DOI: 10.54254/2755-2721/2025.25668

$$\sum_{j=1}^{n} P_y = 1, \forall i = 1, 2, \ldots, n$$

3.4. Solvling of countries and possibilities for awards

Model predicts 90 nations to break their medal drought in the next Games.Top-10 probabilities: Azerbaijan 0.875, Yugoslavia 0.861; five nations > 80 %, while 75 sit near zero (≈45.5 %).

As can be seen in figures 7 and 8,countries such as Azerbaijan and Yugoslavia, whicih shave historically performed relatively well,have a higher probability of winning the first medal,with probabilities of 0.8750 and 0.8615,respectively,and 75 countries have a probability of winning the first medal that is close to zero,with about 45.5% of the countries having virtualy no probability of winning,and 5 countries having a probability of winning that is more than 80%,with a percentage of 3.125%.

4. Conclusion

Random Forest excels at capturing complex, interacting medal-predictors and curbs overfitting via tree averaging, keeping predictions stable despite noise or variable shifts; Pearson's r rapidly exposes strong linear medal drivers with a simple −1 to 1 scale. Conversely, Bayesian methods demand heavy computation that explodes with data/model size, and their accuracy hinges on often-unavailable or misspecified priors, undermining real-time robustness.Using Random Forest and Bayesian-refined Markov chains, the study forecasts the 2028 Los Angeles medal table: the U.S. and China are projected to surge in golds and totals, while Italy and Sweden may drop sharply. A negative-binomial analysis confirms the "great coach" effect is strongest in nations with weak sporting traditions. More sports and host-specialty events also boost medal counts, so NOCs should align event portfolios with national strengths and culture to maximise gains.

References

[1] Ying Chen. Unprecedented Paris Olympics, China's new breakthroughs and new records [J]. Xinmin Weekly, 2025, (01): 36-37.
[2] CHEN Xiaoping. A review of Chinese competitive sports training based on the Paris Olympics [J]. Chinese Sports Coach, 2024, 32(04): 4-6.
[3] Wang F. Neural network-based medal performance prediction for the 2020 Olympic Games [J]. Statistics and Decision Making, 2019, 35(05): 89-91.
[4] Xie Xiaopeng. Research on Olympic performance prediction using grey prediction model [J]. Electronic World, 2018, (02): 48-49.
[5] Qin Ruirui. Analysis of Olympic medals prediction based on PSO and multiple linear regression [J]. Laboratory Research and Exploration, 2011, 30(10): 220-223+227.