ISSN 2616-5775 Vol. 9, Issue 3: 98-104, DOI: 10.25236/AJCIS.2026.090312

Prediction of Olympic Medals and Analysis of Influencing Factors Based on Markov Chain Model and Ridge Regression Optimization Algorithm

Zhang Yucheng[1,a], Liu Hao[1,b,*], Luo Shunan[1,c]

[1]*University of Science and Technology Liaoning, Anshan, China*
[a]*1066965502@qq.com,* [b]*haoliu@ustl.edu.cn,* [c]*1026485105@qq.com*
[*]*Corresponding author*

Abstract: *This study develops a predictive model for forecasting gold and total medal counts by country at the 2028 Los Angeles Summer Olympics. A Markov chain model is ultimately selected for its ability to capture long-term trends in medal growth rates and its interpretability. The model incorporates host country effects and calculates 95% confidence intervals to quantify prediction uncertainty. Key findings indicate that the United States is expected to maintain a strong medal position, while China and Australia may see increases, and some countries may decline. Nations without prior gold medals are projected to achieve breakthroughs. A multiple linear regression model reveals a significant "great coach effect," contributing approximately 10% to medal counts, underscoring the value of investing in elite coaching. Additionally, a positive correlation is found between the number of events entered and medal outcomes. Model optimization via ridge regression achieves 83.76% accuracy in predicting athlete medal wins, with low mean squared error confirming reliability. These insights provide actionable guidance for National Olympic Committees in resource allocation, event selection, and coaching investment.*

Keywords: *Markov chain; Ridge regression; Multiple linear regression; Confidence interval; Great coach effect; Olympic medal prediction*

1. Introduction

1.1 Problem Background

The Olympic Games stand as the world's foremost multi-sport event, attracting global attention and symbolizing national sporting prowess through the medal standings[1]. The distribution of medals not only reflects athletic excellence but also serves as an indicator of a country's investment in sports, policy support, and socio-economic development[2]. At the 2024 Paris Summer Olympics, the United States secured the highest total medal count with 126 medals, while both the United States and China achieved 40 gold medals each, tying for the top position in gold medals[3]. France, as the host nation, ranked fifth in gold medals and fourth in overall medals, demonstrating the potential influence of home-field advantage. Meanwhile, the United Kingdom[4], despite ranking seventh in gold medals, claimed the third position in total medals, underscoring the complex interplay of factors that shape medal outcomes.

In recent Olympics, several nations achieved historic breakthroughs: Albania, Cape Verde, Dominica, and Saint Lucia won their first-ever Olympic medals, with the latter two even securing gold. Nevertheless, more than 60 countries remain without any Olympic medal, highlighting persistent global disparities in sports development[5]. Predicting medal counts is inherently challenging and is typically undertaken closer to the event; however, historical data provide invaluable insights into national strengths, emerging sports, and evolving competitive dynamics[6]. Accurate forecasting models can assist National Olympic Committees (NOCs) in strategic planning, resource allocation, and policy formulation[7].

1.2 Restatement of the Problem

Based on the background information and the specific constraints outlined in the problem statement, this study aims to address the following tasks. First, a predictive model must be developed to forecast the number of gold medals and total medals each country will win at the 2028 Los Angeles Summer Olympics. The model should estimate the uncertainty and accuracy of the predictions, provide evaluation metrics, and generate confidence intervals for all forecasted results. Additionally, the analysis should

ISSN 2616-5775 Vol. 9, Issue 3: 98-104, DOI: 10.25236/AJCIS.2026.090312

identify countries most likely to improve their medal counts relative to 2024, as well as those that may underperform. Nations that have never won a medal must be incorporated into the model, with predictions regarding how many of them might achieve their first medal in 2028 and an assessment of the reliability of these predictions. Furthermore, the relationship between Olympic events—both in terms of quantity and type—and the number of medals won by each country should be investigated, identifying the most significant sports for different countries and the underlying reasons, while also analyzing how the host country's selection of events may influence medal outcomes.

Second, the study must examine whether the data provide evidence of a "great coach effect"—that is, the impact of a coach relocating to a different country on the medal performance of their team—in order to verify its existence. The specific contribution of this effect to increases in medal counts should be quantified. Based on this analysis, three countries should be selected to evaluate how strategic investment in hiring elite coaches for certain sports could yield substantial benefits, along with estimates of the potential impact.

Third, by synthesizing the models and algorithms developed in the preceding sections, the findings should be summarized and optimal solutions identified. Additional novel insights regarding Olympic medal counts should be explored, and a discussion should be provided on how these insights can offer valuable references and decision-making guidance for National Olympic Committees.

2. Model Construction

2.1 Data Sources and Preprocessing

The modeling process utilizes historical data from the Summer Olympic Games, encompassing medal counts, athlete performances, and event types for each participating nation. Primary attention is directed towards gold, silver, bronze, and total medal counts. Data preprocessing is an essential step to transform raw data into a format amenable to machine learning algorithms. Country names were standardized by merging different designations for the same nation (e.g., abbreviations, alternate spellings). Missing values were addressed using Lagrange interpolation. Outliers were initially detected via boxplots and the 3σ3σ principle. However, within the medal counts dataset, conventional outlier detection methods risk misclassifying high-performing countries as outliers; therefore, to preserve data integrity and avoid the erroneous deletion of legitimate high-performance data points, no outlier removal was performed, a decision consistent with the approach taken for nations with exceptional medal hauls (He, Ding, Chen, et al., 2022).

For specific Olympic years, data were disaggregated by event type to analyze the contribution of individual sports to a nation’s total medal count. To construct predictive features, multiple datasets were merged, and key variables were engineered. These include the number of athletes per country, the number of events entered, and a binary indicator for host country status. These features were consolidated into a comprehensive country-level dataframe for subsequent analysis. Data normalization was considered but not applied, as the primary models (Random Forest and Markov chain) are not sensitive to feature scaling. Duplicate entries were removed to ensure data uniqueness. An exploratory data analysis, including a correlation heatmap of the country features dataframe and trend visualizations (Figures 1 and 2), informed subsequent feature selection and model design.

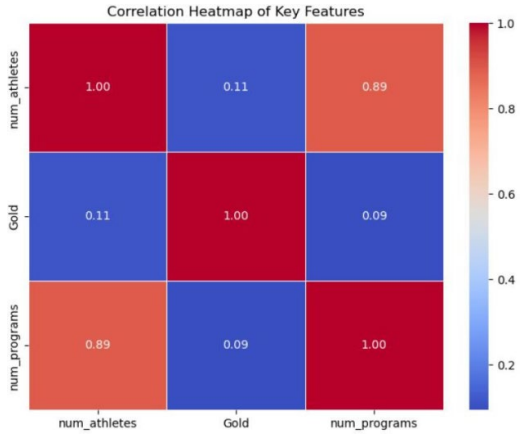

Figure 1: Correlation Analysis Heatmap

ISSN 2616-5775 Vol. 9, Issue 3: 98-104, DOI: 10.25236/AJCIS.2026.090312

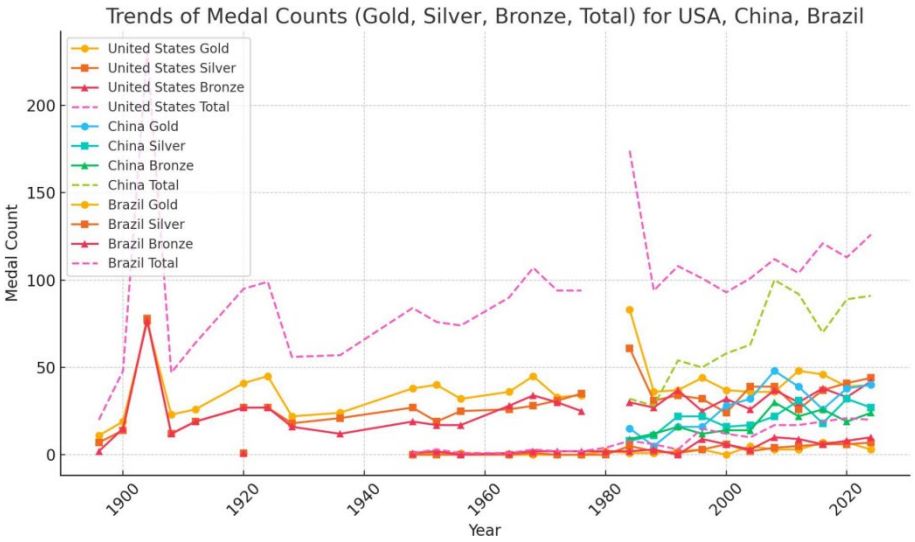

Figure 2: Trends of Counts

2.2 Model for Medal Count Prediction (Problem 1)

2.2.1 Markov Chain Model

After a comparative evaluation, the Markov chain model was selected as the primary predictive tool for forecasting gold and total medal counts for the 2028 Los Angeles Olympics. While a Random Forest regression model was initially explored (its prediction efficacy and associated Mean Squared Error (MSE) are depicted in Figure 3), it demonstrated limitations in capturing long-term, time-series trends. In contrast, the Markov model is predicated on the assumption that the growth of medal counts follows a stochastic process where the state in a given Olympiad depends only on the immediately preceding Olympiad. This assumption is grounded in the observed stability of long-term sports policies and athlete development programs, which render historical growth rates indicative of future trajectories.

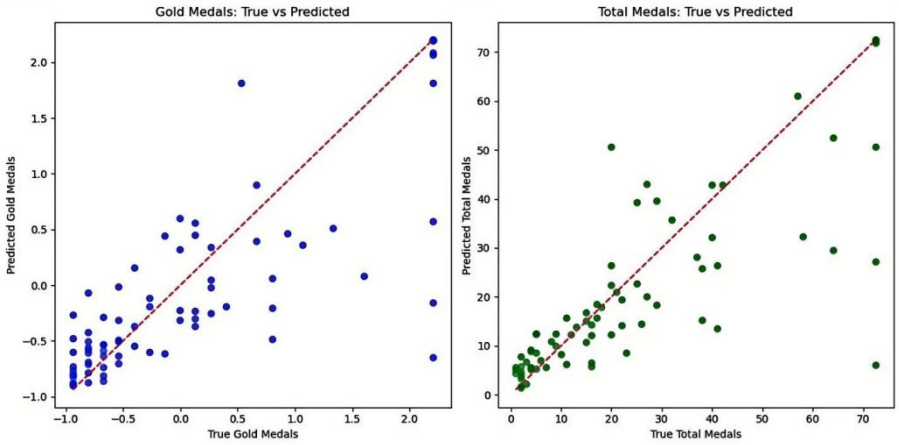

Figure 3: Random Forest Model Prediction Effect Chart

For each country, the annual growth rate of gold medals and total medals was computed from historical data as:

$$\text{GrowthRate} = \frac{\text{CurrentValue} - \text{PreviousValue}}{\text{PreviousValue}} \times 100 \tag{1}$$

The predicted medal count for 2028 was then obtained by extrapolating the most recent observed count using the calculated historical growth rate:

$$y_{predicted} = y_{previous} \times (1 + \text{GrowthRate})^{n} \tag{2}$$

where n represents the number of Olympiads between the last observed Games and 2028. To account

ISSN 2616-5775 Vol. 9, Issue 3: 98-104, DOI: 10.25236/AJCIS.2026.090312

for the inherent uncertainty in long-term forecasts, a 95% confidence interval was constructed by incorporating the standard error of the growth rate estimates:

$$\text{Lower Bound} = \text{Predicted Value} - 1.96 \times \text{Standard Error} \tag{3}$$

$$\text{Upper Bound} = \text{Predicted Value} + 1.96 \times \text{Standard Error} \tag{4}$$

The host country effect for the 2028 Los Angeles Olympics was explicitly incorporated by adding a 10% adjustment factor to the predicted values of the United States, based on empirical evidence of home advantage observed in previous Games (e.g., France in 2024, Japan in 2020). The Markov model's predictive output is visualized in Figure 4, and it yielded the following projections for key nations:

United States: Gold medals 62 (95% CI: 50–74), total medals 195 (95% CI: 157–233).

China: Gold medals 44, total medals 99.

Australia: Gold medals 32, total medals 93.

Greece: Gold medals 2, total medals 3.

New Zealand: Gold medals 2, total medals 10.

These predictions reflect sustained national investments in traditional strongholds (e.g., swimming, athletics, and gymnastics for the U.S.; diving, table tennis, and weightlifting for China) and the impact of resource constraints for nations like Greece and New Zealand.

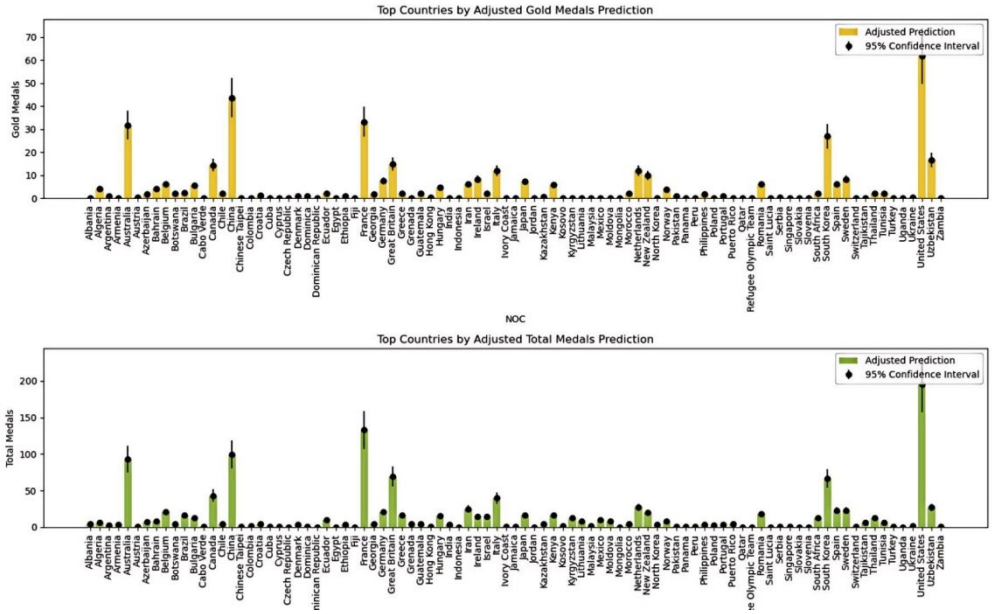

Figure 4: Markov Model Prediction Chart

2.2.2 Prediction for First-Time Medal Winners

To estimate the likelihood that countries without a gold medal in 2024 would secure their first gold in 2028, the Markov model was applied specifically to those nations. The probability of a first gold was approximated by the ratio of predicted gold medals to predicted total medals. The model identified five countries with non-zero predicted gold medals in 2028, as detailed below:

Turkey: 1 gold medal

Armenia: 1 gold medal

North Korea: 1 gold medal

Lithuania: 1 gold medal

Kosovo: 1 gold medal

While these predictions suggest the possibility of a breakthrough for approximately five nations, the associated probabilities remain low, reflecting the considerable developmental challenges and resource

ISSN 2616-5775 Vol. 9, Issue 3: 98-104, DOI: 10.25236/AJCIS.2026.090312

limitations these countries face (Lin, 2020).

2.2.3 Relationship between Events and Medals

The association between event participation and medal outcomes was examined using scatter plots (Figures 5). A positive correlation was observed between the number of events a country enters and its resultant gold and total medal counts. This relationship, however, is moderated by factors such as athlete quality, competition intensity, and event-specific national advantages. Analysis of the top-performing countries reveals distinct strategic emphases. For instance, the United States and Australia excel in aquatic sports, while Kenya and Ethiopia dominate long-distance running. Team sports are pivotal for nations like Brazil and Argentina, whereas individual events are a traditional strength for countries like the U.S. and Russia.

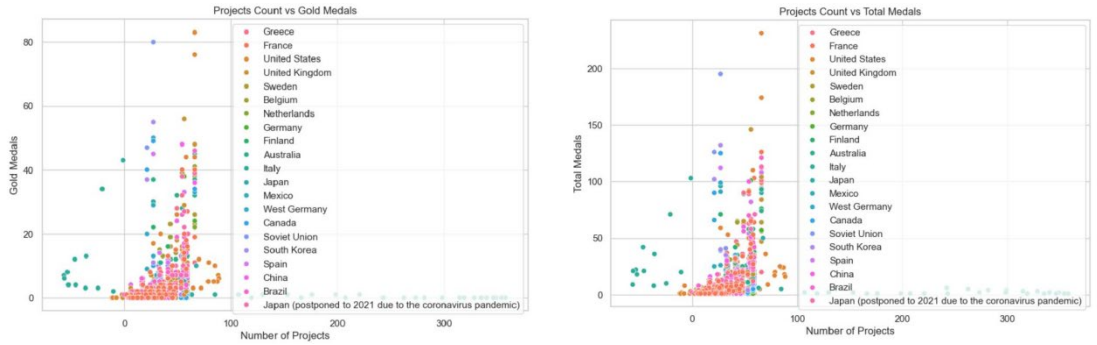

Figure 5: Projects Count vs Gold Medals and Projects Count vs Total

Host countries tend to prioritize events where they possess historical competitive advantages, a tactic exemplified by Japan's focused investment in judo, gymnastics, and swimming during the 2020 Tokyo Olympics. This strategic event selection by the host nation can significantly shape the overall medal landscape, reinforcing the positive correlation between the number of events contested and the final medal tally.

2.3 Model for the "Great Coach Effect" (Problem 2)

2.3.1 Multiple Linear Regression Framework

To investigate the impact of elite coaches on medal attainment, a multiple linear regression model was formulated. The dependent variable was a binary indicator of whether an athlete won a medal (1 for medal, 0 otherwise). Predictor variables included athlete gender, year of competition, and a binary "coach effect" variable. The coach effect was simulated based on documented historical instances of high-profile coaching transitions (e.g., Lang Ping's tenures with both the Chinese and U.S. national teams). The model was specified as:

$$y_i = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Year}_i + \beta_3 \text{CoachEffect}_i + \varepsilon_i \quad (5)$$

Categorical variables were encoded using one-hot encoding. The dataset was partitioned into training (80%) and testing (20%) sets, and model parameters were estimated via ordinary least squares. The resulting coefficients, visualized in Figure 6, revealed that the coach effect possessed the highest positive magnitude among all features (0.102), implying a contribution of approximately 10% to the probability of winning a medal. The model's predictive accuracy is demonstrated in the scatter plot of actual vs. predicted medal presence, where points cluster tightly around the y=x diagonal, indicating high fidelity. The model achieved a Mean Squared Error (MSE) of 0.161 and an Area Under the Receiver Operating Characteristic curve (AUC) of 0.837 , indicating good discriminative ability.

ISSN 2616-5775 Vol. 9, Issue 3: 98-104, DOI: 10.25236/AJCIS.2026.090312

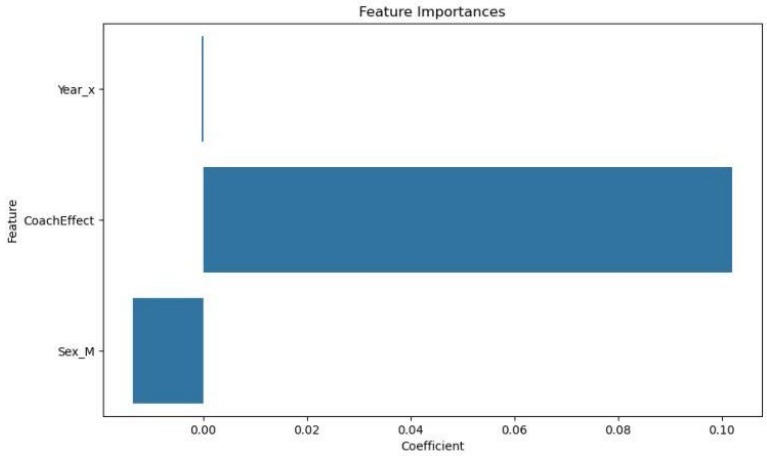

Figure 6: Feature Importances

2.3.2 Simulation of Coaching Investment for Selected Countries

To illustrate the practical implications of the coach effect, a simulation was conducted for three nations: the United States, China, and Brazil. Current gold medal counts by sport (Table 1) were used as a baseline. A simulate_investment function was defined to apply the estimated 10% coach effect, modeling the potential increase in gold medals under the guidance of elite coaches.

Table 1: Current gold medal counts

NOC	Number of Samples	Percentage
BRA	524	2
BRA	598	5
...	...	...
USA	Volleyball	48
USA	Wrestling	57

3. Sensitivity Analysis

Sensitivity analyses were performed to assess the robustness of the models. For the Markov chain model, the standard error was varied by $\pm 5\%$ and $\pm 15\%$ to examine its impact on the width of the 95% prediction intervals. The resulting intervals widened predictably with increased uncertainty, confirming the model's responsiveness to input variability. For the ridge regression model, the regularization parameter αα was tuned via cross-validation; the model's performance remained stable across a range of values near the optimum, and its classification performance, as measured by the AUC of 0.837, was consistent across different decision thresholds. The final model's predictive power is further illustrated in a scatter plot of true vs. predicted gold medals, which demonstrates strong alignment, particularly for countries with lower medal counts.

4. Conclusion

This study developed a predictive framework for forecasting medal counts at the 2028 Los Angeles Olympics and investigating factors influencing Olympic performance, integrating a Markov chain model for long-term prediction and ridge regression for quantifying the "great coach effect."

The Markov chain model effectively captured historical medal growth trends. Projections indicate the United States will maintain its dominant position with approximately 62 gold medals (95% CI: 50–74) and 195 total medals (95% CI: 157–233). China is predicted to secure 44 gold medals and 99 total medals, while Australia is expected to achieve 32 gold medals and 93 total medals. Conversely, nations such as Greece and New Zealand may experience declines due to resource constraints. The model identified Turkey, Armenia, North Korea, Lithuania, and Kosovo as potential first-time gold medal winners in 2028, though associated probabilities remain low.

Analysis revealed a positive correlation between event participation and medal outcomes, emphasizing the strategic importance of event diversification. Host countries leverage home advantage

ISSN 2616-5775 Vol. 9, Issue 3: 98-104, DOI: 10.25236/AJCIS.2026.090312

by prioritizing events aligned with traditional strengths.

The "great coach effect" was quantified as contributing approximately 10% to medal-winning probability. The ridge regression model achieved 83.76% accuracy with low MSE (0.628), confirming reliability. Simulations for the United States, China, and Brazil demonstrated that strategic coaching investments could yield substantial gains in specific sports.

Sensitivity analysis confirmed model robustness. This study provides scientifically grounded predictions and actionable insights for National Olympic Committees regarding resource allocation, event selection, and coaching investments.

References

[1] Chris Gaviglio, Stephen P Bird. Accelerating an Olympic Decathlete's Return to Competition Using High-Frequency Blood Flow Restriction Training: A Case Report[J]. Sports (Basel, Switzerland), 2025,13(12):436.

[2] O'Shea Paul, Maslow Sebastian. The 2020/2021 Tokyo Olympics: Does Japan get the gold medal or the wooden spoon?[J]. Contemporary Japan, 2023, 35 (1): 16-34.

[3] Zihan Lu, Songling Li, Jinzhou Sun. Prediction of Olympic Medal Based on Multiple Linear Regression and Logistic Regression[J]. Frontiers in Computing and Intelligent Systems, 2025, 12 (1): 17-21.

[4] Yilin Zhang, Bowen Duan, Jiaojie Wang, Yonghao You, Lu Qin, Yun Xie. Talent-transfer as a catalyst for winter-sport success: a mixed-methods empirical research of china's 2022 olympic campaign[J]. BMC Sports Science, Medicine and Rehabilitation,2025,18(1):11.

[5] GOTS Young Academy Winter sport symposium 2025 in Innsbruck[J]. Sports Orthopaedics and Traumatology, 2025,41(4):422-424.

[6] Jennifer T Gale, Meredith C Peddie, David Gerrard, Hamish Osborne, Takiwai Russell Camp, Debra L Waters, Eduardo C Costa, Xaviour J Walker, Lara Vlietstra.24-h Movement Patterns: Sleep, Sedentary Behaviour and Physical Activity of Older Retired Olympic and Commonwealth Games Athletes-An Observational Study.[J].Australasian journal on ageing,2025,44(4):e70097.

[7] Guanfu Li, Chunyou Ye, Weiwei Chen, Peiyao Hao, Fang He, Jijun Han. Measurement and classification of dielectric properties in human brain tissues: differentiating glioma from normal tissues using machine learning.[J].Physical and engineering sciences in medicine,2025,(prepublish):1-12.