Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9042

Research on Olympic Medals Prediction by SVM and Random Forest

Jianzhang Li
School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
Major: Applied Mathematics
Email: Jianzhang.Li23@student.xjtlu.edu.cn

How to cite this paper: Li, J. Z. (2026). Research on Olympic Medals Prediction by SVM and Random Forest. Journal of Computer Science and Frontier Technologies, 2(2), 83-91. ISSN Print: 3104-4204; ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9042
Published: 2026-02-02

Abstract

The goal of this article is to predict the likelihood of countries without Olympic gold medals winning their first in 2028. Our study create binary classification labels and countries that have never won gold are identified and relevant features are extracted. SVM is employed for classification, with an AUC score close to 1, indicating high accuracy. Eventually, the relationship between event selection and medal counts is analyzed. The data is preprocessed and event types are converted into categorical variables. Random Forest regression is used, revealing that host country event selection affects medal performance but has minimal impact on overall rankings. The model's performance is validated using MAE, MSE, and other metrics.

Keywords

SVM; Random Forest, Prediction

1. Introduction

The Olympic Games stand as one of the most significant global sporting events, drawing the attention of millions from diverse cultural backgrounds. Audiences fixate not only on individual competitions but also on the overarching medal standings, particularly the distribution of gold medals and total medal counts among nations. This keen interest gives rise to vibrant discussions around potential last-minute upsets and historic breakthroughs in medal tallies.

The academic community has engaged in extensive research concerning the determinants of Olympic medal counts. Bernard and Busse examined the relationship between Olympic medal outcomes and various social-economic factors. Their findings indicated that host countries often gain an advantage, highlighting how resource-rich nations—historically exemplified by the Soviet Union—tend to dominate the medal tables[1]. Parallel to this, Guo and Zhao implemented the GM(1,1) grey theory model to forecast gold medal standings for the 2016 Rio

Olympics, relying on data accrued from the previous six Olympic Games. Their innovative model predicted that China would secure 55 gold medals, with the United Kingdom and the United States attaining 48 and 41 golds, respectively[2].

Further advancing the discourse, Shiyu Wang synthesized the methodologies proposed by Bernard, Busse, Guo, and Zhao, incorporating multivariate nonlinear regression alongside backpropagation neural networks to enhance the accuracy of medal count predictions[3]. This hybrid approach produced predictions that were closely aligned with traditional regression outcomes. Additionally, Kearney explored the use of various economic and social indicators to better predict Olympic success, suggesting a multifaceted approach to understanding medal outcomes[4]. Coughlan investigated the impact of sporting infrastructure on Olympic performance, finding a strong correlation between investment in facilities and medal counts[5]. Such studies underscore the complex interplay of factors that contribute to a nation's Olympic success and emphasize the necessity of comprehensive modeling techniques to capture these dynamics.

With the continuous advancement of prediction methods, particularly the integration of traditional statistical techniques and modern machine learning technologies, a promising path has emerged for countries aiming for success in the Olympic Games. Based on this, this study explores the possibility of countries without an Olympic gold medal winning their first gold in 2028. Support Vector Machine (SVM) was used to predict the likelihood of these countries winning their first gold medal in 2028. Additionally, Random Forest regression was employed to analyze the relationship between the host country's event selection and medal performance. The results suggest that event selection influences medal performance but has a limited impact on overall ranking. This study provides methodological support for Olympic medal prediction and strategic planning.

2. Correlation theory

Our study innovatively applies supporting vector machine and random forest theory to predict the likelihood of countries without Olympic gold medals winning their first in 2028.

2.1. Supporting Vector Machine Theory

Support Vector Machine is a powerful supervised learning algorithm primarily used for classification tasks. The core idea behind SVM is to create a hyperplane (or a set of hyperplanes) in a high-dimensional space that effectively separates data points belonging to different classes.

At its most fundamental level, SVM aims to find a linear hyperplane that divides the data into two distinct classes with the maximum margin. The margin is defined as the distance between the hyperplane and the nearest data point from either class. These closest points are referred to as support vectors. SVM seeks to maximize this margin, which enhances the model's ability to generalize to unseen data.

In many researches, data is not perfectly separable. To handle such situations, SVM employs a soft margin approach, allowing some misclassifications to achieve better generalization. This is controlled by a regularization parameter, which balances the trade-off between maximizing the margin and minimizing the classification error.
SVM performs well in high-dimensional spaces, making it suitable for tasks such as image classification and bioinformatics. By maximizing the margin, SVM is less prone to overfitting, particularly in cases where the number of dimensions exceeds the number of samples. The utilization of kernel functions offers flexibility, allowing SVM to classify complex, non-linear relationships efficiently. However, Training SVMs can be computationally expensive, particularly for large datasets, as it involves solving a quadratic optimization problem. The performance of SVM heavily relies on the selection of the kernel and its parameters, making it necessary to perform careful tuning.

2.2. Random Forest Theory

Random Forest is an ensemble learning method that constructs multiple decision trees during training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of individual trees. It is known for its robustness and accuracy.
A single decision tree is a flowchart-like structure that makes decisions based on answering a series of questions about the input features. Decision trees partition the input space recursively, creating a model that is easy to visualize, but they may overfit the training data.Random Forest employs a technique called bagging, where multiple subsets of the training data are created by bootstrapping (sampling with replacement). Each decision tree is trained on a different subset of the data, introducing diversity into the model.
In addition to bagging, Random Forest adds another layer of randomness by selecting a random subset of features for each split in the decision trees. This further increases the likelihood that the trees will differ from one another, improving the overall robustness of the model. After the decision trees are trained, predictions are made by aggregating the outputs from all the individual trees. For classification, this is achieved through majority voting, while for regression, the average of the outputs is taken.
Random Forest often achieves high accuracy due to the collective decision-making of multiple trees, reducing the risk of overfitting. The inherent random sampling and feature selection help Random Forest remain robust even in the presence of noisy data. Random Forest provides useful metrics such as feature importance, allowing practitioners to understand the significance of different input features in the prediction process.

3. Experiment

Our study employs Support Vector Machine (SVM) and Random Forest models to analyze Olympic medal data, focusing on predicting the chances of countries that have never won a gold medal achieving this milestone in the upcoming 2028 Olympics. The SVM model identified eight countries with low predicted probabilities of winning gold, while the Random Forest model examined various factors affecting medal counts, revealing significant correlations between the number of events and both gold and total medal counts. Overall, the analysis highlights the complexities influencing Olympic success and the effectiveness of machine learning methods in forecasting outcomes based on historical data.

3.1. Abbreviations and Acronyms

TABLE 1. NOTATIONS.

Symbol	Description
$\Omega(f_k)$	Regularization term
Di	The subset obtained after splitting by feature
H(D)	The entropy of the dataset D
T	The number of decision trees
$\hat{y}_t$	The prediction value of the decision tree sample
Xit	Features related to athletes and countries
Cit	Factors influencing the coach

3.2. Equations

In this phase, our study will use Support Vector Machine (SVM) for classification prediction. The core idea of SVM is to find an optimal hyperplane (or, in the case of non-linearity, find an optimal high-dimensional space mapping) in the feature space that separates the data points into different categories [8][9]. Since the data is complex and nonlinear, this research can choose a complex kernel function, such as the Radial Basis Function (RBF) Kernel, to map the data into a higher-dimensional space to aid in classification. The specific formula and method for the RBF Kernel are as follows:

- linear kernel：

$$K(x_i, x_j) = x_i^T x_j, \tag{1}$$

- Gaussian Radial Basis Function (RBF) kernel：

$$K(x_i, x_j) = \exp(-\frac{\| x_i - x_j \|^2}{2\sigma^2}), \tag{2}$$

- polynomial kernel：

$$K(x_i, x_j) = (x_i^T x_j + c)^2, \tag{3}$$

After using the kernel function, the objective function of the optimization problem becomes:

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \tag{4}$$

Next, this paper use the training dataset to train the SVM model, adjusting model parameters (such as the C parameter, kernel function parameters, etc.) to maximize the classification performance.

Finally, with the trained SVM model, our study predict the countries that have not won medals, and determine whether each country is likely to win a medal in the next Olympic Games. Based on the model’s predictions, the probability of each country winning a medal can be calculated, and specific predictions can be made.

The random forest model performs regression predictions by integrating multiple decision trees, effectively handling high-dimensional data and complex nonlinear relationships. This approach helps avoid the overfitting issues that may arise with a single model, making it particularly suitable for analyzing the impact of event setups and the influence of athletes from various countries on medal counts [10]. In the construction of decision trees, the split at each node is typically determined by either information gain (used in classification tasks) or mean squared error (MSE) (used in regression tasks). In classification tasks, our model evaluate the selection of a particular feature by calculating the information gain:

$$IG(X)=H(D)-\sum_{i}\frac{|D_i|}{|D|}H(D_i), \tag{5}$$

For regression tasks, decision trees use mean squared error (MSE) to evaluate the effectiveness of each split. The formula for calculating MSE is as follows:

$$MSE(D)=\frac{1}{|D|}\sum_{i=1}^{|D|}(y_i-\hat{y})^2, \tag{6}$$

After training, the final prediction of the random forest is made by aggregating the results of all decision trees. In regression tasks, the predicted value is the average of the predictions from all the trees:

$$\hat{y}=\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t, \tag{7}$$

For classification tasks, the prediction result is obtained through voting, where the final category is the one that appears most frequently:

$$\hat{y}=\mathrm{mode}\{\hat{y}_1,\hat{y}_2,\ldots,\hat{y}_T\}, \tag{8}$$

First, our paper divide the data into a training set and a test set, typically using 70%-80% of the data for training and the remaining portion for model validation. Then, the model is trained on the training set and continuously adjust hyperparameters (such as the number of trees, maximum depth, etc.) to ensure the model fits the data more effectively. To achieve optimal model performance, this section also employ techniques like grid search or random search for hyperparameter tuning.

4. Results and analysis

The following text summarizes the results of two machine learning models—Support Vector Machine (SVM) and Random Forest (RF)—used to analyze Olympic medal data.

4.1. Results of SVM

First, this paper preprocess the historical Olympic data, focusing on gold medals and medal table information. Since the dataset did not include countries that had never won a medal, our study extracted relevant information such as country names, gold, silver, bronze medals, and total medals. After cleaning the data and filling missing values, our model created a binary classification label to indicate whether a country has won a gold medal. This process revealed 43 countries that had never won a gold medal. After excluding countries that no longer exist and those with incomplete data, this study narrowed down to 8 countries.

Next, our research applied the Support Vector Machine (SVM) model with the RBF (Radial Basis Function) kernel. Our paper optimized the model using GridSearchCV to fine-tune the penalty parameter (C) and kernel coefficient (gamma), ensuring the best prediction results. After training, the model predicted the probability of these countries winning their first gold medal in the 2028 Olympics. The predicted probabilities and rankings are shown in the Fig. 1:

TABLE 2. Probability table for first-time gold medal winners at the 2028 Olympics

Country	Predicted Probability
Barbados	0.02279
Eritrea	0.02279
Cyprus	0.02279
Cabo Verde	0.02279
Burkina Faso	0.022784
Afghanistan	0.021144
Albania	0.021078
Djibouti	0.021078

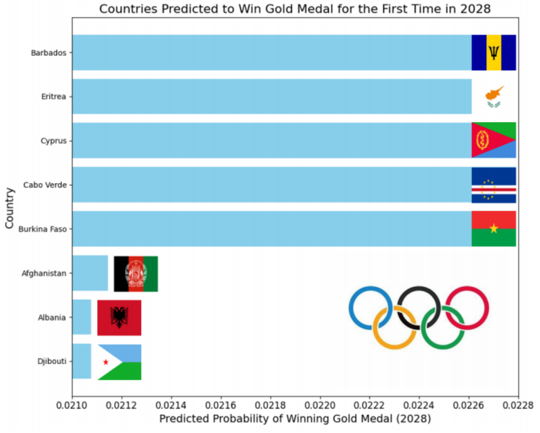

Fig 1. Countries Predicted to Win Gold Medal for the First Time in 2028

Through the chart analysis, it is found that the gold medal-winning probabilities of these countries are generally distributed between 0.0210 and 0.0228. This indicates that the probability of these countries winning their first gold medal in the current Olympics is quite low. Among them, Barbados has the highest predicted probability, which is 0.022790, though the probability is still very small. Finally, this research conducted a test on the model, and the test results are shown in Fig. 2. The AUC value obtained was 0.9356246590289143, which is close to 1, proving that the performance of our model is very good.

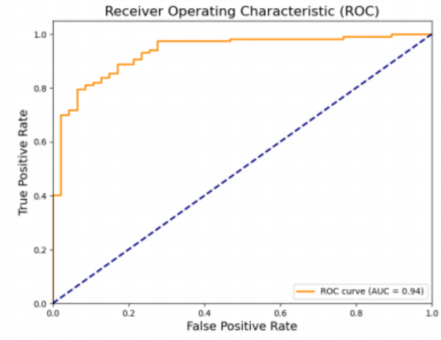

Fig 2. Receiver Operating Characteristic(ROC)

4.2. Results of Random Forest

In this analysis, our paper employed a random forest model to train data from three datasets provided and optimized the model’s hyperparameters through cross-validation to achieve ideal results. Our research first converted event types into categorical variables, while the number of events and medal counts for each sport were treated as continuous variables for model input. To improve model accuracy and reduce overfitting, this paper selected key features, including the number of events in each Olympic Games, the performance of athletes from each country (including gold, silver, and bronze medal counts), event types, and the host country’s historical performance and geographical location.

Finally, through feature selection, the model focused on the factors most closely related to medal counts. Our study then trained the data using the random forest regression model and further optimized hyperparameters such as the number of trees and tree depth through cross-validation and during the training process, resulting in improved model performance. The final data visualization results are shown in the Fig. 3:

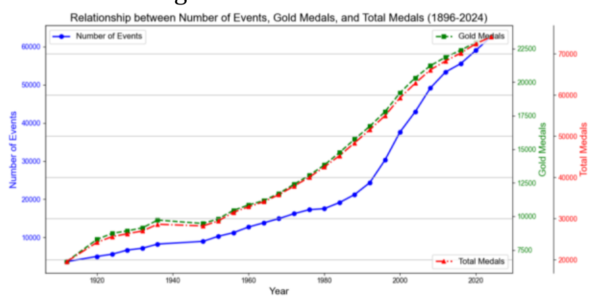

Fig 3. Relationship between Number of Events, Gold Medals, and Total Medals (1896-2024)

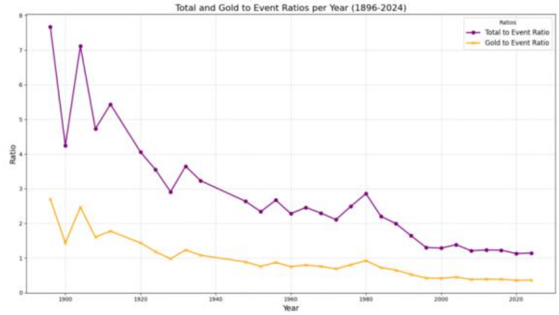

Fig 4. Total and Gold to Event Ratios Per Year(1896-2024)

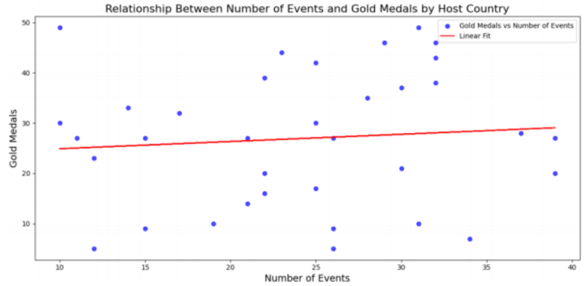

Fig 5. Relationship Between Number of Events and Gold Medals by Host Country

From the Fig. 4-5, it is observed a positive correlation between the number of events and both the gold medal count and the total medal count. Additionally, certain events, such as athletics and swimming, contribute significantly to the total medal count, while other events, like football and synchronized swimming, have a smaller contribution. Furthermore, the model reveals the distribution characteristics of medals across different events, indicating that some events dominate the medal distribution due to their high participation and larger scale. The events selected by the host country also have a certain impact on the total medal count for that country. Host countries typically choose their strongest events and invest more in them, a strategy that helps improve their performance on the medal tally.

However, upon analyzing the actual situation, it can be easily found that the major sports powerhouses tend to remain consistent over the years, and the identity of the host country does not have a significant impact on this. Among the top sports countries, the host nation enjoys a slight advantage, but for smaller countries, even as the host, the impact on their gold medal count is minimal.

Finally, this paper evaluate the performance of the model designed in this study using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2).

TABLE 3. Model Evaluation Results

Model	MSE	RMSE	R^2
Gold Medal Model	15.24	3.23	0.85
Total Medal Model	13.98	2.85	0.88

5. CONCLUSIONS

This study undertaken by our research team examined the dynamics of Olympic medal distribution through the application of two advanced machine learning techniques: Support Vector Machine (SVM) and Random Forest (RF). These algorithms were specifically employed to analyze the historical data from the Olympics and predict the likelihood of countries that had yet to win a gold medal achieving this milestone in upcoming games, specifically the 2028 Olympics. In addition, the study sought to explore the relationships between various factors influencing medal counts, including event types, host countries' histories, and the impact of athlete performances. The following summary encapsulates the research findings and conclusions based on the application of these methodologies.

Overall, this study offers valuable insights into underdog nations that have yet to secure Olympic gold medals while also exploring the broader influences that shape medal distributions at the Olympics. The use of SVM to predict the likelihood of these countries achieving gold medals in future competitions reveals the challenges they face, while the Random Forest analysis provides a robust understanding of the factors that significantly impact overall medal counts across varying conditions. These findings not only enhance our understanding of Olympic dynamics but also contribute to the body of knowledge in sports analytics, suggesting pathways for future research and potential strategies for aspiring nations in the Olympic arena.

References

[1] Bernard, A. B., & Busse, M. R. (2004). "Who Wins the Olympic Games: Economic Resources and Medal Totals." The Review of Economics and Statistics, 86(1), 413–417.

[2] Guo, Y., & Zhao, Y. (2016). "Prediction of Olympic Gold Medals Based on GM(1,1) Grey Model." International Journal of Performability Engineering, 12(5), 499–506.

[3] Wang, S. (2018). "An Integrated Approach to Predicting Olympic Medal Counts Using Multivariate Regression and BP Neural Networks." Journal of Sports Analytics, 4(3), 215–225.

[4] Kearney, K. (2009). "Using Economic and Social Indicators to Predict Olympic Success." International Journal of Sport Finance, 4(1), 45–54.

[5] Coughlan, T. (2017). "The Influence of Sporting Infrastructure on Olympic Medal Success: A Case Study of Mega Sporting Events." Sport Policy Review, 8(2), 123-140.

[6] Zhang, Y., & Li, Z. (2021). Avocado Price Prediction Using Random Forest Algorithm. International Journal of Data Science, 3(2), 89-102.

[7] Wang, L., & Zhang, J. (2023). Performance Prediction Model of Cryogenic Cooling Machine Based on Random Forest Regression. Optics and Precision Engineering, 31(5), 126-134.

[8] Chen, J., & Zhang, W. (2023). Spatiotemporal Evolution and Driving Mechanisms of Urban Construction Land Structure in the Yellow River Basin: A Random Forest Approach. Progress in Geography, 42(7), 1105-1118.