Problem Chosen	2025	Team Control Number
C	MCM/ICM	2524273

Summary Sheet

We developed a novel XGBoost-based prediction model to accurately forecast Olympic medal tally from 1960 to 2024 for gold, silver, bronze, and total medals , achieving **Normalized Root Mean Square Error(NRMSE)** values around **0.05** and **R-square** around **0.9**. The model outperformed random forest(RF) and Support Vector Machine(SVM). Based on this, we predicted the 2028 medal table, projecting the U.S. to excel while France and Australia may face setbacks.

For countries without prior medals, a specialized model incorporating the major country medal shares was proposed to predict the numbers of nations making breakthrough and achieving an **NRMSE** of **0.0589**. Our model predicts that the most likely number of countries making breakthrough in 2028 will be **three**, with a probability of **0.2655(or odds, 2.77)**.

Our model also shows the importance of each project for each country, and we are convinced that the number of medals won by a country, along with the number of medals available in a particular event, both influences the importance of events for the country. We also analyze the impact of important projects in each country on the results. Meanwhile, we analyzed the impact of events selected by home country on the results based on the principles of XGBoost.

Additionally, we assessed the great coach effect, demonstrating its significant influence on medal counts, whose weight is 0.501, surpassing host nation, events and athlete factors. Based on this we predict that in 2028: **Swimming (Australia), Swimming(Italy), and Judo (France)** will be identified as top opportunities to invest coaches.

Finally, we presented our original insights about the influence of host nations in newly introduced events.

Contents

1 Problem Restatement

1.1 Prediction Model

Develop a model to predict medal counts for each country, focusing on gold and total medals. Provide uncertainty estimates, prediction precision, and performance metrics for the model.

Based on the model, predict the medal table for the 2028 Los Angeles Olympics, including prediction intervals. Identify countries likely to improve or decline compared to 2024. Concentrate on countries with no prior medals and make prediction on the number of countries earning the first medal in your model and estimate the likelihood of these predictions. Analyze how event types and numbers influence medal tally, identify sports for different countries and evaluate how home country event selections affect results.

1.2 Great Coach Effect

Examine the impact of the great coach effect, where coaches move between countries without citizenship restrictions. Analyze data for evidence of this effect on medal counts, citing examples like Lang Ping and Béla Károlyi. Estimate the contribution of such an effect and identify three countries and sports where investing in a great coach could significantly improve results.

1.3 Original Insights

List the original insights about Olympic medal counts that your model reveal. Explain how these insight(s) can inform country Olympic committees.

2 Introduction

The Olympic Games represent a global event that garners immense attention every four years. Among the various aspects of the Games, the medal tally, particularly the gold medal tally, attract the most attention and generates a high economic impact.

To this end, we developed a prediction model based on XGBoost. Although there are many studies on predicting medal tables, existing studies typically focus on economic data (such as GDP) for prediction ([Jia et al.(2020)Jia, Zhao, Chang, Zhang and Yoshigoe]). Instead, we use historical medal tally data, athlete information, and Olympic events (types and numbers). This study predicts the medal tally for the next Olympic Games, including the total number of medals, the medals of each kind and events, the number of athletes and events, and the host country.Compared to models based on economic data, it is simpler and more reliable, because accurately assessing a country's economic conditions can be challenging. Ultimately, our model performed excellently, with the normalized root mean square

error (NRMSE) not exceeding 0.1 and R-squared(R^2) around 0.9, outperforming previous studies ([Jia et al.(2020)Jia, Zhao, Chang, Zhang and Yoshigoe], [Condon et al.(1999)Condon, Golden and Wasil]).

Additionally, we focused on countries that did not win a medal and use our model to predict their probability of achieving their first-ever Olympic medal in the next Olympics. Using our model, we estimated the probability of each country breaking its zero-medal record and derived the probability distribution of the total number of such breakthroughs. Our prediction suggests that in 2028 Olympics, the most likely number of breakthrough countries will be three, with a probability of 0.2655. The corresponding odds ratios were also calculated.

Based on this, we analyzed the variations in the importance of different events across countries and identified the most important for each nations. In our study, we selected the top five most important events for each country. Our model revealed that the most important events are not entirely correlated with the number of medals countries earn. It also depends on the number of medals set by International Olympic Committee(IOC). Simultaneously, we analyzed the impact of events selected by different countries on the forecast results.

Moreover, coaches play a significant role in determining the athletic performances of competing nations. For instance, coaches like Lang Ping, who led both the Chinese and U.S. women's volleyball teams to championships, are categorized as great coaches. However, coaching changes occur frequently and often unexpectedly, making the impact of a coach difficult to estimate. Therefore, we developed a model using XGBoost to evaluate their influence on the medal tally. Based on our model, we selected three countries and identified the events in which they should invest large in coaches, along with predictions of the potential impact on their medal performance.

Additionally, we observed that the host country's influence on the number of medals won in newly set events is strongly correlated with the total number of medals newly introduced, with the correlation coefficient of 0.699. Additionally, among all nations, Japan is the country that won the most medals in its own introduced events, as well as the highest proportion of such medals. Based on this, we provided recommendations to the host country Olympic Committee.

In summary, we developed a series of prediction and evaluation models based on XGBoost, and the flowchart of our study is shown below.

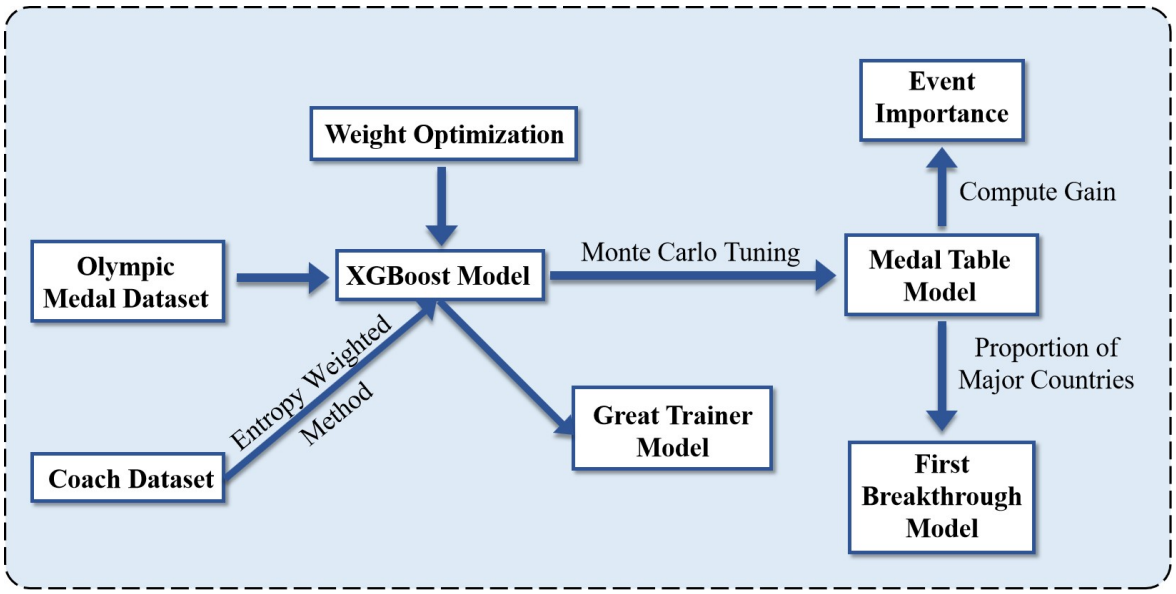

Figure 1: Our Model

The remainder of this paper is organized as follows: We examine previous research on Olympic medal predictions and related methodologies in literature review. We also detailed how we used XGBoost to predict and analyze the medal table, medal-less countries, and the importance of specific events. Additionally, we used a similar model to estimate the impact of the great coach effect and identified the most valuable events for investment in three countries. Moreover, we presented our original insights about the influence of host nations in newly introduced events. Finally, the section Conclusion and Discussion section provides potential explanations for our findings and suggests possible improvements.

3 Literature Review

The analysis of the distribution of medal counts in the Summer Olympic Games, along with the corresponding total medal counts for nations, is not a novel area of research. Research on this topic dates to the 1970s. In 2004, a significant breakthrough was made when [Bernard and Busse(2004)] compared various econometric methods and concluded that the Tobit model consistently yields better results.

Subsequently, the use of the Tobit model to estimate the distribution of medal attainment became the standard approach ([Forrest et al.(2010)Forrest, Sanz and Tena]).
Moreover, since [Bernard and Busse(2004)]also focused on prediction, Tobit regression has become the traditional method for forecasting national total medal counts.

Building on their work, researchers have incorporated several additional variables, such as economic level, inertia from the previous Olympic Games, and national sports expenditure

([Maennig and Wellbrock(2008)]). In Addition, other models were also utilized. For instance, [Nevill et al.(2012)Nevill, Balmer and Winter] employed a logistic regression model to accurately predict the performance of United Kingdom in the 2012 Olympics.

Notably, [Lowen et al.(2014)Lowen, Deaner and Schmitt] have examined the impact of gender on sports performance and found that gender ratio and imbalance have no significant effect on the medal standings.

With the development of technologies such as machine learning, new prediction algorithms have emerged, often relying on historical medal data and other fundamental variables, with reduced data dimensionality and improved prediction accuracy. For example,
[Jia et al.(2020)Jia, Zhao, Chang, Zhang and Yoshigoe] employed Random Forest for prediction, while [Zhang et al.(2024)Zhang, Zhou and Bai] used neural network, whose the R-squared value reaches 87%. [Condon et al.(1999)Condon, Golden and Wasil] proposes a point-based system for medals and uses neural networks for prediction. Additionally,
[Nagpal et al.(2023)Nagpal, Gupta, Verma and Kirar]compared several classical machine learning models, such as Bayesian and Ridge regression, concluding that no model currently exhibits a significant advantage. And [He and Wang(2024)]conducted their research based on robust time series analysis.

Among them, random forests(RF) are a widely used algorithm. [Horvat and Job(2020)] highlighted their significant role in sports-related predictions. And
[Schlembach et al.(2022)Schlembach, Schmidt, Schreyer and Wunderlich] utilized random forests to predict Olympic medal standings using economic data. Similarly,[Shailaja(2020)] employed random forests to analyze the performance of India in the Olympics. In summary, method related to random forests have been widely applied in Olympic-related research and proven to be highly effective.

Compared with random forests, XGBoost employs gradient boosting algorithms, often resulting in better performance ([Chen and Guestrin(2016)]). [Passi and Pandey(2018)] used this model to predict cricket match outcomes, outperforming the RF model, while studies by [Sagala and Ibrahim(2022)] and [Huimin et al.(2024)Huimin, Dongying and Yonghui] on Olympic medal prediction suggested that XGBoost was the optimal model for this field of research.

Significant research has also been conducted on individual sporting events. For instance,
[Nevill et al.(2012)Nevill, Balmer and Winter] examined the impact of competitions such as the Judo World Championships, while [Iyer and Sharda(2009)] focused on the performance of cricket players. [Huang and Chen(2011)] employed an MLP model to predict World Cup matches.
Moreover, [Gu et al.(2019)Gu, Foster, Shang and Wei] and [Pischedda(2014)] utilized Support Vector Machines (SVM) to forecast rankings in the National Hockey League (NHL), whereas [Bednar and Bauer(2011)] applied various neural networks, including feedforward, radial basis, and generalized regression neural networks.

Additionally, [Csurilla and Fertő(2024)] focuses on the countries earning their first medal. It employs a zero-inflated beta regression and suggests that the probability of a zero-medal country winning

a medal is related to the expected medal count of superpowers (such as China, the U.S., and Russia).

Systematic analysis of the great coach effect in the Olympics is relatively scarce. However the influence of coaches on athletes has been extensively studied across various sports. [Barth et al.(2020)Barth, Güllich, Raschner and Emrich] identifies a significant relationship between coach-dominated links and athletes' international performance through various tree-based algorithms. Moreover, [Cook et al.(2021)Cook, Fletcher and Peyrebrune] revealed differences among coaches in swimming. In conclusion, it is widely believed that coaches impact athletes' ability to win medals, but the precise nature of this influence remains inconclusive.

4 Result

In this study, we employed the XGBoost model to predict the medal tally, focusing on the total medal counts for each country and on the countries likely to win their first Olympic medals. Our model performed excellently in predicting medal counts, including gold, silver, bronze, and total medals, with NMSE values below 0.1, indicating an accuracy rate of over 90%(Table1,10), better than RF and SVM. The model also performed well in predicting the number of countries achieving their first-ever medals, with an NMSE of 0.0589, significantly outperforming the reference model (gray prediction), which had an NMSE of 0.369. Based on our model, we predict that the most likely scenario in the 2028 Olympics will involve three countries winning their first gold medals, with a probability of 0.2655(Figure3,4). The corresponding odds are 2.77 (or 3.77, depending on the type of odds used).

We also analyzed the importance of different events in various countries and found distinct patterns between sports powerhouses and other nations. We believe that the importance of different events is related to both the number of medals set by IOC and the proportion of medals that contribute to a country's total medal count.

Simultaneously, we developed an XGBoost model to assess the impact of the great coach effect. Ultimately, after using the difference method to exclude the effect of national activity level, we found that the great coach effect significantly impacted the medal results of individual events, with a weight of 0.501, higher than that of the host country (0.439) and GDP (0.039). The number of events in that Olympic Games and the participation of the country in those events had a minimal impact. Finally, we selected the three most affected countries and their events, and assessed the potential impact of the great coach effect(Table6).

Moreover, we found a correlation between the host nation's performance in newly introduced events and the number of medals awarded in these events, with a correlation coefficient of 0.669. Notably, among the host nation, Japan performed exceptionally well in this regard. We provided recommendations to the host nation's Olympic committee based on these findings.

5 Method

5.1 Medal Tally Prediction

In this section, we perform data pre-processing. We then present a model based on XGBoost to predict the medal table, the countries that won their first medals, and assess the importance of events.

5.1.1 Data Pre-processing and Predict Medal Tally

First, we pre-process the data. We verify the authenticity of our data, integrate different numbers from the same country(We consider teams with the same National and Regional Olympic Committees (NOCs) as the same team) and check the completeness of dataset. Next, we filter out the factors we need from the data set. Generally, factors influencing medals include previous medal counts, the structure of Olympic events, and the number of participating athletes ([Forrest et al.(2010)Forrest, Sanz and Tena]). To this end, we gathered relevant data from the inception of Olympics. Owing to the underdeveloped event structure and the limited number of participating countries prior to 1960, we started our analysis from the 1960 Summer Olympics. We compiled data on the number of medals earned by each country and event annually, including gold,silver, and bronze medals. We also compiled the number of participating athletes, the number of events, and the proportion of medals won by major countries (the China, USA, Russia, France, and the UK) in each event. Moreover, we considered certain historical factors, such as the Soviet Union and West Germany teams, which, for specific reasons, could no longer exist. Additionally, we assigned weights of 10, 5 and 2 to gold, silver, and bronze medals respectively, and we calculate the medal score(Hereafter referred to as score) to differentiate between the various medal types and replace the total medals:

$$score = 10N_{gold} + 5N_{silver} + 2N_{bronze} \tag{1}$$

In the formula, N is the number of medal(gold,silver or bronze) This method has also been applied in [Horvat and Job(2020)].

EXtreme Gradient Boosting (XGBoost) is a machine learning algorithm based on the Gradient Boosting Decision Tree (GBDT), which is known for its efficiency, flexibility, and accuracy.

For a data set containing n dimensions, the XGBoost model can be expressed as(2):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F(i = 1, 2, ...n) \tag{2}$$

The objective function of the XGBoost model are as follows(3):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}i^{(t)}) + \sum k = 1^t \Omega(f_k) \tag{3}$$

In the equation $ell(y_i, \hat{y}_i^{(t)})$ is a loss function, that is often used to measure models predict ($haty_i(t)$) and the error between the actual label (y_i). In regression problems, the common loss function is the square error loss(4):

$$ell(y_i, \hat{y}_i^{(t)}) = \frac{1}{2}(y_i - \hat{y}_i^{(t)})^2 \tag{4}$$

$\Omega(f_k)$is a regularization term that penalizes model complexity to prevent over-fitting. The regularization terms of XGBoost used in this study are defined as equation5:

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} w_j^2 \tag{5}$$

Among them, T_k is the number of leaf nodes in k tree. w_j is the weight of the leaf node. γ is the penalty factor that controls the number of leaf nodes. λ is the L2 regularization coefficient that controls the weight of leaf nodes.

To optimize the objective function, XGBoost employs the gradient boosting method, incrementally adding a new tree (f_t) in each iteration to fit the residuals of the previous model. Specifically, XGBoost uses a Taylor expansion to approimate the objective function up to the second order, enabling more precise optimization(6):

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[\ell(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t) \tag{6}$$

Among them: $g_i = \frac{\partial \ell(y_i, \hat{y}_i^{(t-1)})}{\partial y_i^{(t-1)}}$is one of the loss function derivative (gradient). $h_i = \frac{\partial^2 \ell(y_i, \hat{y}_i^{(t-1)})}{\partial (y_i(t-1))^2}$ denotes the loss function of the second derivative (Hessian matrix). This second order gradient optimization method enables XGBoost to converge to the global optimal solution faster, and has higher robustness and accuracy during model training. The prediction results are presented below. Due to the large number of countries and years, only the top eight predictions for the 2028 Summer Olympics(Figure2, and the corresponding model performance(Table1) are presented here. The result of silver and bronze is shown is appendix(Table10) The learning rate of our model is 0.1 ,the maximum depth is 5 and the n_estimators is 100.

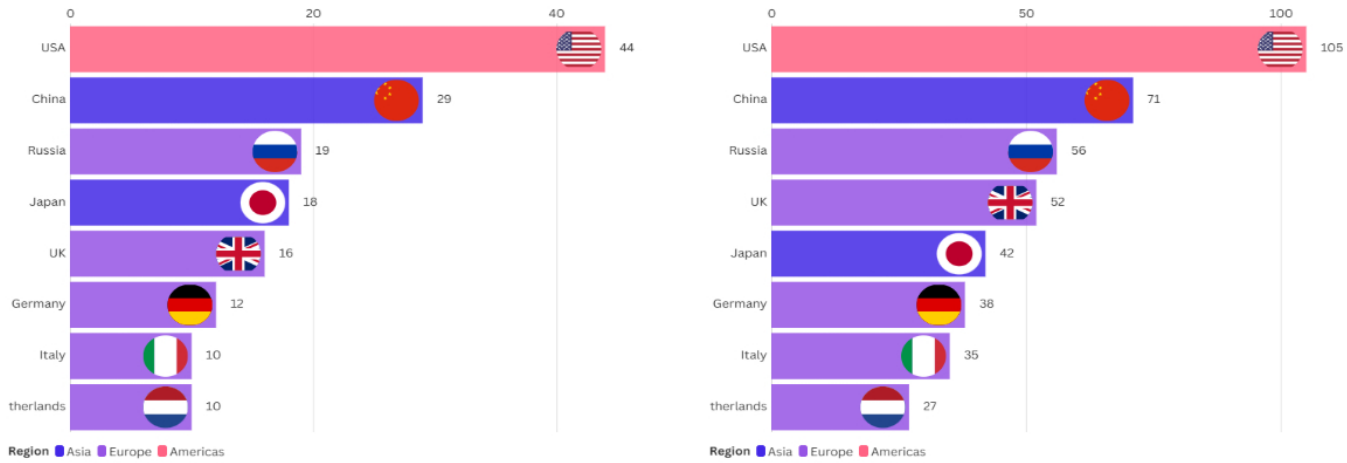

Figure 2: The prediction medal tally

The left figure shows the gold medal tally and the right denotes total medal tally.

Table 1: The performance of XGBoost

Type	**NRMSE**	R^2
Gold	0.072	0.830
Silver	0.061	0.849
Bronze	0.082	0.818
Total	0.052	0.911

NRMSE denotes the normalized root mean square error, which is defined as follows(7):

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2 - y_i^2}}{\overline{y}} \tag{7}$$

Where y_i represents the true value, $\hat{y}_i$ denotes the predicted value, and $\overline{y}$ represents the average value of the variable.

Simultaneously, we employed random forest(RF) to predict the medal standings. RF is a powerful and versatile machine learning method that combines the strengths of multiple decision trees through ensemble learning, bootstrap aggregation, and random feature selection. Its ability to handle high-dimensional data, mitigate over-fitting, and provide insights into feature importance makes it a widely used technique in various domains([Biau and Scornet(2016)]). In this study, its n_estimators is 100, the maximum depth of the random forest is set to 15, and the minimum number of samples for leaf nodes and split nodes is two. These parameters are tuned using the Monte Carlo method and we employed 5-fold cross-validation. The result is shown in Table2:

Table 2: The performance of RF

Type	NRMSE	R^2
Gold	0.085	0.762
Silver	0.072	0.800
Bronze	0.097	0.820
Total	0.085	0.877

Clearly, XGBoost outperforms RF. We assume that is owing to the regularization of its loss function and its boosting-based optimization, which contribute to better generalization and stronger robustness. We also tested back propagation neural networks and support vector machines, whose R^2 is below 0.6, and could not make valuable predictions. Based on our predictions, the United States is expected to see an increase in the number of medals in the 2028 Los Angeles Olympics compared to the 2024 Paris Olympics, whereas France and Australia are predicted to experience a significant decline.

5.1.2 Predicting Medal Probabilities for Countries with No Medals

Countries with their first medals are also worth attention. In addition to the top-ranking countries on the medal table, we also focus on these countries

Then, we introduced a new variable: the total number of medals won by the top 20 countries and enabled the model to output probabilities for each class. Then, we retrained the model specifically for them, keeping the same validation set division and parameters. The probability calculation of the RF model is as follows(8):

$$p_x = \frac{N_i}{N} \tag{8}$$

Where N_i is the number of leaf nodes of event i and N is the total number of leaf nodes.

Next, we use the probabilities of these countries winning medals to calculate the probability distribution for countries achieving a breakthrough of zero. This distribution is similar to a binomial distribution. However, because the probabilities differ for each country, we employ computational methods to solve it. We then used this probability distribution to predict the number of countries achieving breakthroughs in the past five Olympic Games and the 2028 Olympics. The results are displayed below(figure4,figure3):

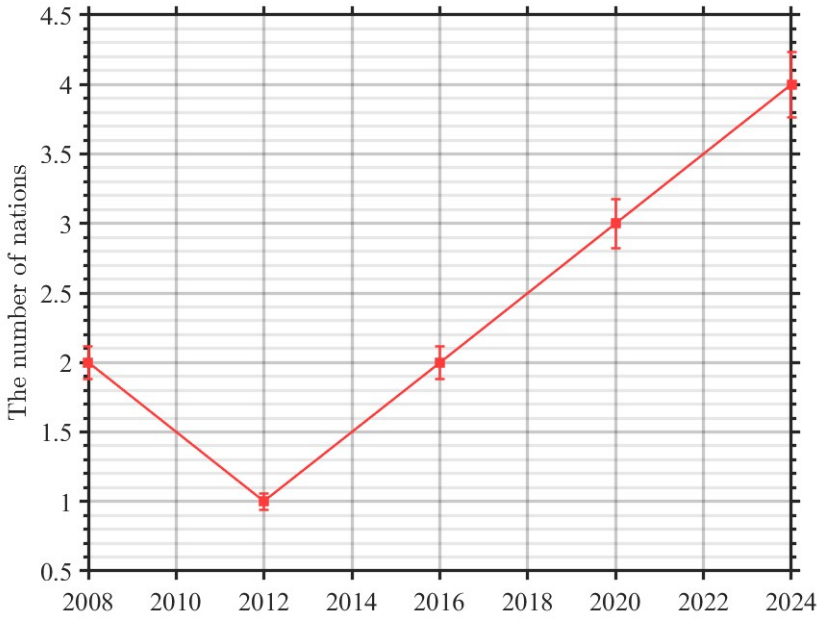

Figure 3: The prediction of past 5 Olympics

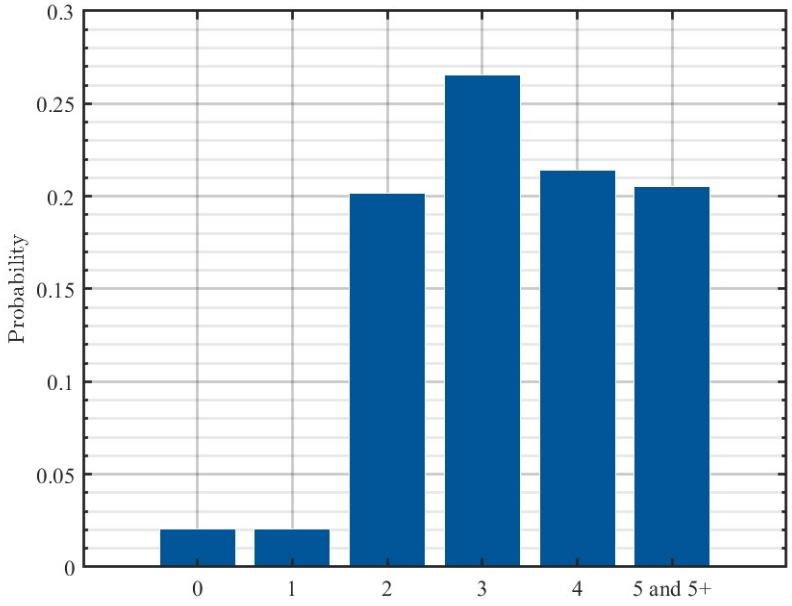

Figure 4: The 2028 Olympic Games probability distribution

In Figure 3, the x pixel denotes time, the y pixel is the number of countries earning the first medal. The error bars indicate the range of deviations observed in the multiple simulations, with an average relative error of 0.0589. Figure 4 presents the probability distribution of the number of countries

winning medals in the 2028 Olympics. The x pixel denotes the number and the y pixel denotes the probability.

The most likely scenario is that three countries achieving breakthroughs, with a probability of 0.2655. Based on this, we can calculate the corresponding odds:

$$Odd_P = \frac{1-P}{P} \tag{9}$$

In the formula, P is the probability of the occurrence of an event.

Upon calculation, the odds are 2.77, meaning that for every dollar invested, a profit of 2.77 dollars would be generated upon victory. Alternatively, this can be interpreted as a return of 3.77 dollars.

Furthermore, because our model provides the breakthrough probability for each country, we also identified the country or region most likely to achieve a breakthrough. Our prediction indicates that North Yemen has the highest likelihood, with a probability of 0.409. All other potential countries are listed in the Appendix(Table8).

For comparison, we applied a gray prediction model. Gray prediction is a forecasting method used for systems with uncertainty and limited data([Julong et al.(1989)]). In this paper, we use model GM(1,1), which uses data accumulation and differential equations to make predictions based on small datasets. It is widely applied in fields such as economics and engineering. Ultimately, its NMSE was 0.369, which is significantly higher than ours, demonstrating the effectiveness of our model.

5.1.3 Most Important Events for Different Countries

XGBoost can also identify the most important sports for each country. In our model, the importance of nodes is assessed based on the gain, which is defined in XGBoost as follows:

$$\text{Gain} = \frac{1}{2}\left(\frac{(\text{sum}left)^2}{\text{count}left+\lambda} + \frac{(\text{sum}right)^2}{\text{count}right+\lambda} - \frac{(\text{sum}parent)^2}{\text{count}parent+\lambda}\right) - \gamma \tag{10}$$

Where sum$left$ and count$left$ are the sum of the target values of the left child node and the number of samples; sum$right$ and count$right$ are the sum of the target values of the right child node and the number of samples; sum$parent$ and count$parent$ are the sum of the target values of the current node $parent$ and the number of samples; λ is the L2 regularization term used to control model complexity (to prevent overfitting); γ is the minimum gain required for splitting, and splitting is performed only if the gain is greater than γ.

Typically, nodes with higher gains contribute more to an outcome. In other words, the selected events were the most influential in predicting medals.

Taking China as an example, our model identifies gymnastics as the event most likely to influence medal predictions, indicating that if China performed well in gymnastics during the previous Olympics,

we anticipate that their overall performance in the next Olympics will improve.
Here, we applied random forests to select the top five most important sports for each country, and we present four countries as examples: China,the U.S., Cuba and Norway(Figure5).

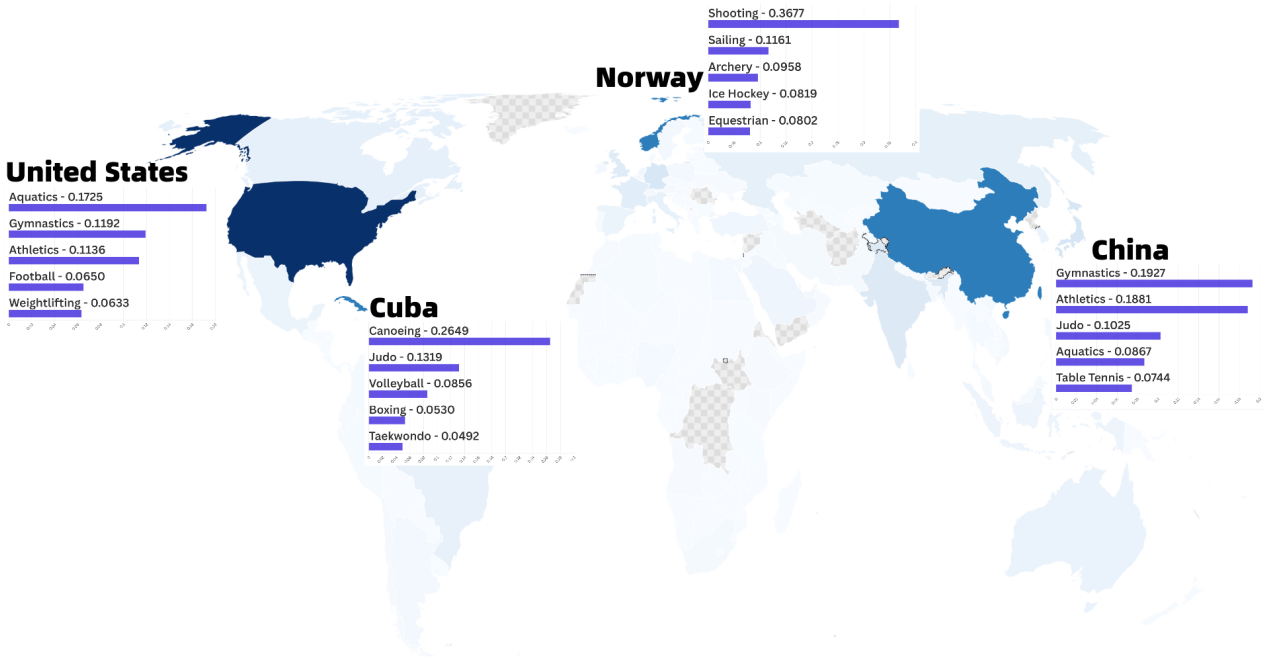

Figure 5: The importance of example countries

In the figure, we exhibit the importance of example countries with bar chart. We observed that the most significant events are those with a higher number of medals set by the IOC. At the 2024 Paris Olympics, the top eight events in terms of the number of medals awarded are as follows: Aquatics (49 medals), Athletics (48 medals), Cycling (22 medals), Gymnastics (18 medals), Wrestling (18 medals), Judo (15 medals), and Shooting (15 medals). Clearly, these events are more frequently featured among the most important events. Furthermore, events with a high proportion of medals in a given country are typically more significant. For instance, aquatics are highly prioritized in China, which has a similar importance in the United States. It account for a substantial share of their medal tallies(most for China and the second most for the U.S.).

We also conducted a correlation test on the event importance and number of events and the proportion of medals. The correlation coefficient is between 0.3 and 0.4, and the P-value is less than 0.05, which indicates that there is a certain correlation between them. In conclusion, we believe that the number of medals won by a country, along with the number of medals available in a particular

event, jointly influence the importance of events for the country.

5.2 Great Coach Effect

However, owing to limitations in the dataset, our model overlooks the role of coaches. In this section, we introduce a novel model based on entropy weighted method(EWM) and XGBoost regression model to evaluate the impact of coaches and analyze the influence of 14 great coaches(Given in Table9). EWM is a technique used in multi-criteria decision making (MCDM) to determine the weights of various decision criteria. In this method, entropy is used to measure the uncertainty or diversity of information. Information entropy is expressed as follows(11):

$$E_j = -k \sum_{i=1}^{m} p_{ij} \ln(p_{ij}) \tag{11}$$

Where E_j is the information entropy of the random variable X. p_{ij} is the probability of event occurring. k is a constant, usually taken $\frac{1}{\ln(m)}$, where m is the number of samples.

Generally, a lower entropy means that the information is more concentrated and important. Therefore its weight should be higher. This can be calculated as follows(12):

$$w_j = \frac{1 - E_j}{\sum_{j=1}^{n}(1 - E_j)} \tag{12}$$

Wherew_j denotes the weight.
Thus, we obtained the preliminary weights of the variables(Table3):

Table 3: Preliminary weights

Variable	Information entropy value	Information utility value	Weight(%)
Coach	0.605	0.395	36.905
Host	0.488	0.512	47.857
Events	0.957	0.043	4.029
Athletes	0.988	0.012	1.086
GDP	0.892	0.108	10.123

Thus, we can quantify the effects of various variables, as expressed by the following equation, C represents total contribution, and N_i is the value of variable i :

$$C = \sum w_i N_i \tag{13}$$

However, the high proportion of zero values in the data from host countries and great coaches,

may cause data distortion. Consequently, we introduce the Correction coefficient of these variables. Then, we use XGBoost to optimize them. Moreover,to eliminate the effect of each country's sport level, we used the change value of its individual MEDALS as the dependent variable.

$$C = \sum w_i N_i + p w_c N_c + q w_h N_h \tag{14}$$

Where $\sum w_i N_i$ do not include the coach and host. Subsequently, to reduce the disparity in athletic performance between countries, we applied a difference transformation to the number of medals

Next, we calculate the R-squared and NRMSE of XGBoost to optimize p and q using grid search, with the following efficiency function f(15):

$$f = R^2 + NMSE \tag{15}$$

We set the range for p and q as $[w_i - 5, w_i + 5]$,in which w_i is the weight calculated by EWM. and te step size of 0.001. Then, we traverse the entire space and obtain the weights for p and q , 3.48 and 2.35 respectively. Normalize the weight, the new weights of variables are as follows(Table4).

Table 4: Final weights

Variables	**Weights**
Coach	0.501
Host	0.439
Athletes	0.005
event	0.014
GDP	0.039

It is evident that, for individual events, the influence of great coaches is most significant, followed by the host country advantage, while GDP, number of athletes, and the number of events have minimal impact. For robust test, we calculate their Variance inflation factors(VIF) and they are all below 10, indicating no multicollinearity problems.

Then, we applied the optimized parameters to other models for validation, yielding the following results(Table5):

Table 5: Performance of other models

Model	R^2
Random forest(RF)	0.833
Gradient boosting regression(GBR)	0.802
Gradient Boosting Decision Tree(GBDT)	0.63.

Where the n_neighbors of KNN is 3. For GBR,GBDT and RF, the n_estimators is 100, its maximum depth set to 15, and the minimum number of samples for leaf nodes and split nodes is 2. The leaning rate of the GBR and GBDT is 0.1. Evidently, the results of RF and GBR indicate that the applied weights are generalizable, reflecting some fundamental characteristics of the data. Moreover, the poorer performance of GBDT highlights the necessity of using XGBoost. Notably, the R-squared value for KNN is 0.537, while that for SVM is 0.281. These results are consistent with previous findings using other datasets, suggesting that SVM and KNN may not be as effective as initially anticipated for this task. We believe this is due to the inability to obtain a sufficiently large sample size for medal predictions.

Building on this, we sequentially assume that they have a great coach and predict the number of medals. The relevant data for 2028 is sourced from forecasts published by the International Monetary Fund (IMF) and the official Olympic website.Subsequently, we selected the events from the three countries with the most significant improvements. Details are presented in Table6.

Table 6: Nations and events recommended to invest coaches

Nation	**Event**	**Predict gold medal difference**
Italy	Swimming	8.02
France	Judo	6.03
Australia	Swimming	5.80

Interestingly, the presence of a great coach has a greater impact on teams that are weaker in the event and have performed poorly in recent years. For example, the impact on the Italian and Australian swimming teams (approximately 6 medals) was greater than on the U.S. and Chinese teams (4.5 and 2.0 medals, respectively), suggesting that a great coach may be more helpful in achieving breakthroughs in specific events, rather than maintaining a leader.

5.3 Other Insights

Additionally, we observed that the influence of host country on the number of medals won in newly established events is strongly related to the total number of medals introduced. In Olympics with more new events, the host country Generally achieves a higher medal percentage(Figure6). Where red line denotes the total score, and the black line represent the host.

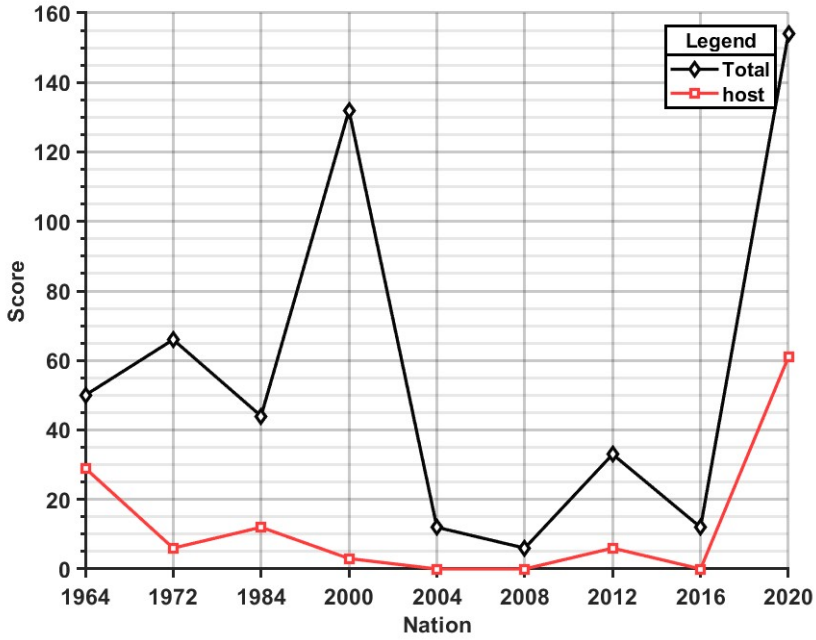

Figure 6: Number of newly set medals

Clearly, they have a similar trend. Additionally, we conducted a correlation analysis between the two variables, and found that with a p-value below 0.05, the Pearson correlation coefficient was 0.6693, indicating a statistically significant relationship. Therefore, From the perspective of the medal tally, we recommend that the country Olympic Committee of the host nation introduce more new events and medals, particularly in sports that were not featured in previous Olympic Games. For example, Japan excelled in this regard. In 2020 Tokyo Olympics, through flexibly adjusting the new events such as skateboarding, karate, and 3x3 basketball, they earned over 30% of the newly awarded medals, significantly surpassing the average value of 7.5% for non-host countries.

6 Conclusion

This study introduces an XGBoost-based prediction model built solely on historical medal data, without incorporating additional factors. Our main contribution is the accurate prediction of the medal table, surpassing predictions based on economic data ([Condon et al.(1999)Condon, Golden and Wasil]) in terms of both simplicity and robustness.

Another key contribution is the development of a model capable of accurately predicting both top-ranking and medal-less countries, achieved by incorporating the medal share ratio of leading nations.Additionally, by leveraging XGBoost's features, we identified the five most important events for each country and analyzed the relationship between selected events and outcomes.

Moreover, a significant contribution is estimating the great coach effect, achieved by introducing an XGBoost-based parameter optimization algorithm to determine its weight. Using this algorithm, we identified the top three countries and the most valuable projects for coaching investment.

Additionally, we proposed original insights related to the host nation's performance in newly introduced events and the number of medals awarded in these events, and we provided recommendations based on this insight.

7 Discussion

We used an XGBoost-based model, which outperformed comparison models such as RF, KNN, GBR, and SVM, likely owing to the advantage of XGBoost in handling high-dimensional, complex data. Moreover, our model typically selects events with a high medal share and a large number of total medals across countries, which we believe is reasonable because these events often reflect a country's athletic prowess.

Additionally, We analyzed the impact of great coach effect on individual events and found it to be far greater than factors such as the host country or GDP. We assume that the reason is that the other variables influences all events, whereas a great coach directly impacts a single event. If we analyze the overall medal tally, the coach's influence on the total medal count may not exceed these factors. However, this requires dealing with numerous confounding factors, such as the impact of other coaches, which necessitates more data and advanced algorithms.

Of course, there is still room for improvement in proposed model. We could use more modern models such as LSTM and CNN, but owing to the small sample size and high-dimensional nature of the dataset, we remain cautious about the performance of neural network in this context.

Additionally, we could incorporate athlete strength indicators, such as rankings and past medal counts, into our model, which may be a direction for future improvements. Furthermore, integrating the great coach effect more effectively with our prediction model could be our key focus of our future research.

In addition, Using momentum, including psychological and strategic momentum ([Gilovich et al.(1985)Gilovich, Vallone and Tversky]), to process our data is another intriguing direction. This may require additional data, such as national sports investments and World Championship results.

8 Acknowledgment

Thanks to the following websites for providing us with related data(Table7):

Table 7: Website providing data

Web	Data
https://www.olympics.com	Most of coach information
https://www.atptour.com	About tennis coaches
https://www.sport.gov.cn	Chinese coaches
https://www.stats.gov.cn	GDP of China
https://www.bea.gov	GDP of USA
https://lib.nju.edu.cn	GDP of other countries
https://www.wikisport.eu	Supplement and correct some coaching information

9 Appendix

The full forecast for first-time medalists in 2028 is shown here(8)(The probability for countries not given is less than 0.05):

Table 8: Full forecast for first-time medalists

Nation	Probability
North Yemen	0.409
South Sudan	0.272
Malaya	0.265
Bosnia and Herzegovina	0.24
Angola	0.24
Dansk Idrts Forbund	0.186
Crete	0.186
Saar	0.18
Congo (Brazzaville)	0.15
South Vietnam	0.13

Table 9: Coaching History

Name	Years	Country
Alejandro Villa	1988-2000	Brazil
Don Talbot	1996-2008	Australia
Martha Karolyi	1984-2012	U.S.
Bob Bowman	2004-2012	U.S.
LI Yongbo	2000-2016	China
Park Joo-bong	2008-2016	Japan
Bernardo Rezende	2004-2016	Brazil
Toni Nadal	1990–	Spain
Marta Károlyi	1996-2016	U.S.
Ping Lang	2005-2008	U.S.
Ping Lang	2012-2020	China
YiFu Wang	2008–	China
Cerioni Stefano	2008-2012	Italy
Cerioni Stefano	2016–	Russia
George H. Morris	2008–	U.S.

Table 10: The predicted tally of 2028 Olympics

States	Gold	Sliver	Bronze	Total
U.S.	44	35	26	105
China	29	23	17	71
Russian	19	17	19	56
Japan	18	11	16	42
Great Britain	16	17	19	52
Germany	12	12	14	38
Italy	10	11	14	35
Netherlands	10	8	10	27

References

[Barth et al.(2020)Barth, Güllich, Raschner and Emrich] Barth, M., Güllich, A., Raschner, C., Emrich, E., 2020. The path to international medals: a supervised machine learning approach to explore the impact of coach-led sport-specific and non-specific practice. PLoS One 15, e0239378.

[Bednar and Bauer(2011)] Bednar, E.M., Bauer, K.W., 2011. Journal of quantitative analysis in sports predicting nba games using neural networks. URL: https://api.semanticscholar.org/CorpusID:121658250.

[Bernard and Busse(2004)] Bernard, A., Busse, M., 2004. Who wins the olympic games: Economic resources and medal totals. The Review of Economics and Statistics 86, 413–417. doi:10.1162/003465304774201824.

[Biau and Scornet(2016)] Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227.

[Chen and Guestrin(2016)] Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

[Condon et al.(1999)Condon, Golden and Wasil] Condon, E.M., Golden, B.L., Wasil, E.A., 1999. Predicting the success of nations at the summer olympics using neural networks. Comput. Oper. Res. 26, 1243–1265. URL: https://api.semanticscholar.org/CorpusID:29759917.

[Cook et al.(2021)Cook, Fletcher and Peyrebrune] Cook, G.M., Fletcher, D., Peyrebrune, M., 2021. Olympic coaching excellence: A quantitative study of psychological aspects of olympic swimming coaches. Psychology of Sport and Exercise 53, 101876. URL: https://www.sciencedirect.com/science/article/pii/S146902922030861X, doi:https://doi.org/10.1016/j.psychsport.2020.101876.

[Csurilla and Fertő(2024)] Csurilla, G., Fertő, I., 2024. How to win the first olympic medal? and the second? Social Science Quarterly URL: https://api.semanticscholar.org/CorpusID:272235010.

[Forrest et al.(2010)Forrest, Sanz and Tena] Forrest, D., Sanz, I., Tena, J., 2010. Forecasting national team medal totals at the summer olympic games. International Journal of Forecasting 26, 576–588. URL: https://www.sciencedirect.com/science/article/pii/S0169207009002088, doi:https://doi.org/10.1016/j.ijforecast.2009.12.007. sports Forecasting.

[Gilovich et al.(1985)Gilovich, Vallone and Tversky] Gilovich, T., Vallone, R., Tversky, A., 1985. The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology 17, 295–314. URL: https://api.semanticscholar.org/CorpusID:317235.

[Gu et al.(2019)Gu, Foster, Shang and Wei] Gu, W., Foster, K., Shang, J.S., Wei, L., 2019. A game-predicting expert system using big data and machine learning. Expert Syst. Appl. 130, 293–305. URL: https://api.semanticscholar.org/CorpusID:145847209.

[He and Wang(2024)] He, Z., Wang, Z., 2024. Prediction of olympic medal count for usa based on robust time series model and computer implementation, p. 165. doi:10.1117/12.3033012.

[Horvat and Job(2020)] Horvat, T., Job, J., 2020. The use of machine learning in sport outcome prediction: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10, e1380. doi:10.1002/widm.1380.

[Huang and Chen(2011)] Huang, K.Y., Chen, K.J., 2011. Multilayer perceptron for prediction of 2006 world cup football game. Adv. Artif. Neural Syst. 2011, 374816:1–374816:8. URL: https://api.semanticscholar.org/CorpusID:11426237.

[Huimin et al.(2024)Huimin, Dongying and Yonghui] Huimin, S., Dongying, Z., Yonghui, Z., 2024. Can olympic medals be predicted? based on the interpretable machine learning perspective. Journal of Shanghai University of Sport 48, 26–36.

[Iyer and Sharda(2009)] Iyer, S.R., Sharda, R., 2009. Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications 36, 5510–5522. URL: https://www.sciencedirect.com/science/article/pii/S095741740800420X, doi:https://doi.org/10.1016/j.eswa.2008.06.088.

[Jia et al.(2020)Jia, Zhao, Chang, Zhang and Yoshigoe] Jia, M., Zhao, Y.L., Chang, F., Zhang, B., Yoshigoe, K., 2020. A random forest regression model predicting the winners of summer olympic events. Proceedings of the 2020 2nd International Conference on Big Data Engineering URL: https://api.semanticscholar.org/CorpusID:220347807.

[Julong et al.(1989)] Julong, D., et al., 1989. Introduction to grey system theory. The Journal of grey system 1, 1–24.

[Lowen et al.(2014)Lowen, Deaner and Schmitt] Lowen, A., Deaner, R., Schmitt, E., 2014. Guys and gals going for gold: The role of gender empowerment in olympic success. Journal of Sports Economics 17. doi:10.1177/1527002514531791.

[Maennig and Wellbrock(2008)] Maennig, W., Wellbrock, C.M., 2008. Sozioökonomische schätzungen olympischer medaillengewinne. analyse-, prognose- und benchmarkmöglichkeiten (socio-economic estimations of winning olympic medals: Analysis, prognosis and benchmark possibilities). Sportwissenschaft 38, 131–148. doi:10.1007/BF03356075.

[Nagpal et al.(2023)Nagpal, Gupta, Verma and Kirar] Nagpal, P., Gupta, K., Verma, Y., Kirar, J., 2023. Paris Olympic (2024) Medal Tally Prediction. pp. 249–267. doi:10.1007/978-981-99-1414-2_20.

[Nevill et al.(2012)Nevill, Balmer and Winter] Nevill, A.M., Balmer, N.J., Winter, E.M., 2012. Congratulations to team gb, but why should we be so surprised? olympic medal count can be predicted using logit regression models that include 'home advantage'.

[Passi and Pandey(2018)] Passi, K., Pandey, N., 2018. Predicting players' performance in one day international cricket matches using machine learning. Computer Science and Information Technology URL: https://api.semanticscholar.org/CorpusID:57356317.

[Pischedda(2014)] Pischedda, G., 2014. Predicting nhl match outcomes with ml models. International Journal of Computer Applications 101, 15–22. URL: https://api.semanticscholar.org/CorpusID:949008.

[Sagala and Ibrahim(2022)] Sagala, N.T., Ibrahim, M.A., 2022. A comparative study of different boosting algorithms for predicting olympic medal, in: 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), IEEE. pp. 1–4.

[Schlembach et al.(2022)Schlembach, Schmidt, Schreyer and Wunderlich] Schlembach, C., Schmidt, S.L., Schreyer, D., Wunderlich, L., 2022. Forecasting the olympic medal distribu-

tion – a socioeconomic machine learning model. Technological Forecasting and Social Change 175, 121314. URL: https://www.sciencedirect.com/science/article/pii/S0040162521007459, doi:https://doi.org/10.1016/j.techfore.2021.121314.

[Shailaja(2020)] Shailaja, V., 2020. Predictive analytics of performance of india in the olympics using machine learning algorithms. International Journal of Emerging Trends in Engineering Research URL: https://api.semanticscholar.org/CorpusID:225831191.

[Zhang et al.(2024)Zhang, Zhou and Bai] Zhang, J., Zhou, L., Bai, W., 2024. Innovative methods for integrating translation memory and cat tools: Enhancing intelligent support in human translation processes. Applied and Computational Engineering 90, 1–7. doi:10.54254/2755-2721/90/2024MELB0054.