Beyond the Podium: Medal Counts Prediction, Great Coaches and Host Effect*

S. Tan†[1], M. Feng[1], and J. Xu[1]

[1]Shanghai University

Jan 2025 New!

Abstract

"The Olympic Games are a platform for nations to showcase their strength, unity, and the human spirit, transcending politics and differences," said Thomas Bach. Undoubtedly, medals and effective Olympic strategies are crucial for nations. Can the final medal counts be predicted? To effectively address this question and offer valuable guidance for national Olympic committees, we present the following insights.

For Task 1, it is requisite to establish a model to predict the medal counts for each country in the Los Angeles, USA summer Olympics in 2028, while also exploring the impact of specific events on a nation's total medal count. We propose an improved Random Forest Model that includes both training and test datasets, which reflects a country's medal count based on the performance of its athletes. We calculated the correlation coefficients to identify several features highly correlated with medal counts. Additionally, we account for changes in the athletes' composition by classifying them as Continuing Athletes

*This paper is a revision of the paper submitted by the three authors when they participated in MCM during January 23-27 2025. The copyright of the article before it is officially published still belongs to the three authors. More information, please visit COMAP official website: click.

†Email: sepinetam@gmail.com, Address: No. 333, Nanchen Road, Shanghai, China.

and New Athletes, which facilitates a more accurate reflection of individual medal performance, thereby making the features fed into the Random Forest Model more precise. We finally use a table to provide visual representation of the 2028 medal table, and the list of countries with their first-time medalists is presented. Furthermore, we identify several key events where countries have a significant advantage, and calculate their impact in that table.

For Task 2, we need to explore the ”great coach” effect on a country's medal count and determine whether certain countries should prioritize the development of specific sports. Owing to the free movement of coaches, a Difference-in-Differences (DID) model was constructed to explain the effect of great coaches, using the coaching careers of Lang Ping, Béla Károlyi, and Jon Urbanchek as key examples. The regression results indicate that the arrival of a great coach significantly increases the likelihood of a team winning between 0.5 to 1.5 additional medal tiers, while the departure of a great coach significantly reduces the team's chances of winning medals.

For Task 3, it requires us to find and explain the unique and insightful perspectives proposed by the model. Based on the results of the previous questions, a significant host-country effect was found, and we further explored it. We still used the Random Forest Model and filtered the data from 1960 onwards. The data were divided into two datasets for model training: one with $Host_{c,y} = 1$ and the other with $Host_{c,y} = 0$. We also put forward some effective perspectives, such as attaching importance to first-time participation and emphasizing the continuation of historical medals.

After that, in order to test our model, we compared our Random Forest Model with the traditional Ordered Logistic Regression Model with the same dataset. The outcomes indicate that our model exhibits a high level of robustness.

Keywords: Random Forest Model, Difference-in-Differences, Medal Prediction, Great Coach, Host Effect

1 Introduction

1.1 Background

In recent years, the Olympic medal table has become a focal point of global attention. The performances of traditional sporting powerhouses such as the

United States, China, and Australia, in particular, have drawn significant interest. However, the achievements of athletes from other nations on the Olympic stage should not be overlooked, as their results often become topics of lively discussion.

Medal prediction projects are commonplace, yet these analyses typically focus on individual athlete data, such as past performance in their respective disciplines and their current form. However, a country's total medal count at the Olympics is influenced by a broader set of factors. Beyond the abilities of individual athletes, hosting the Games is a critical determinant, as host nations often leverage the home-field advantage to achieve better results. Additionally, the "great coach" effect cannot be underestimated; the presence of renowned coaches can significantly enhance the performance of entire teams. National policies and public opinion supporting sports also play a pivotal role, particularly in specific disciplines, thereby indirectly affecting the overall medal tally.

Medal counts serve not only as a key metric of a nation's athlete training level but also as a tangible reflection of its sportsmanship. Thus, accurately predicting and analyzing medal trends holds considerable importance. By delving into these influencing factors, we can better understand the current state and future trajectory of sports development in various countries. Such insights also offer valuable guidance for formulating more effective sports development strategies. This not only enhances a nation's competitiveness in international sporting events but also promotes public health through increased participation in sports, fosters enthusiasm for physical activity, and facilitates the widespread dissemination of sports culture.

1.2 Restatement of the Problem

In light of the background information and constraints highlighted in the problem statement, we need to tackle the following tasks:

1. Based on athletes' historical data, establish a mathematical model for medal allocation for each country and conduct an accuracy test to evaluate the predictive power of the model. Use the calibrated model to predict the medal standings for the 2028 Los Angeles Olympics in the United States. Compare the number of medals from the 2028 Olympics with those from 2024 to determine whether countries have achieved breakthroughs or experienced declines in performance. Additionally,

for countries that have never won a medal, predict the total number of medals they might win in 2028 and the probability of achieving this.

2. The model analyzes the impact of the “great coach” factor on the distribution of Olympic medals. It selects three countries to evaluate the potential medal enhancement effects that could result from introducing elite coaches in specific sports.

3. Evaluate the strengths and areas for improvement of the model, and assess its overall performance.

1.3 Work Flow

Our approach consists of four modules: constructing two core models to address the issue and comparing their results with the OLR model to validate the model’s performance and plausibility, whcih is conducted as Figure 1 shows.

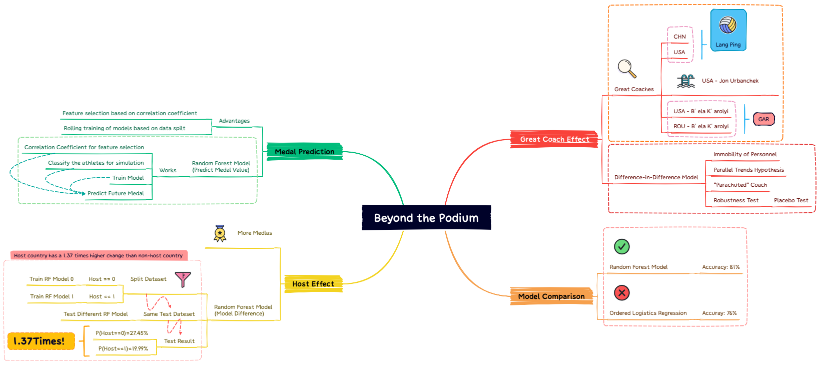

Figure 1: Workflow of the Model

2 Preparations of the Models

2.1 Assumptions and Explanations

- Assume that the athletes’ performance does not deviate significantly from the predicted values, and that exceptional outcomes, such as “upsets,” are not considered.

Explanations: In most cases, athletes demonstrate consistent performance, with relevant data being relatively concentrated, making it a reliable reference for predictions.

- It is assumed that there are no sharp fluctuations in the number of athletes sent by each country, i.e., the number of athletes will not experience significant increases or decreases.
  Explanations: This assumption ensures that the model remains stable and avoids large-scale disruptions in the number of athletes, which could otherwise lead to skewed results

- It is assumed that environmental factors, such as weather and venue conditions, do not have a significant impact on athletes' performance.
  Explanations: While environmental factors can influence competition outcomes, these factors are not included in the model due to limitations in available data.

- It is assumed that historical data provides stable predictive power for future performance.
  Explanations: In most cases, athletes' performances are influenced by their training and past experiences. Therefore, historical data is considered to be a stable predictor of future competition outcomes.

2.2 Notations

2.3 Data Pre-processing

- The attached data includes detailed information on all Olympic Games in history, such as medal distributions, hosting details, event information, and comprehensive data on participating athletes, which includes their years of participation, countries, events, and the types of medals they won. We merged and cleaned the data.

- We converted textual data into numerical form. For instance, medal achievements are represented by values from 0 to 3, and other awards are also numerically coded.

- We identified the years when "outstanding coaches" appeared in the dataset for further analysis, facilitating breakpoint regressions at those points.

Table 1: Summary of Key Indicators and Their Explanations

Indicators	Explanations
i	The athlete i
$c(i)$	The country c of the athlete i
P_i	The probability of athlete i winning a medal
$I(P_i)$	Whether the athlete i winning a medal
τ^+	The probability threshold for winning a medal
r	The correlation coefficients of different features
$Medal$	The rank of the medals
CAR	Continuing athletes rate
NAR	New athletes rate
$APE_{i,y}$	The individual features of athlete i in year y
$CE_{c,y}$	The country-level features of country c belongs in year y
$year^-$	A set of years the Olympic Games held before $year$

3 Model Overview

3.1 Improved Random Forest Model

Random Forest is an machine learning algorithm that makes classification or regression predictions by taking the majority vote or average of the results from multiple decision trees[1]. A Random Forest Model has been established for Olympic medal predictions, which extends the classic forests developed by Breiman[2]and Wager and Athey[3].

The partitioning method prevents data leakage by properly partitioning the training data and has the following advantages:

1. It allows for the prediction of future data based on historical training data.

2. During model training, different countries are separated to avoid data mixing, which may cause bias in the results.

3. By incorporating additional variables, it helps to distinguish between different sports or events.

We have organized the flowchart of the model as Figure 2.

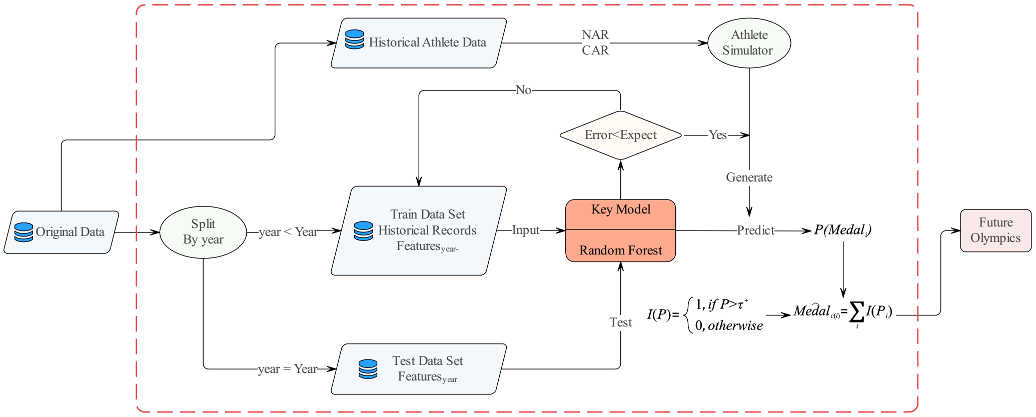
Historical Athlete Data
NAR
CAR
Athlete Simulator
No
Error<Expect
Yes
Generate
Original Data
Split By year
year < Year
Train Data Set Historical Records $\text{Features}_{\text{year-}}$
Input
Key Model
Random Forest
Predict
$P(Medal_i)$
Future Olympics
$I(P)=\begin{cases}1, \text{if } P>\tau^* \\ 0, \text{otherwise}\end{cases}$
$\widehat{Medal}_{c(i)}=\sum_i I(P_i)$
Test
year = Year
Test Data Set $\text{Features}_{\text{year}}$

Figure 2: Workflow of Random Forest Model

3.2 Difference-in-Differences Model

Difference-in-Differences (DID) is a widely used quasi-experimental research method for identifying causal effects by comparing changes in outcomes between treatment and control groups before and after an intervention. The key assumption of this method is the “parallel trends assumption,” which states that, in the absence of the intervention, the treatment and control groups would follow similar trends over time. Given the context of this study, where the athletes participating in the competition are largely stable over a short period, but the coach may change, DID provides a viable approach to evaluate the “great coach” effect. By examining the differences between the two groups across different time periods, DID effectively accounts for pre-existing disparities between groups as well as macro-level or time-specific factors influencing both groups, leading to more robust causal estimates.

In this study, to address the second research question, we apply the DID method to compare the outcome differences between groups influenced by “great coaches” and those without such influence, focusing on changes before and after the intervention. First, we calculate the outcome gap between the two groups prior to the intervention. Then, we measure the gap after the intervention. By subtracting the pre-intervention difference from the post-intervention difference, we isolate external time trends and group-invariant factors, allowing us to estimate the net impact of “great coaches.” Based on these findings, corresponding strategies and recommendations can be proposed to inform policy formulation and management practices.

3.3 Ordered Logit Regression Model

Ordered Logit Regression (OLR) is a statistical technique designed for situations where the dependent variable is ordinal, meaning its categories follow a natural order but the distances between categories are not necessarily equal. This method assumes the existence of an unobserved latent variable y^* that represents the underlying propensity or tendency associated with the outcome. The observed variable y is derived from y^* through thresholds that partition the latent variable into ordered categories[4].

The primary goal of OLR is to estimate the regression coefficients that explain the relationship between the independent variables and the latent variable y^*. By incorporating the cumulative distribution function (CDF), the model computes cumulative probabilities for each category, which represent the likelihood that an observation falls within or below a specific category. This approach ensures that the ordinal nature of the dependent variable is preserved while capturing the influence of predictor variables[5].

OLR is widely used in fields such as social sciences, economics, and medicine, where ordinal outcomes frequently arise—for instance, levels of customer satisfaction, disease severity, or educational attainment. By modeling the cumulative probabilities, OLR provides a nuanced understanding of how changes in independent variables affect the likelihood of an outcome occurring in higher or lower categories, making it a powerful tool for analyzing ordered data[6].

4 Model Building and Processing

4.1 Prediction Model Based on RF

4.1.1 Correlation and Feature Selection

The logistic correlation coefficient is a metric based on the logistic regression model, designed to evaluate the relationship between predicted probabilities and actual classifications. The regression coefficients further quantify the marginal impact of each independent variable on the dependent variable, elucidating their contribution to the classification outcomes in terms of magnitude and direction.

As a preliminary step, the Pearson correlation coefficient between two datasets is calculated[7]. The formula for the Pearson correlation coefficient r is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{1}$$

Where X_i and Y_i represent the values of the i -th sample point for the variables X and Y , respectively. $\bar{X}$ and $\bar{Y}$ denote the sample means of the variables X and Y , respectively.

Additionally, the value of r ranges from [-1, 1].

Finally, a correlation coefficient matrix between multiple variables is generated, which is then transformed to obtain the Logit correlation coefficient matrix.

The Logit transformation is calculated using the following formula:

$$logit(r) = log\left(\frac{1+r}{1-r}\right) \tag{2}$$

Here,r represents the Pearson correlation coefficient.

Based on the results of the correlation coefficients[8], a Logit correlation heatmap is generated. The heatmap's horizontal axis represents the dependent variables, specifically whether a certain medal is achieved, while the vertical axis showcases the individual features of the athletes. Darker shades of the squares indicate stronger correlations.

It was observed that the variable “Gender” has an insignificant impact on medal types. Therefore, this variable will not be considered in subsequent research. Variables such as “Host” will be used as features for the Random Forest prediction in the following tasks.

4.1.2 Some Parameters and Formulas

To more clearly analyze the impact of age on Olympic athletes' retirement and the transition between veteran and new athletes, we categorize the athletes for each Olympic Games into two groups: Continuing Athletes and

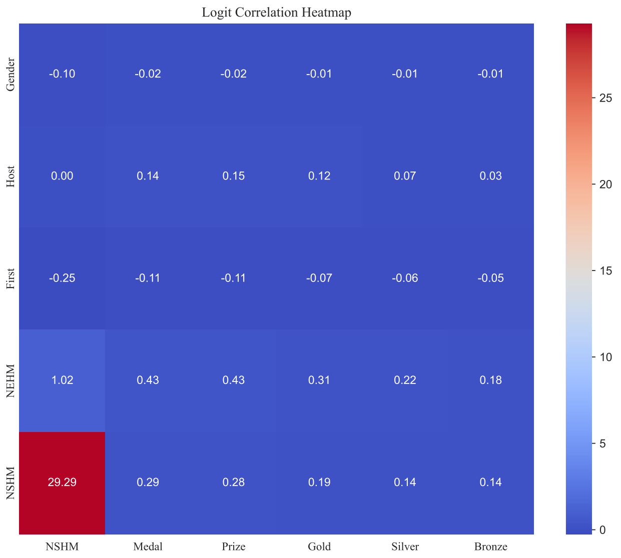

Figure 3: Logit Correlation Coefficient Heatmap for Medal Prize Distribution

New Athletes. This classification method helps to more accurately predict the potential contribution of athletes to the country's total medal count.

Let A_{y_1} denote the set of athletes participating in the Olympic Games in the y_1 - th year, A_{y_2} denote the set of athletes participating in the Olympic Games in the y_2 - th year, U denote the universal set of all possible athletes, and N_{y_1} denote the total number of athletes participating in the Olympic Games in the y_1 - th year. The number of continuing - participating athletes:

$$C = |A_{y_1} \cap A_{y_2}| \tag{3}$$

New athletes are those who participate in the $y_2 - th$ year but not in the y_1 - th year. The number of new athletes:

$$N = |A_{y_2} \cap \overline{A_{y_1}}| \tag{4}$$

The proportion of continuing - participating athletes:

$$CAR = \begin{cases} \frac{C}{N_{y_1}}, & N_{y_1} \neq 0 \\ 0, & N_{y_1} = 0 \end{cases} \tag{5}$$

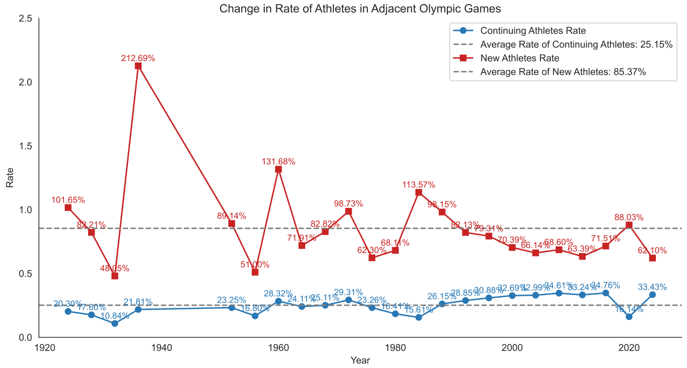

Figure 4: Change in Rate of Athletes in Adjacent Olympic Games

The proportion of new athletes:

$$NAR = \begin{cases} \frac{N}{N_{y_1}}, & N_{y_1} \neq 0 \\ 0, & N_{y_1} = 0 \end{cases} \tag{6}$$

The average proportion of veteran athletes $\overline{CAR}$ When considering $n-1$ groups of adjacent Olympic Games (since we are comparing adjacent two - year periods, there are $n-1$ groups), the formula for calculating the average proportion of veteran athletes $\overline{CAR}$ is:

$$\overline{CAR} = \frac{1}{n-1}\sum_{i=1}^{n-1} CAR_{i,i+1} = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{|A_{y_i} \cap A_{y_{i+1}}|}{|A_{y_i}|} \tag{7}$$

The average proportion of new athletes $\overline{NAR}$ Similarly, the formula for calculating the average proportion of new athletes $\overline{NAR}$ is:

$$\overline{NAR} = \frac{1}{n-1}\sum_{i=1}^{n-1} NAR_{i,i+1} = \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{|A_{y_{i+1}} \setminus A_{y_i}|}{|A_{y_i}|} \tag{8}$$

According to the formula above,we calculate that the Continuing Athletes Rate (CAR) is 25%, while the New Athletes Rate(NAR) is 85%. Based on the initial assumptions and comparisons with the correlation coefficients, an athlete's medal performance is determined by a set of features. To simulate the sample size of athletes participating in the 2028 Olympics, we proceed with random resampling through the following steps:

- *100x* is defined as the total number of athletes participating in the last Olympic Games, and $CAR \cdot x$ athletes are randomly selected from *100x* individuals to form a continuous sample of athletes.*100x* represents the number of athletes who participated in the previous Olympic Games.
- The remaining *85x* individuals are randomly assigned as the sample of New Athletes.
- The two samples are combined to form the total sample set of $CAR \cdot x + NAR \cdot x$.

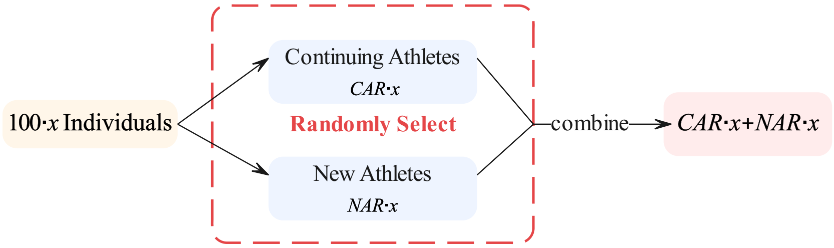

Figure 5: Workflow of Athlete Sampling

This method allows us to simulate the expected composition of athletes for the 2028 Olympics.

To improve the realism and accuracy of the simulation, we introduce the following constraints during the sampling process:

Participation Experience Control: Based on whether an athlete is participating in the Olympics for the first time, it will determine whether they can be selected as a Continuing Athlete or a New Athlete.

Medal History Impact: Historical data on whether the country has won medals in a specific event will be used as a reference condition, influencing the distribution of medal potential among new and veteran athletes.

We conduct n random repeated experiments based on the previous year's data. Define

$$N_{athletes} := CAR \cdot x + NAR \cdot x \tag{9}$$

The final output is an $n \times$ Leng matrix, where each row represents the simulation result of a single experiment.and each value indicates the medal contribution of an athlete. For example, when *med=0* , the new combination for "value" is an array of size n, with each value being the sum of the simulation results where med = 0 . This process is similarly applied for other values of med .

The individual medal outcomes are used as feature values x for input into the Random Forest Model. Based on the statistical data of athletes' different results, we aggregate and generate predictions for the total number of medals at the national level.

The individual medal outcomes are used as feature values x and are input into the Random Forest Model to construct multiple decision trees.

$$P_i^{Medal} = RF(APE_{i,year^-}, ACE_{c(i),year^-}) + Host_{c(i)} + \varepsilon_{i,c(i),year} \tag{10}$$

$year^-$ represents a set of years the Olympic Games held before $year$
To calculate the mean of a large number of predicted results, we use:

$$M\hat{e}dal_{c(i)} = \sum_i I(P_i) \tag{11}$$

where $I(P)$ is a threshold function that determines whether a medal is obtained based on the probability P and a threshold τ^+. Specifically,

$$I(P) = \begin{cases} 1, & \text{if } P > \tau^+, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Based on the output of the RF, the different results for the athletes are counted and the data is summarized to generate the predicted total number of medals at the country level.

4.1.3 Model Training

Considering that the size of the training set can influence the accuracy of the model, an optimal training set size is identified through an analysis of different errors corresponding to various models, where the error is represented by ε in formula 10.

To demonstrate that models trained on overly small samples tend to have significant errors, scatter plots of the sum of errors and the sum of squared errors are presented in Figure 6, respectively.

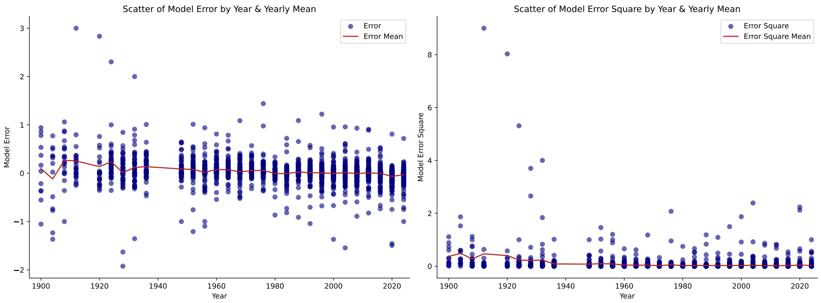

Figure 6: Error Scatter and Fitness Line by Year

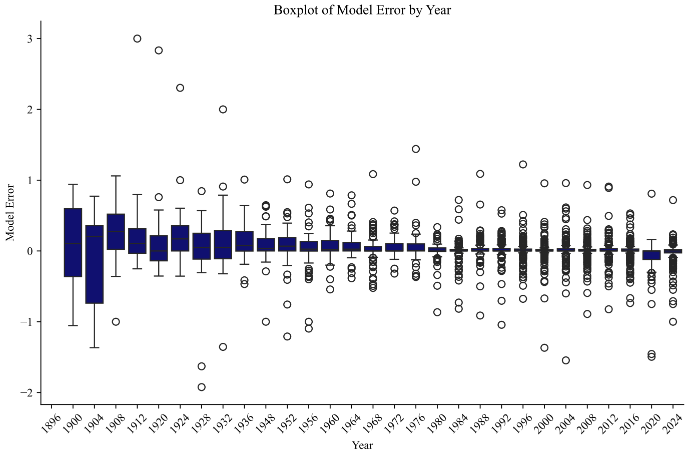

Figure 7: Boxplot of Model Error by Year

However, these do not provide evidence of the effect of larger sample sizes on errors. Therefore, boxplots were created to observe the distribution of model errors under different training set sizes, as shown in Figure 7.

Based on these figures, it can be observed that the optimal training dataset size is approximately 15 cycles[9;10]. The model is trained using data from 1960 to 2024. Simultaneously, predictions regarding the number of participants are also based on data from 1960 onwards.

4.2 Effect Evaluation Based on DID

4.2.1 Fundamental Principle

The traditional Difference-in-Differences model has been widely applied in policy evaluation[11;12] and has subsequently been employed to assess individual effects[13]. This provides a theoretical foundation for using the DID approach to evaluate the great coach effect.

We use the coaching experience data of Lang Ping, Béla Károlyi, and Jon Urbanchek, identified through the IOC Coaches Lifetime Achievement Awards, to conduct a DID regression analysis on the changes in the results of related events in different countries. Lang Ping coached the Chinese women's volleyball team before moving to the U.S. 2008 and later returned to coach China again. Béla Károlyi, originally from Romania, became the national gymnastics coach for the U.S. after his defection. Jon Urbanchek served as the USA synchronized swimming coach from 1982 to 2005.The parallel trends assumption allows us to use DID Model to estimate the degree of the "great coach" effect.

The first dimension reflects the difference over time. As time progresses, the performance of different athletes or teams may improve. The calculation formulas is:

$$\Delta Treat = \overline{Y}_{Treat,Post} - \overline{Y}_{Treat,Pre} \tag{13}$$

The second dimension represents the difference in coaching effects. Different coaches have varying impacts, with the treatment group representing teams influenced by the coach and the control group representing teams not under the coach's influence. This is calculated as:

$$\Delta Control = \overline{Y}_{Control,Post} - \overline{Y}_{Control,Pre} \tag{14}$$

Therefore, the regression model is as follows:

$$Y_{i,t} = \beta_0 + \beta_1 \cdot Treat_i + \beta_2 \cdot Post_t + \beta_3 \cdot (Treat_i \times Post_t) + \epsilon_{i,t} \tag{15}$$

In the model, $Y_{i,t}$ represents the medal performance of athlete i in the Olympics during year t, $Treat_i$ is a dummy variable indicating whether the athlete or team is coached by a great coach taking the value of 1 if the athlete or the team is coached by a great coach and 0 otherwise, and $Post_t$ is a time dummy variable, taking the value of 1 for the years after the great coach began coaching. The term $Treat_i \times Post_t$ represents the interaction between $Treat_i$ and $Post_t$. The coefficient β_3 measures the great coach effect.

4.2.2 Data and Model Test

To analyze the impact of outstanding coaches on national performance, we first focus on Lang Ping using the Olympic medal data of the Chinese women's volleyball team during her tenure. We then incorporate her coaching experience in the United States to separately evaluate her influence on team performance in both countries. Similarly, we apply the same methodology to Béla Károlyi and Jon Urbanchek. For Béla Károlyi, we examine his coaching records in Romania and the United States gymnastics teams, conducting cross-country and longitudinal comparisons. For Jon Urbanchek, we analyze his tenure and departure from the USA synchronized swimming team, comparing performance metrics before, during, and after his coaching period. Considering the constraints imposed by athlete nationality, we utilized only the historical medal data of specific countries rather than data from all countries when conducting the DID analysis. The parallel trends assumption in our model allows us to use DID to estimate the magnitude of the "great coach" effect.

In the placebo test for the DID model, we examined two dimensions:

Time Dimension: Select a time period prior to the coach's influence and assume the effect had already occurred during that period, testing whether a significant effect is observed.

Group Dimension: Choose teams that were not affected by the coach and assume they were virtually influenced, testing whether a significant change can be detected.

Based on our test results, the estimated β is not significant, indicating that the model inference is robust and the DID results are credible.

5 Model Solving

5.1 Medal Prediction

Using the trained Random Forest Model, the number of gold medals and the overall medal rankings for the 2028 Summer Olympics in Los Angeles, USA, are predicted. The results are presented in Table 2.

Table 2: Los Angeles Olympics (2028, Predicted) Final Medal Table and Historical Sessions

NOC	Rank	$Gold^*_{2028}$	$Silver^*_{2028}$	$Blonze^*_{2028}$	$Total^*_{2028}$	$Total_{2024}$	$Total_{2020}$	$Total_{2016}$
USA	1	55	43	46	144	126	113	121
CHN	2	44	30	32	106	91	89	70
GBR	3	27	29	30	86	65	64	67
JPN	4	25	29	29	83	45	58	41
AUS	5	19	33	12	64	53	46	29
ITA	6	12	10	22	44	40	40	28
FRA	7	10	20	13	43	64	33	42
GER	8	15	10	16	41	33	37	42

Based on the comparison of the 2028 data with the medal counts from the previous three Olympic Games, the medal tally shows a significant increase for most countries. However, France (FRA) exhibits a decreasing trend compared to the 2024 Olympics.

We observe that historical factors such as wars have contributed to the rise and fall of nations in the Olympics. We have compiled statistics on how many countries won their first medals in each Olympic Games and the total number of different types of medals they achieved.

From the Table 2, we can see that the United States is predicted to win the most medals in the 2028 Olympics, with a total of 144 medals, including 55 gold, 43 silver, and 46 bronze. China is expected to rank second with 106 medals, followed by Great Britain, Japan, and Australia. In terms of growth, the United States, as the host country, is expected to increase its total medal count by 18 compared to the 2024 Olympics. However, the last host country, France, is expected to experience a decrease in total medals by 21 compared to the 2024 Olympics. It is an obvious host effect, which will be discussed in Section 5.3.

Based on the above data, we filtered the medal-winning records of countries after 1960, focusing on nations that did not win any medals between 1960 and 2024. Using the Random Forest method, we predicted the prob-

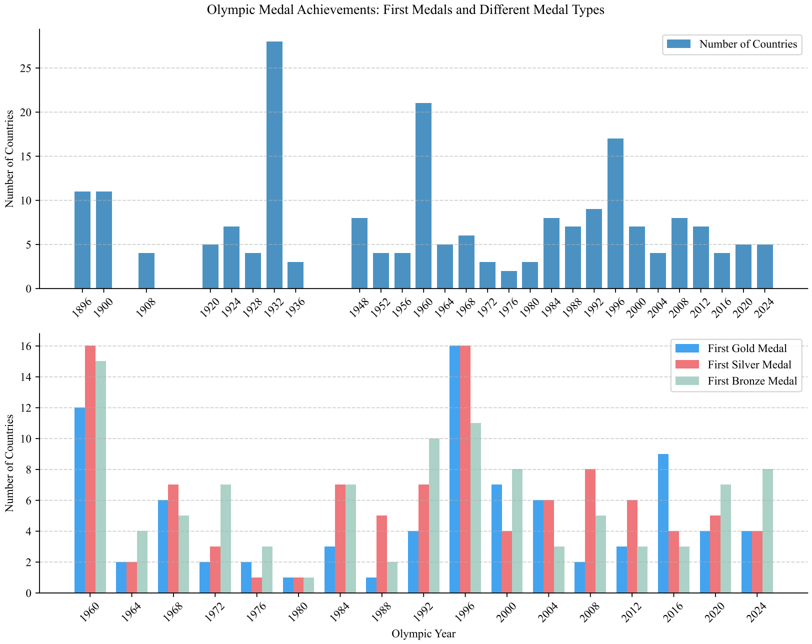

Figure 8: Olympic Medal Achievements: First Medals and Different Medal Types

ability of these countries winning at least one medal in the 2028 Olympics. The results are presented in the Table 3.

Table 3: Probability of Winning Medals (Non-Medaling Countries)

NOC	Prob(Total)	Prob(Gold)	Prob(Silver)	Prob(Bronze)
DEN	81.82%	72.73%	90.91%	81.82%
RSA	75.76%	72.73%	81.82%	72.73%
CRO	60.61%	63.64%	54.55%	63.64%
SLO	60.61%	54.55%	63.64%	63.64%
NGR	45.45%	45.45%	45.45%	45.45%
POR	42.42%	45.45%	45.45%	36.36%

Also, in order to find out whether there is a strong correlation between country and sport, we have calculated the Nation-Sport Interaction Scores by the historical data sorted by the dataset offered.

The input data of the original Random Forest Model was modified by replacing individual athlete characteristics with sport-specific features as training and prediction inputs. To achieve this, we isolated different Nations and Sports, generating a unique interaction ID (NSid). The final Score represents the absolute advantage of a nation in a specific sport. The results are reported in the Table 4. A higher Score value indicates a greater proficiency of the Nation in that particular sport.

5.2 Great Coach Effect

Outstanding coaches often have a significant impact on athletes. Generally, since athletes can only represent their country based on their nationality in the Olympics, it can be assumed that there will be little short-term change in the athletes themselves. However, coaches are not constrained by nationality.

We selected the United States, China, and Romania as examples to study the influence of renowned coaches, such as Lang Ping, Béla Károlyi, and Jon Urbanchek, on the performance of their respective national teams.

To further assess the robustness of the estimated results, a placebo test was conducted. This test compared the performance of different teams within

Table 4: Nation-Sport Interaction Scores: Evaluating Proficiency in Olympic Sports

NSid	**NOC**	**Sport**	**Score**	**Rank**
3486	SUI	Aeronautics	100.00	1
1178	CHN	Table Tennis	98.00	2
1542	ESP	Basque Pelota	79.50	3
3487	GER	Alpinism	70.50	4
2384	USA	Basketball	64.03	5
625	CAN	Breaking	58.33	6
1364	FRA	Motorboating	57.14	7
4011	USA	Fencing	54.17	8
4074	VEN	Cycling BMX Freestyle	54.17	9

the same country and analyzed the data under the same DID framework. Specifically, a difference analysis was performed on the performance of different teams during the same time period.

Lang Ping, who served as the head coach of the Chinese women's volleyball team in 1995, resigned in 1998. She then coached the U.S. team from 2005 to 2008, leading them to win the gold medal in 2008. In 2013, she returned to the Chinese team and successfully revived the long-struggling squad, guiding them to another championship. From the regression results, it can be seen that Lang Ping's arrival in China increased the probability of the Chinese team winning medals by a factor of 0.483. Meanwhile, her coaching tenure with the USA team resulted in a 1.555-fold increase in the likelihood of winning a gold medal. Béla Károlyi's coaching career began with Romania in 1976, and he later coached the USA starting in 1981. The DID coefficients for these two periods were 0.442 and 0.529, both statistically significant at the 99% confidence level. Jon served as the head coach of the USA synchronized swimming team before their first participation in the 1984 Olympics. After leaving the national team in 2005, the performance of USA synchronized swimming significantly declined, with the level of awards won dropping by one and a half tiers.

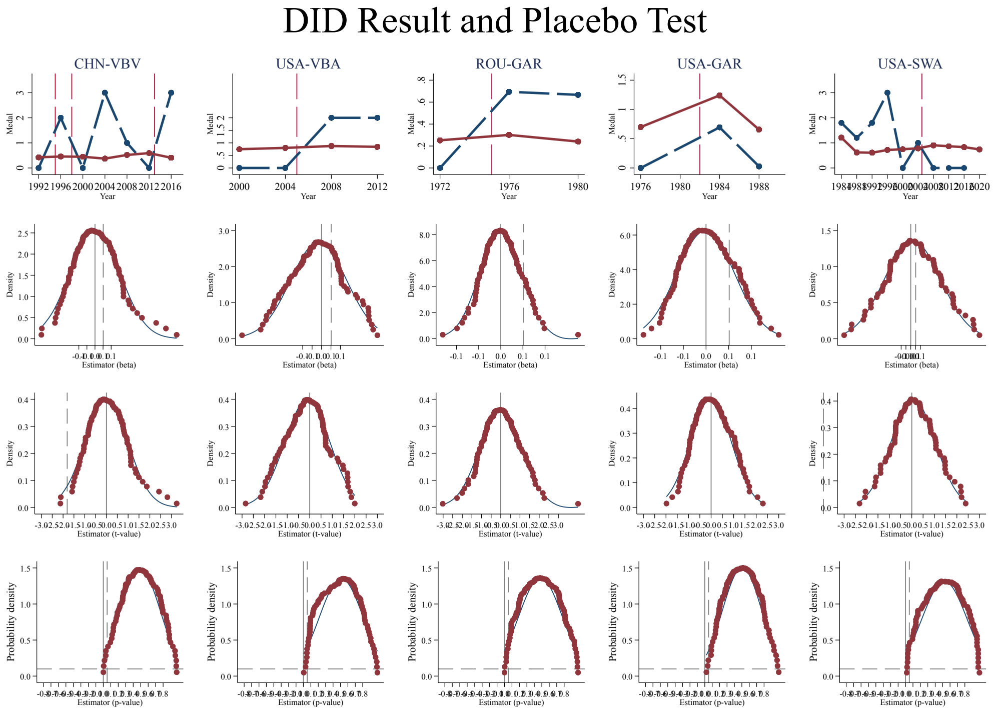

Figure 9: DID Result and Placebo Test

According to the results in Figure 9, the placebo test did not reveal any significant DID coefficients, indicating that the influence of outstanding coaches is indeed significant. The placebo test is passed, further validating the reliability and robustness of the subsequent analysis results.

The significance of excellent coaches for the sports development of various countries is self - evident. During Lang Ping's tenure as the coach of the US volleyball team, she not only increased the probability of winning gold medals but also introduced advanced concepts and tactics, thus promoting the long - term development of the sport. Jon designed a scientific training program that enhanced the competitiveness of the US synchronized swimming team on the international stage. Béla Károlyi, with a comprehensive training model, contributed to the rise of US gymnastics.

In China, Lang Ping coached the women's volleyball team twice, significantly improving the team's performance. Meanwhile, she established a stable development system, which promoted the popularization of volleyball.

Table 5: DID Regression Results

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Medal	Medal	Medal	Medal	Medal
Trade	-0.049	-0.357***	-0.112**	-0.699***	0.656***
	(0.211)	(0.118)	(0.057)	(0.069)	(0.169)
Post	0.010	0.070***	0.170***	0.080***	0.097***
	(0.039)	(0.0213)	(0.028)	(0.018)	(0.0212)
Trade × Post	0.483*	1.555***	0.442***	0.529***	-1.480***
	(0.271)	(0.190)	(0.072)	(0.098)	(0.344)
Constant	0.549***	0.732***	0.175***	0.726***	0.727***
	(0.030)	(0.010)	(0.023)	(0.013)	(0.010)
Coach	Lang	Lang	Béla	Béla	Jon
NOC	CHN	USA	ROU	USA	USA
Observations	2,869	16,774	3,994	16,774	16,774
R-squared	0.002	0.005	0.034	0.008	0.002

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In Romania, Károlyi used personalized training methods to elevate the level of gymnastics and reserve talented athletes. Excellent coaches can improve sports performance in the short term and drive the sustainable development of sports in the long run.

5.3 Host Effect

What has ever been reffered as before, the host effect is a significant factor that influences the performance of athletes and the medal of countries in the Olympics. The host effect is a phenomenon in which the host country of the Olympic Games achieves better results than usual. We attempt to use the Random Forest Model to quantify the host effect and determine the extent of this effect.

The flowchart for this part is shown in Figure 10. We selected data from countries that hosted the Olympic Games between 1960 and 2024. We trained separate Random Forest Models using data from athletes in the countries that hosted the Games in each respective year, as well as other relevant data. We then used 1,000,000 randomly generated data sets to predict the outcomes of the models.

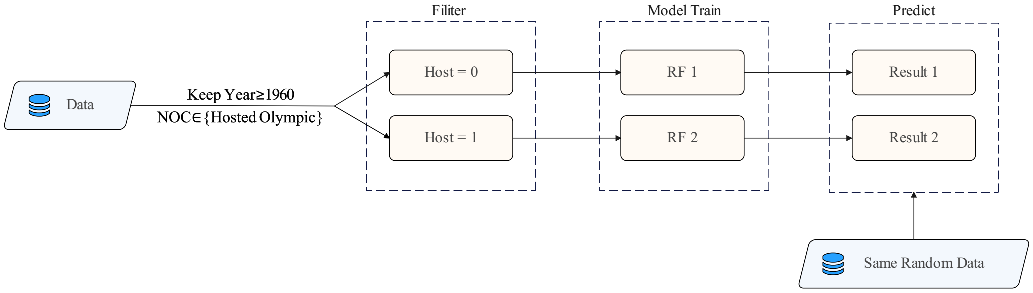

Figure 10: Workflow of Estimating the Host Effect

The final prediction results indicate a significant decline for France and a notable rise for the United States in 2028. A pronounced host country effect is observed in the model results, warranting further investigation.

Data from 1960 onwards were filtered, retaining only countries that had hosted the Olympics to prevent the overall national strength from influencing the outcomes. The filtered data were then divided into two datasets for model training: one with ($\boldsymbol{Host}_{c,y} = 1$) and the other with ($\boldsymbol{Host}_{c,y} = 0$).

Using the same randomly generated samples, we compared the models and made predictions. The results demonstrate that, on average, the probability of winning for host countries is approximately 1.37 times that of non-host countries. To validate this, we generated a new dataset containing one million samples and produced the following Table 6.

Table 6: Descriptive Statistics (Model Comparison)

	Count	**Mean**	**Std**	**Min**	**Q1**	**Mid**	**Q3**	**Max**
Model (Host = 1)	1,000,000	27.45%	13.78%	13.06%	13.06%	14.35%	40.08%	42.33%
Model (Host = 0)	1,000,000	19.99%	8.41%	10.81%	10.81%	13.02%	25%	31.17%
Times	-	1.37	-	1.21	1.21	1.10	1.60	1.36

6 Model Comparison

6.1 Ordered Logit Regression Result

The Ordered Logit Regression (OLR) model results indicate that factors such as whether it is the athlete's first time participating in the Olympics, whether the athlete's country is the host country, and whether the country has previously won medals in a particular sport have a significant impact on the Olympic medal rank.

Specifically, the negative coefficient for first-time Olympic participation (-0.2117) suggests that, compared to the baseline group, first-time participants perform worse in terms of medal rank. In contrast, the country being the host is positively correlated with higher medal ranks (0.7049), indicating that athletes from the host country are more likely to achieve higher medal ranks. The coefficient of 1.1686 for prior medal wins in the sport indicates that a country's previous success in a particular sport significantly increases the likelihood of its athletes achieving higher medal ranks. This also provides an explanation for the features fed into the Random Forest Model based on their direction.

6.2 Significant Differences

While it offers an explanation for the direction of the Random Forest Model, its performance does not match that of the Random Forest in terms of accuracy.

We utilized data from 2016 and earlier to fit both the Random Forest Model and the OLR model, and subsequently validated the models using data from 2020 and 2024. The performance evaluation of the models indicates that the Random Forest Model achieved an accuracy of 81%, while the traditional OLR model exhibited an accuracy of only 76%.

To assess the validity of the proportional odds assumption, several classical tests were performed, including the Wolfe-Gould test, Brant test, score test, likelihood ratio test, and Wald test. The chi-square values for all tests were significant, suggesting that the independent variables exert consistent effects across the different categories of the dependent variable, thereby confirming the validity of the proportional odds assumption in the ordered probability regression model.

To clearly visualize the model's difference, we plotted the confusion ma-

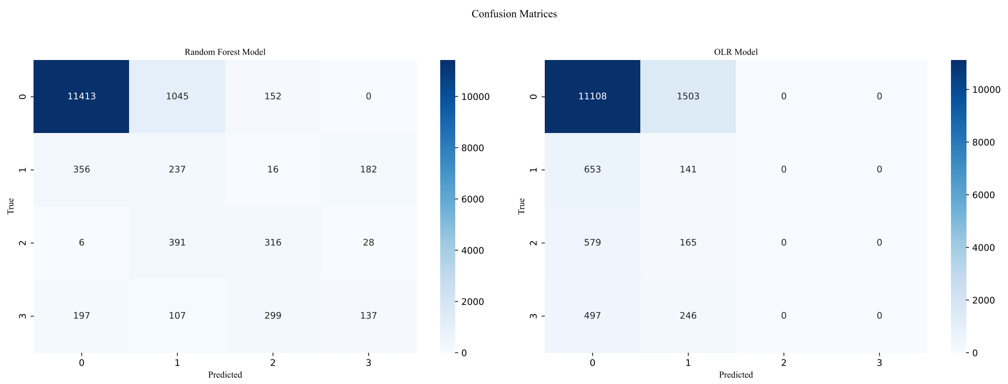

Figure 11: Confusion Matrices (Model Comparison)

trix, as shown in Figure 11. The confusion matrix reveals that the OLR model failed to predict any gold or silver medals. However, due to the large number of actual outcomes being no medals, its accuracy appears inflated, whereas in reality is not. In contrast, the Random Forest Model successfully predicted medals across all categories, with its accuracy reflecting genuine performance rather than inflated results. This further underscores the superiority of our Random Forest Model in comparison.

7 Strengths and Improvements

7.1 Strengths

- **Split-Based Data Segmentation:** To facilitate the observation of future data, we employ a split-based method for data segmentation. Using random splitting, the dataset is divided into training and testing sets. By comparing the model's predicted results with the known data, any observed deviation prompts a return to the training set for further calibration. This iterative process allows for multiple validations of the model, thereby enhancing the accuracy of the final predictions.

- **Country-Specific Training:** During the model training process, data from different countries is separated, and independent models are trained for each nation. This approach avoids the mixing of data across countries, ensuring more precise predictions while better capturing the unique characteristics of each country.

- **Athlete Classification:** We categorize athletes into two groups: new athletes and those with sustained participation. This classification takes into account factors such as the athlete's competition experience and age, which significantly affect their performance in competitions. By incorporating these factors, the model can more accurately reflect the dynamic realities of athlete performance.

7.2 Improvements

- The model exhibits a significant effect from the host country, which can be further improved. For example, France, as the host for the 2024 Olympics, saw a notable decrease in medal counts when it was not the host in 2028. Similarly, the United States experienced a marked increase in medal counts after hosting in 2028.

- The variation in Olympic medal counts is influenced by various factors such as GDP and national policies. However, due to constraints in the scope of the research question and the available data, the model lacks sufficient variables. Only data directly related to sports competitions were used, which limits the generalizability of the results. Future improvements will involve adjusting the model parameters to account for changes in policies and other relevant factors.

8 Conclusion

In this study, we first developed an improved Random Forest method to address the first task. Our improved Random Forest Model offers several notable advantages. It employs a partition-based data segmentation approach to divide the dataset into a training set and a test set, followed by iterative calibration and validation. This process significantly enhances the accuracy of our predictions. The country-specific training method, which trains independent models for each country, effectively captures the unique characteristics of different countries, avoids data mixing, and provides more precise forecasts. Moreover, we classified athletes into continuing athletes and new athletes, taking into account key factors such as competition experience and age. This classification enriches the input features of the model, enabling it to more accurately reflect the dynamic nature of athletes' performance and ultimately contributing to more accurate medal-count predictions.

Compared with the traditional Ordered Logit Regression Model (OLR), our Random Forest Model demonstrates clear superiority. In the model performance evaluation, where data from 2016 and earlier were used for training and data from 2020 - 2024 were used for validation, the Random Forest Model achieved an accuracy of 81%, outperforming the OLR model's 76% accuracy. The confusion matrix further shows that the Random Forest Model can successfully predict medals in all categories, while the OLR model fails to accurately predict gold and silver medals.

This study also employed the Difference in Differences (DID) model to evaluate the impact of "great coaches." By using data on the coaching experiences of renowned coaches such as Lang Ping, Béla Károlyi, and Jon Urbanchek, we were able to estimate the influence of these coaches on a country's medal counts in specific sports. The results of the DID regression analysis and the passing of the placebo test confirm the significant impact of excellent coaches on sports performance. This not only provides valuable insights into the role of coaches in sports development but also offers practical guidance for countries to improve their sports performance through coaching strategies.

Furthermore, our model identified and verified the host effect. Based on the data we processed, the probability of the host country winning medals is significantly higher, with the average winning probability being approximately 1.37 times that of non-host countries. This finding not only validates the importance of considering the host factor in medal-count predictions but also provides a reference for future Olympic organizers and participating countries. In summary, our model has achieved favorable results in predicting Olympic medal counts and analyzing influencing factors. It can offer valuable guidance for national Olympic committees in formulating strategies, allocating resources, and predicting competition outcomes.

References

[1] JunHao Chen, XueLi Wang, and Fei Lei. Data-driven multinomial random forest: a new random forest variant with strong consistency. *Journal of Big Data*, 11(1):34, February 2024.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[3] Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, July 2017. arXiv:1510.04342 [stat].

[4] Peter McCullagh. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 42(2):109–127, January 1980.

[5] Michael R Frone. Regression models for discrete and limited dependent variables. In *Research Methods Forum*, volume 2, pages 1–10, 1997.

[6] Alan Agresti. *Analysis of Ordinal Categorical Data.* Wiley Series in Probability and Statistics. Wiley, 1 edition, March 2010.

[7] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, 2018.

[8] Sheldon M. Ross. Chapter 12 - linear regression. In Sheldon M. Ross, editor, *Introductory Statistics (Fourth Edition)*, pages 519–584. Academic Press, Oxford, fourth edition edition, 2017.

[9] Carlo Corinaldesi, Daniel Schwabeneder, Georg Lettner, and Hans Auer. A rolling horizon approach for real-time trading and portfolio optimization of end-user flexibilities. *Sustainable Energy, Grids and Networks*, 24:100392, 2020.

[10] Weilin Hou, Zhaoxi Liu, Li Ma, and Lingfeng Wang. A real-time rolling horizon chance constrained optimization model for energy hub scheduling. *Sustainable Cities and Society*, 62:102417, 2020.

[11] Orley Ashenfelter and David Card. Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs, November 1984.

[12] David Card and Alan Krueger. Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania. Technical Report w4509, National Bureau of Economic Research, Cambridge, MA, October 1993.

[13] Jonathan Roth, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244, 2023.