# Models for Predicting Olympic Medal Tables

Xiaobei Wang[1], Yang Sun[2], Zuocheng Zhang[3]

[1] Jinan University, Guangzhou, 511443, China;

[2] The University of Sydney, Sydney, 2006, Australia;

[3] Simon Fraser University , Burnaby, V5A 1S6, Canada;

Xiaobei Wang, Yang Sun, Zuocheng Zhang are co-first authors

**Abstract.** During the 2024 Paris Olympics, the medal table attracted intense attention, and there was a surge of enthusiasm for predicting the medals of the 2028 Olympics. This study focuses on constructing a model to indicate the number of Olympic medals, aiming to explore the factors influencing a country's performance in winning Olympic medals and make accurate predictions. First, we conducted rigorous data cleaning operations on the collected data to ensure that the data input into the model met the requirements. Then we adopted a two-stage model construction approach to predict the number of medals these countries could obtain. The virtual medal prediction leaderboard shows that Bangladesh and Benin might win their first Olympic medals in 2028. Next, we used the Kmeans++ clustering model to group similar countries together. We also analyzed the relationship between sports events and medal counts. We identified their advantageous sports events and visualized the clustering results through data visualization. Overall, this model provides a practical tool for reasonably predicting the number of medals. By considering multiple factors, it offers insights for countries on how to win more medals.

**Keywords:** Olympic games; “Great coach” effect; Random Forest; XGBoost; Kmea.

## 1. Introduction

### 1.1 Problem Background

With the excitement and anticipation surrounding each Olympic Games reaching a fever pitch, the global audience, from sports enthusiasts to statisticians, is gripped by a profound curiosity. There is a significant interest in predicting the medal tallies of various nations and individual athletes. This isn't just a matter of idle speculation; it has far-reaching implications for sports marketing, national pride, and strategic sports planning.

The prediction problem, in its complexity, aims to leverage multiple layers of data. Therefore, this article aims to construct a predictive model for medal counts for each country in the Olympic Games, focusing specifically on gold medals and total medals. This model should incorporate measures to estimate the uncertainty and precision of its predictions, as well as metrics to assess its overall performance. Using this model, we need to provide projections for the medal standings in the 2028 Los Angeles Summer Olympics. For each country, include prediction intervals to indicate the range of possible outcomes. Then give additional unique perspectives and explain how it benefits the country's Olympic committees.

Sports are closely related to national prestige and international relations[1]. The fields of sociology and economics have been employed to forecast the quantity of Olympic medals a nation is likely to obtain. Past evil-related incidents have also had a significant impact[2]. Countries with larger populations and more abundant economic resources tend to perform better in the Olympics. Some studies have considered the economic determinants of Olympic performance and predicted the number of medals to be won at the Beijing Olympics[3]. Politically "unfree" countries generally perform better in the Olympics and win more medals[4]. In the early stages, research in this regard was carried out by scholars such as Ball in 1972, Grimes AR and others in 1974, and Levine N also in 1974[5]. Among them, the importance of causal forces in time-series extrapolation is applied[6]. The relationships between population, economic development level, and the number of Olympic

medals won are further studied[7]. Around 2000, Johnson and Ali, along with Bernard and Bussé, among others, advanced the development of specialized and systematic research on medal prediction.

## 2. Model Preparation: Data visualization and analysis

Figure 1-4 respectively illustrate the trend in the number of athletes participating in the Olympics over the years, the top 10 countries with the most Olympic medals, countries with noticeable declines in medal counts, and the year-by-year medal counts for the top 10 countries. Figure 1 highlights that the number has significantly increased, especially in the 21st century, stabilizing around 15,000 athletes. With the Olympics continuing to grow in global influence and participation, the number of athletes is expected to rise further, potentially surpassing 16,000 in 2028.

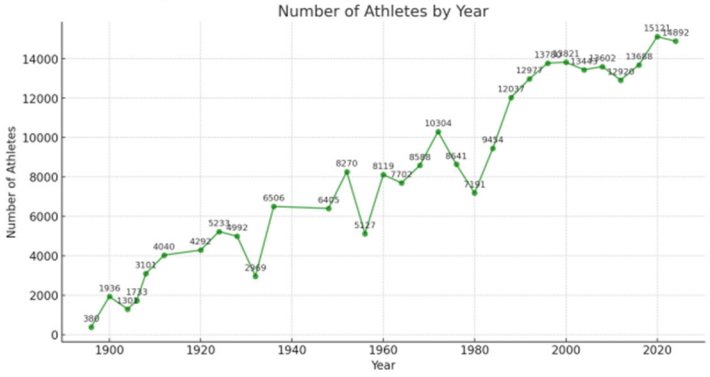

Fig. 1 Number of athletes by year

## 3. Two-Stage Medal Prediction Model based on PSO

### 3.1 Data Collection and Processing

To get extensive and accurate data, we carried out several data collection initiatives and rigorous data purification processes on the collected data. The data gathering and cleaning techniques are explained in full below.

**(1) Medal statistics**

**Country**: Document each country and region's involvement in every Olympic Games. These countries and regions play essential roles in the Olympics, and their medal accomplishments are one of the primary research topics.

**Year**: Clearly state the Olympic year to which the data corresponds, including the dates of all previous Olympic Games, to ensure that the temporal component of the data is well established.

**Types of medals**:

- **Gold**: Counting the number of gold medals won by each country in these Olympic Games is one of the most important markers of a country's competitive performance.
- **Silver**: The number of silver medals each country earns shows athletes' performance levels during contests.
- **Bronze**: The number of bronze medals won by each country adds to the thoroughness of the medal data.
- **Total**: The formula "Total = Gold + Silver + Bronze" determines and accurately depicts a country's overall medal result during the Olympic Games.

**Olympic host city:** Clearly identifying the host city for each Olympic Games is important not just geographically but also in terms of local sports culture, resources, and other characteristics. This can provide great insights for studying medal distribution patterns.

**Number of participating athletes**: The statistics on the number of athletes from each country participating in this Olympic Games reflect the scale of involvement and the level of importance each nation places on the event.

**(2) Sports event medal data**

Specific statistics on the number of medals in various sports, such as athletics and swimming, have been gathered. By studying medal counts across events, we can acquire insights into multiple

countries' strengths and weaknesses in different sports disciplines, revealing the intrinsic relationship between sports events and national medal distribution.

**(3) Athlete-related data**

The quantity and quality of national athletes: Assessing the size and competitive level of each country's athletic team, the number and quality of athletes are directly related to a nation's ability to secure medals at the Olympics.

Training facilities: Understanding the status of training facilities for athletes in various countries, advanced training facilities enhance training effectiveness and improve competitive performance.

### 3.2 Medal Prediction Model

For the model to accurately judge whether a country has won and how many medals it has won, it is necessary to pick the appropriate characteristics. Combined with the data dimensions collected earlier, this paper selects national population, GDP, sports investment, number of participating athletes, number and quality of national athletes, and sports medal data as characteristics. Taking whether or not a country has won medals and the number of medals won as dependent variables, the specific label construction is divided into two stages: the first stage is to use whether a country has won medals as a label, and the medals won are recorded as 1, and the number of medals won is recorded as 0, to construct the data structure of the binary classification problem. Phase 2: For countries judged to be able to win medals in the first stage, the actual number of medals won is used as a label to construct the data structure of the regression problem.

**Model Construction (Classification Model)**: A binary classification model uses a random forest to determine whether a country can win a medal.

The second stage of model construction (regression model): Based on the data of the countries judged to be able to win medals in the first stage, the XGBoost algorithm is used to build a regression model to predict the number of medals that these countries can win.

#### 3.2.1 Random Forest Classification Model

Random forest regression is an ensemble learning method that makes predictions by building multiple decision trees and averaging their results. Each decision tree is trained using a different data subset and built using a bootstrap (with put-back sampling) approach. The final prediction is the average of all decision tree predictions. The specific algorithm process steps are as follows:

**Step 1**: Data preparation and preprocessing

Collect datasets with traits and labels. Each sample contains features $\{x_1, x_2, \cdots, x_n\}$ and a categorical label y.

**Step 2**: Build multiple decision trees

**(1) Bootstrap sampling**: Sample from training set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ A to generate T training subsets $D^{(t)}$ : $D^{(t)} = \{\left(x_1^{(t)}, y_1^{(t)}\right), \left(x_2^{(t)}, y_2^{(t)}\right), \ldots, \left(x_N^{(t)}, y_N^{(t)}\right)\}, t = 1,2,\ldots,T$

**(2) Build each decision tree**: Each tree is split using randomly selected features. For the split selection of each node, use Gini Impurity: $Gini(t) = 1 - \sum_{k=1}^{C} p_k^2$, where $p_k$ is the proportion of samples belonging to category $k$.

Split point selection:

$$arg\min_{split} \sum_{i=1}^{k} \frac{N_i}{N} Gini(t_i) \quad (1)$$

where $N_i$ is the number of samples $t_i$ the split subset, and $N$ is the total number of samples.

**Step 3**: Integrate the results of multiple trees

The output of each tree is the prediction of category $y$, and the final classification result is a majority vote for all tree predictions:

$$\hat{y} = mode\left(f_1(x), f_2(x), \ldots, f_T(x)\right) \quad (2)$$

where $f_t(x)$ is the prediction for the $t$-th tree.

Step 4: Model evaluation

1. Accuracy:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i) \tag{3}$$

where $\mathbb{I}(\hat{y}_i = y_i)$ is the indicator function, taking 1 when the prediction is correct and 0 otherwise.

2. Precision:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

where TP is the true example and FP is the false positive.

### 3.2.2 XGBoost Regression Model

XGBoost (Extreme Gradient Boosting) is an optimized version of Gradient Boosting Decision Trees (GBDT), which is an effective ensemble learning method, especially for regression and classification problems. XGBoost has the advantage of improving model prediction by integrating multiple decision trees and optimizing it by gradient descent of the loss function. The algorithm process is as follows：

**Step 1**: Data preparation and preprocessing

Each sample contains the feature $\{x_1,x_2,...,x_n\}$ and the target variable $y \in R$。

**Step 2**: Building an Additive Model (Gradient Boosting Tree)

**(1) Initialize the model**: Initialize a constant value $F_0(x)$, which is usually the mean of the target variable:

$$F_0(x) = arg\min_{\gamma}\sum_{i=1}^{N} L(y_i, \gamma) \tag{5}$$

where $L(y_i, \gamma)$ is the loss function, and the commonly used mean square error (MSE):

$$L(y_i, \gamma) = (y_i - \gamma)^2 \tag{6}$$

**(2) Objective Function**: The objective function of XGBoost optimization consists of data loss and regularization terms:

$$L = \sum_{i=1}^{N} L\left(y_i, F(x_i)\right) + \Omega(F) \tag{7}$$

where the data loss $L\left(y_i, F(x_i)\right)$ is the mean square error (MSE):

$$L\left(y_i, F(x_i)\right) = \left(y_i - F(x_i)\right)^2 \tag{8}$$

The regularization term $\Omega(F)$ is used to control the complexity of the model:

$$\Omega(F) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \tag{9}$$

where $T$ is the number of leaf nodes of the tree, $w_j$is the weight of the$j$-th leaf, and $\gamma$ and $\lambda$ are the regularized hyperparameters.

**(3) Gradient calculations**: In each iteration, XGBoost uses gradient boosting to optimize the model. For each sample $i$, calculate the gradient of the loss function:

$$g_i = \frac{\partial L\left(y_i, F(x_i)\right)}{\partial F(x_i)} \tag{10}$$

where $g_i$ is the gradient of the i-th sample.

**(4) Construction of new trees**: The structure of the new trees is optimized by minimizing the objective function. The output of $h_m(x)$ per tree is updated from the current model $F_{m-1}(x)$:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \tag{11}$$

Among them, $\eta$ is the learning rate, which controls the impact of each tree on the final model. For each tree, the second-order Taylor expansion is used to approximate the objective function to obtain the weights of the leaf nodes:

$$w_j^{(m)} = -\frac{\sum_{i\in R_j} g_i}{\sum_{i\in R_j} h_i + \lambda} \tag{12}$$

where $R_j$ is the sample set of the $j$-th leaf node, and $g_i$and $h_i$ are the gradient and second derivatives of the $i$-th sample, respectively.

**Step 3**: Update the model

Minimize the objective function: The final model is updated through multiple rounds of iterations:

$$F(x) = F_0(x) + \sum_{m=1}^{M} \eta\, h_m(x) \tag{13}$$

where $M$ is the total number of trees.

**Step 4**: Model evaluation. We use MSE and RMSE to evaluate the performance of model.

### 3.3 Hyperparameter optimization

The particle swarm optimization (PSO) algorithm is used to optimize the key hyperparameters of the model (number of trees, maximum depth, minimum number of sample segments, and learning rate), and the algorithm process is as follows:

**1) Particle initialization**: Let the particle swarm size be P, the search space dimension be D (corresponding to the number of trees, the maximum depth of the hyperparameter combination), and each particle position represents a hyperparameter combination $\mathrm{x}_i^k = [u_i, \gamma_i, \eta_i, b_i, l_i]^T \in \mathbb{R}^D$,, where $i = 1,2,\dots,P$,$k$ is the current number of iterations. The particle velocity is initialized to $\mathrm{v}_i^0 = [v_{i1}^0, v_{i2}^0, \dots, v_{iD}^0]^T$, the position is initialized to $\mathrm{x}_i^0 = [x_{i1}^0, x_{i2}^0, \dots, x_{iD}^0]^T$, and the initial velocity and position are randomly generated by a uniform distribution:

$$v_{id}^0 \sim \mathcal{U}(-v_{\max}, v_{\max}), \quad x_{id}^0 \sim \mathcal{U}(x_{\min,d}, x_{\max,d})$$

where $d = 1,2,\dots,D$, $v_{\max}$ is the velocity limit, and $[x_{\min}, d, x_{\max,d}]$ is the value range of the $d$-dimensional hyperparameter.

**2) Fitness function**: The fitness function is used to measure the advantages and disadvantages of the corresponding hyperparameter combination of particles, and is defined as the loss value based on the validation set $\mathcal{L}_{\mathrm{val}}$:

$$f_i = \mathcal{L}_{\mathrm{val}}(\mathrm{x}_i^k) \tag{14}$$

where $\mathcal{L}_{\mathrm{val}}$: represents the average loss value (e.g., mean square error) on the validation set. The smaller the fitness, the better the performance of the hyperparameter combination.

**3) Velocity and Position Updat**e: The velocity and position update formula for particles is:

$$\begin{aligned} v_i^{k+1} &= \omega v_i^k + c_1 r_1 (p_i^k - x_i^k) + c_2 r_2 (g^k - x_i^k) \\ x_i^{k+1} &= x_i^k + v_i^{k+1} \end{aligned} \tag{15}$$

Among them, $v_i^k$ and $x_i^k$ are the velocity and position of particle $i$ in the $k$-th iteration, $\omega$ is the inertia weight, $c_1, c_2$ are the acceleration factors, $r_1, r_2$ are random numbers, $p_i^k$ and $p_i^k$ are the local and global optimal solutions of the particles, respectively.

**4) Termination condition**: stop iteration when the number of iterations reaches the maximum value $k_{\max}$ or the improvement of the $\mathcal{L}_{\mathrm{val}}$ of the loss value of the validation set is less than the preset threshold $\epsilon$, and the final output of the optimal hyperparameter combination is the global optimal solution $\mathrm{g}^{k_{\mathrm{opt}}}$, where the $k_{\mathrm{opt}}$ is the optimal number of iterations when the iteration is stopped.

### 3.4 Analysis of Results

#### 3.4.1 Model Performance

In the first stage of the prediction process, i.e., the training of different models, we use machine learning models to predict the number of medals. For the classification model, we included support vector machines (SVMs), decision trees, and random forests. After a series of tests and analyses, we found that the key indicator of accurately predicting the number of medals was being accurate. The random forest model performed significantly better than other models. This advantage is visually reflected in the area under the receiver operating characteristic (ROC) curve, i.e., AUC of 0.95 (see Figure 5 for details), which indicates that the model performs well in terms of classification accuracy.

In the second part of the model construction, regression, after several rounds of comparative experiments, the XGboost algorithm stood out among many algorithms and showed better performance. Based on the above results, we determined the final prediction model as a two-stage algorithm. Firstly, the random forest classifier is used to preliminarily process and classify the data, and then the XGboost regressor is connected to the XGboost regressor for further regression analysis, and the PSO algorithm is used to find the optimal hyperparameters, so as to give full play to the respective advantages of the two models and achieve more accurate prediction of the number of medals.

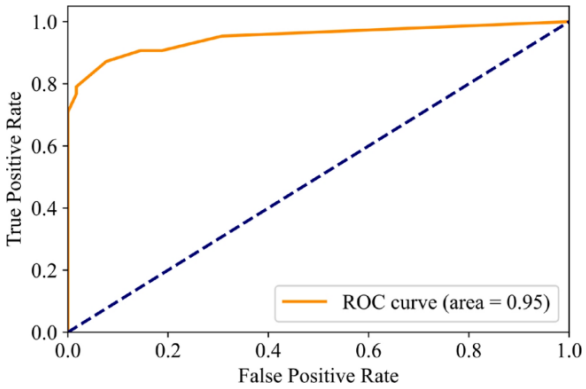

Fig. 2 ROC curve of random forest classifier

### 3.4.2 Medal Prediction Results

Based on the model we built, here are the predictions for the medal table for the 2028 Summer Olympics. We have predicted each country's medal count using a variety of data sources, historical performance, and related characteristics and provide a 95% prediction interval for each prediction, reflecting the range of possible fluctuations in medal counts. Table 1 shows the predictions for the top 10 finishes in the medal table.

Table 1 Predictions for the top 10 medals in the table

| Country | Prediction range for the number of gold medals | Total Medal Prediction Interval |
|---|---|---|
| America | 45 - 55 | 135- 145 |
| China | 37 - 42 | 83 - 94 |
| England | 16 - 24 | 58 - 67 |
| Japan | 12 - 17 | 45 - 55 |
| Australia | 14 - 19 | 55 - 65 |
| France | 18 - 25 | 60 - 70 |
| Netherlands | 16 - 22 | 35 - 45 |
| Britain | 16 - 23 | 60 - 70 |
| Germany | 13 - 18 | 35 - 43 |
| Italy | 11 - 16 | 40 - 50 |
| Canada | 8 - 13 | 15 - 25 |

The U.S. is strong in athletics, swimming, and many other sports, with a projected range of 45-55 gold medals and 135-145 total medals at the 2028 Olympic Games. With breakthroughs in traditional and emerging sports such as diving and table tennis, China is expected to win 37-42 gold medals, and the total number of medals may reach 83-94. Great Britain is competitive in water sports, athletics, and other areas, with a forecast of 16-24 gold medals and 58-67 medals. Japan has strength in gymnastics, swimming, and other events, with a possible 12-17 gold medals and a total of 45-55 medals. Australia has a strong advantage in swimming, with a forecast of 14-19 gold and 55-65 medals. As a sports powerhouse, France has a traditional advantage in cycling, fencing, and other events, with a forecast of 18-25 gold and 60-70 medals. The Netherlands is strong in speed skating, hockey, and other events, with 16-22 gold and 35-45 total medals. Germany has a strong sports heritage, with extraordinary strength in rowing, canoeing, and other events, with a forecast of 13-18

gold medals and a total of 35-43 medals. Italy is competitive in fencing, cycling, and other events, with 11-16 gold and 40-50 total medals. Canada has a particular strength in ice, snow, basketball, and other sports, with 8-13 gold and 15-25 total medals.

The influence of the host country will be crucial to the performance of the Olympic Games, and the United States will have a significant home-field advantage as the host country in 2028. The U.S. will receive enthusiastic support from their home crowds, a morale boost, and possibly even a benefit regarding referee decisions. In addition, new sports such as baseball/softball, lacrosse, cricket, squash, and flag football are among the strengths of the United States, where 15-25 gold medals are expected. On the other hand, India has benefited from the inclusion of cricket, which has a huge fan base in India and is expected to do well in the sport. In contrast, the performance of China and Russia may be affected by project adjustments. China's weightlifting and boxing events are its traditional strengths, but both sports could be eliminated from the 2028 Olympics. The Chinese weightlifting team won five gold medals at the Paris Olympics, and the boxing team won three gold medals. If these two events are no longer held, China's gold medal count is expected to be affected. Russia's Olympic results are likely to be lackluster due to international sports pressures such as bans in recent years, as global sports competition has intensified, and other countries are gradually catching up in their dominant events.

Separately, countries that have never won a medal, such as Bangladesh and Benin, are expected to win their first medal in 2028. Bangladesh has potential in events such as badminton and has government support, with an expected 15%-20% chance of winning a medal. On the other hand, Benin has made significant progress in events such as athletics and boxing, with a 20%-25% chance of winning. Breakthroughs in these countries depend on sustained investment in sports, athlete development, and international cooperation.

Overall, the hosts' home advantage, programme adjustments, and countries' preparedness will directly impact the medal distribution for the 2028 Games.

## 4. Olympic events and medal distribution analysis model

This subsection aims to identify which events are significant to a particular country based on the distribution of medals in different Olympic sports through a clustering approach. Cluster analysis can identify similarities between countries and identify each country's strengths.

### 4.1 K-means++ Clustering Model

K-means++ is an optimization of the traditional K-means algorithm, which accelerates the algorithm's convergence by selecting the initial cluster center more reasonably and reduces the possibility of falling into the local optimal solution. K-means++ improves the initialization process and avoids the problems that may be caused by the traditional K-means random selection of cluster centers, thereby improving the clustering effect and computational efficiency. Here are the steps and how to calculate K-means++:

**Step1**: Select the First Center Point

A point is randomly selected from the data as the first cluster center, denoted as $C_1$.

**Step2**: Calculate Distance to the Nearest Center

For each point $x$ that is not selected as the center, calculate the distance from it to the nearest one in the center of the selected cluster, set to$D(x)$, and the formula is as follows:

$$D(x) = \min_{i=1,\ldots,k'} dist(x, C_i) \tag{16}$$

where $k'$ is the number of currently selected center points, $C_i$ is the $i$-th selected center, and $dist(x, C_i)$ is usually the Euclidean distance.

**Step 3**: Select a New Center Point by Probability

The probability of each data point $x$ being chosen as the new center, proportional to the square of its distance, is defined as：

$$P(x) = \frac{D(x)^2}{\sum_j D(x_j)^2} \quad (17)$$

Based on this probability distribution, a new cluster center is randomly selected, and points farther away from the existing center are more likely to be chosen. The newly selected cluster center is $C_{k'+1}$.

**Step 4**: Repeat *Steps 2* and *3* until K initial cluster centers are selected.

After selecting K initial center points, the algorithm enters the normal K-means clustering step.

**Step 5**: Perform K-means Clustering

The standard clustering process of K-means was carried out using the selected K centers as the starting point.

### 4.2 Result Analysis

In the cluster analysis, three key indicators were used: **the number of medals (the number of gold**, **silver, and bronze medals won by each country in different events), the proportion of medals (the proportion of medals in each event to the country's** total **medals, reflecting its strengths), and the type of events (how many medals each country won in different events, a measure of its overall competitive ability)**. Based on these indicators, the cluster analysis divides countries into various categories, each representing countries with similar Olympic performances. The specific results are as follows:

**Single-advantage project countries**: These countries excel in certain specific projects. For example, the United States excels in swimming, Russia excels in gymnastics, and China has a clear advantage in table tennis, winning a large number of medals, which has become its strong point.

**Diverse countries**: These countries perform well in multiple programs, but do not have a single dominant program. Germany, for example, has contributed to shooting, athletics, swimming, and other events, but none of them stand out in particular.

**Specialized countries**: These countries have significant advantages in some programs, but fewer in others. For example, China has performed exceptionally well in table tennis and diving, but its performance in other sports has been relatively weak.

## 5. Conclusion

Our study shows that past performance is a strong indicator of future success. Countries that have done well in previous Games will likely continue performing well. This means that improving areas where the government has succeeded can help increase medals. We also found that hosting the Games often gives countries an advantage. The host effect helps improve medal counts, which can benefit the country in future Games. When looking at specific sports, swimming, athletics, and gymnastics usually bring in the most medals. Investing in these sports can lead to better results, especially with more resources for training, facilities, and coaches. While no prediction is perfect, and outside factors like politics or economics can affect results, our study clearly shows a way forward. By focusing on successful sports, improving coaching, and using data to make decisions, the National Olympic Committee can enhance performance in future Games.

## References

[1] Allison, L., Monnington, T., 2002. Sport, prestige and international relations. Gov. Oppos. 37, 106–134.

[2] Hermann, A., 2019. The tip of the iceberg: the Russian doping scandal reveals a widespread doping problem. Diagoras: International Academic Journal on Olympic Studies 3, 45–71.

[3] Lee, C., 2021. A review of data analytics in technological forecasting. Technol. Forecast. Soc. Change 166, 120646.

[4] Xun Bian (2005) "Predicting Olympic Medal Counts: the Effects of Economic Development on Olympic Performance" Zhangqiao Scientific Research.

[5] Ball, D.W., 1972. Olympic games competition: structural correlates of national success. Int. J. Comp. Sociol. 13, 186–200.

[6] Armstrong, J.S., Collopy, F., 1993. Causal forces: structuring knowledge for time-series

[7] extrapolation. J. Forecast. 12, 103–115.

[8] Zhao, X., Yan, X., Yu, A., van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models. Travel Behav. Soc. 20, 22–35.