

Medal Prediction Model Based on BP Neural Network

Leqi Xu^{#, 1}, Xinying Zhao^{#, 1, *}, Yifei Gao^{#, 2}

¹College of Sciences, Nanjing Agricultural University, Nanjing 210095, China

²School of New Energy and Materials, Southwest Petroleum University, Chengdu 610500, China

*Corresponding author: Xinying Zhao (Email: zhaoxinying@stu.njau.edu.cn)

[#]These authors contributed equally.

Abstract: This study predicts medal standings for the 2028 Olympics using a two-layer BP neural network model, analyzing historical data and national potential factors. A newly proposed National Potential Index (P) serves as a key parameter to assess countries' likelihood of winning their first medal. Findings suggest Czechoslovakia, Yugoslavia, Australasia, and Uruguay will achieve medal breakthroughs, while China and the United States maintain dominant positions. The research employs a hierarchical analytical approach to enhance prediction reliability, evaluating model performance through precision, recall, F1 score, MSE, and MAE metrics. Results demonstrate the model's effectiveness in forecasting medal distributions, particularly in identifying emerging competitors and confirming traditional powerhouses' advantages. Validation confirms the model's reference value for Olympic medal predictions, successfully integrating classification and regression analyses to balance breakthrough detection with ranking accuracy. The methodology emphasizes stepwise predictions, first determining medal qualification probabilities before estimating specific medal counts, thereby optimizing outcome reliability. This dual-layer structure addresses both categorical (medal attainment) and continuous (medal quantity) prediction challenges, offering comprehensive insights into Olympic performance dynamics.

Keywords: Stepwise prediction, BP neural network, Sliding window method, KS test.

1. Introduction

The Olympic Games, without a doubt, stand as the most preeminent and influential sports extravaganza across the globe. It serves as a magnificent stage where the world's top athletes converge to showcase their extraordinary skills and tenacious spirits. For sports enthusiasts, the allure isn't merely confined to the excitement of watching diverse individual competitions. They are equally captivated by closely tracking the medal standings. Each athlete who manages to bring glory and honor to their respective countries is rightfully showered with respect and admiration. At the same time, the medal-winning feats of athletes from all corners of the world hold great significance, as they reflect the overall strength and development of sports on a global scale.

In recent years, with the remarkable advancements in big data technology and the burgeoning field of artificial intelligence, the application of data-driven models for predicting Olympic medals has emerged as a highly debated and sought-after topic in the research domain. In past research endeavors, numerous scholars have made earnest attempts to analyze the multifaceted factors influencing Olympic medal counts by relying on traditional statistical models. However, these models have inherent limitations when it comes to grappling with the intricate and often non-linear relationships that exist within the relevant data. Although some studies have resorted to linear algorithms, they still fall short in terms of the depth of data mining and the complexity required to accurately capture the real-world scenarios.

Against this backdrop, the present study sets out with a clear and ambitious objective: to construct a far more accurate and interpretable model specifically tailored to predict the number of medals that countries will secure at the 2028 Olympics. The research will embark on an in-depth

exploration of the key factors that play a decisive role in medal predictions. By meticulously sifting through and analyzing a vast amount of data, it aims to unearth the hidden regularities and patterns underlying the numbers. Moreover, through the implementation of this model for predictions, this study endeavors to offer a solid scientific foundation for the sports development planning of various countries. It is hoped that by doing so, the efficiency of sports resource allocation can be significantly enhanced, thereby giving a strong impetus to the further flourishing and development of the global sports industry.

2. Materials and methods

2.1. Data acquisition and preprocessing

The data in this article comes from the official website of the Olympics (add the URL and category) historical medal rankings dataset. Due to the complexity of the data, the age of the records, and factors such as missing and erroneous data.

Missing value processing. Two functions were used to achieve the total number of missing values in the data set of the statistical analysis of the total number of missing values, so as to effectively identify the missing data in the data set. It is recommended to replace them with zero.

Outlier Handling. Outliers are handled by utilizing Emmett interpolation[1]. Emmett interpolation is an interpolation method commonly used to smooth curves and fill in missing data. The method uses data points and their derivatives (or rates of change) to construct interpolating polynomials. This method can deal with outliers that are clearly out of place in the data: the point values around the data are used to infer the location of the outliers, thus restoring the data to its smooth state.

2.2. Introduction to the method and advantages

This study aims to enhance the model's analytical capabilities by dividing the problem into levels and solving them step by step to maximize the reliability of the results. It explores the prediction of the medal situation for different countries in the 2028 Olympics using a two-layer BP neural network model. First, a classification model is constructed to exclude those data that contribute less to the prediction accuracy but are large in quantity, which may negatively affect the model performance. Subsequently, a regression model using BP neural network was used to predict the final results. After obtaining the final results, a model performance study should be carried out. In this study, the accuracy of the former classification BP model is proposed to be verified, and the mean square error (MSE) and the mean absolute error (MAE) are proposed to be used as the performance criteria for the latter regression BP model.

2.3. Model evaluation index

After that, this paper will use MSE and MAE to test the confidence level of the regression model predictions.

Accuracy is the proportion of samples predicted by the classification model to be in the positive category that are actually in the positive category.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Mean Square Error (MSE): Used to represent the mean of the squared error between the predicted and true values of the regression model.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

Mean Absolute Error (MAE): The calculation used is the average of the absolute values of the differences between the predicted and true values of the regression model.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

3. Model building and solving

3.1. Data exploration

This operation is mainly used to mine the hidden mathematical relationships in the original data, in preparation for feature engineering to produce data sets.

3.1.1. The relationship between the number of medals and historical awards

Mining the attached data shows that there is a certain historical dependence on the number of medals won by different countries, i.e., the number of medals won in the next term is correlated with the number of medals won in the past, shown in Figure 1.

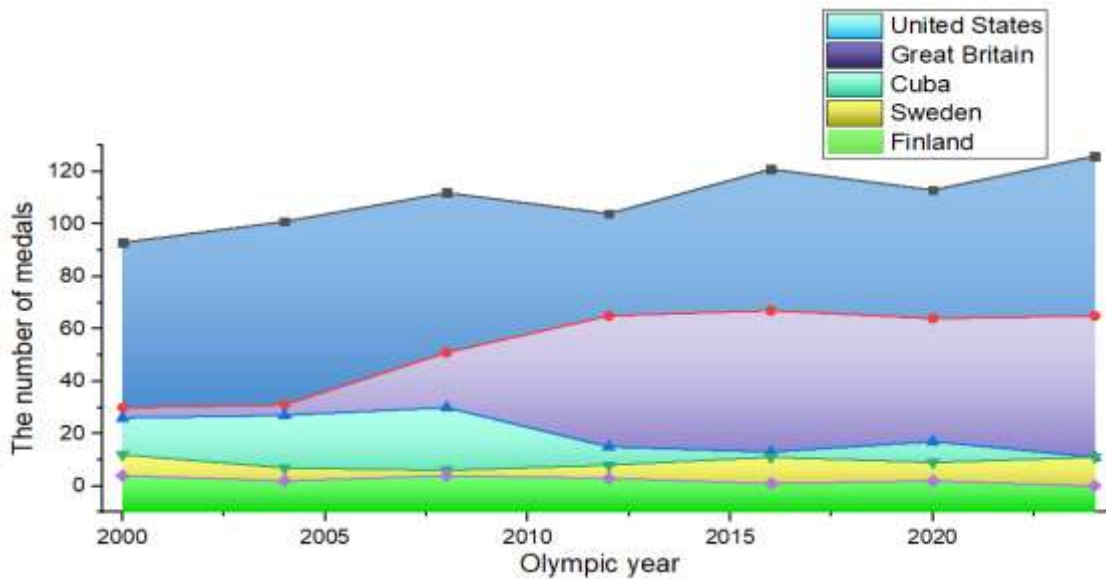


Figure 1. Olympic year and the number of medals

By visualizing and analyzing the award data of the past years, it is revealed that there is a significant positive correlation between the number of medals won by the country in the current year and the number of awards in the history. If the number of historical medals is used as a key indicator to predict the number of awards in the future, it will help to improve the accuracy of the data reliability study.

3.1.2. The relationship between an athlete's years of service and national awards

The average length of service of athletes, as the main body

of the competition, can reflect to a large extent the effectiveness of the country in the cultivation of outstanding athletes and logistical support. The risk of early retirement due to accidental injuries can be judged, which indirectly reflects the investment of the country in athletes and the importance it attaches to the Olympic Games.

By studying the dataset (summerOly_athletes.csv), Figure 2 can be come up with.

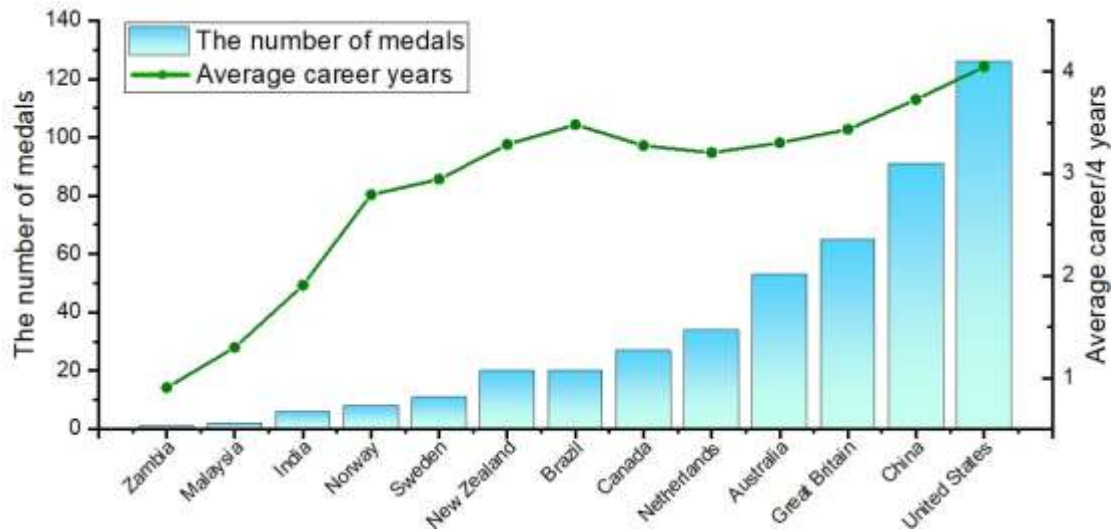


Figure 2. The relationship between average career years and national awards

The bar in the figure shows the service period of national athletes, and the discounted part shows the award rate of the country. It reflects that the number of medals won is positively related to the service period of athletes, so the average service period can be used as one of the bases of the model for the future award rate.

In the data, there are a few events where the number of medals is negatively correlated with the age of the athletes, so additional weighting is needed in the feature engineering.

3.1.3. The number of participating athletes in relation to national awards

Poisson's theorem reveals that increasing the number of trials increases the probability of the event, hypothesizing that the number of athletes from the country is positively correlated with the number of medals won.

From the annex, this paper has counted the number of athletes from some countries and the number of medals won by the countries to get Figure 3.

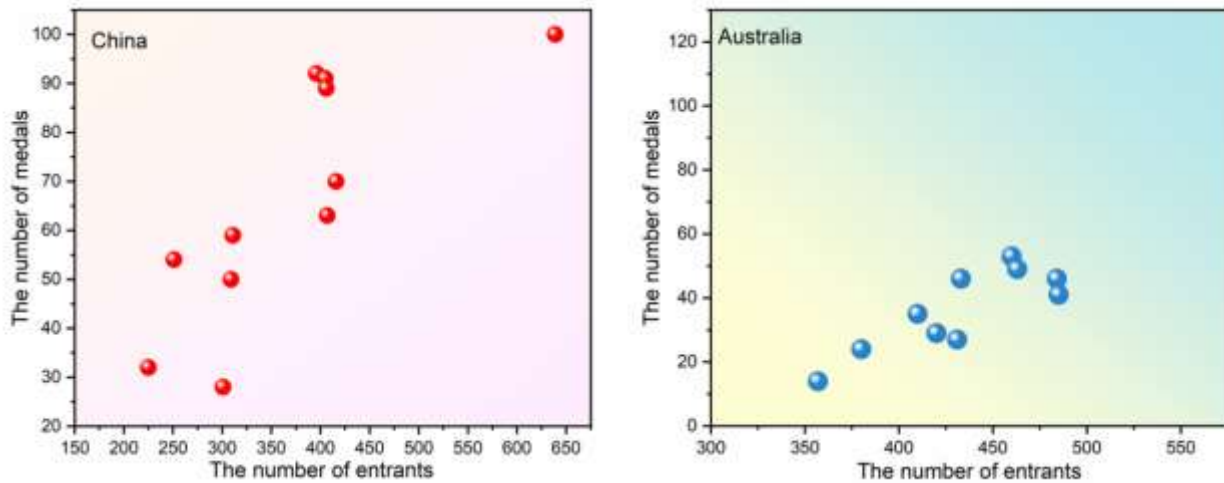


Figure 3. Number of participating athletes in relation to national awards

It can be seen that there is a significant positive correlation between the number of participating athletes and the number of medals won. However, when the number of participating athletes reaches a certain threshold, the number of medals won tends to stabilize, which may be related to the number of excellent athletes produced by the population base or national factors. Nonetheless, the overall positive correlation between the two is still present, and thus can be used as a characteristic variable for model training.

3.1.4. Host effect

Athletes, as the subject of the competition, usually display a higher level of athleticism than usual when playing on their home turf. Host countries have a greater say in the current Olympic cycle, and have added or strengthened specific disciplines to enhance their competitive advantage (e.g., Japan's addition of karate to the 2020 Tokyo Olympics).

The paper intercepted the host country years of the Olympics and plotted them against the previous Olympics as Figure 4.



Figure 4. The host effect on the medals

The graph reflects that the number of medals won by the host country at the Olympic Games shows a significant increase compared to the previous two years, and it can be inferred that the host country effect is an important indicator in future medal projections.

3.2. Characterization works

After clarifying the important influencing features of the medal list, this study can further construct the relationship between the features through feature engineering.

Feature construction generates better quality or more relevant data from the original data through a series of analytical methods. For a large number of unprocessed features in a dataset, new features are constructed and feature structure optimization is performed to achieve the purpose of enhancing the rationality of the dataset.

3.2.1. Feature interactions

Feature interaction plays a key role in revealing deep causal relationships among data. When a single dataset cannot fully show its inherent laws, it is necessary to perform feature interaction operations on the extracted features in order to generate interaction features. The construction of second-order interaction features is an effective means to achieve this goal. This operation can reveal the nonlinear relationship between the data, and effectively reduce the redundancy of the features and alleviate the multicollinearity between the data.

3.2.2. Logarithmic processing

In the face of large data value, logarithmic transformation is used to obtain more advantageous data performance. Compared with the original data, the logarithmically processed data show higher sensitivity to differences, which makes the data more stable and effectively reduces the

covariance and heteroscedasticity of the model.

3.2.3. Pearson correlation coefficient analysis

In order to exclude the interference of data on the negative correlation between the average age of athletes and the number of medals, this study used Pearson's correlation coefficient analysis with the aim of identifying the demand for young athletes in specific sports. Identification was done by exploring the negative correlation between the number of medals and average years of service. Pearson's correlation coefficient[2]:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (4)$$

X_i and Y_i is the historical number of medals in the sport and the average number of years of service in that sport.

$\bar{X}$ and $\bar{Y}$ denotes the mean of the number of historical medals and the average age of athletes in service.

After performing the Pearson correlation coefficient analysis, this work succeeded in revealing a linear correlation between the performance Y taken in the sport and the number of years of service X of the athlete.

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (5)$$

The linear relationship between the two was determined by constructing a linear regression model.

For the above data after the feature selection process, this paper can determine the degree of correlation between the two by Pearson correlation coefficient analysis, which greatly reduces the error.

Table 1. Pearson correlation coefficient analysis

Sports event	Gymnastics	Sprint	Diving	Skiing	Boxing
Avg cycle	1.2	1.6	1.5	1.8	1.7
r	-0.82	-0.81	-0.79	-0.76	-0.76
294	601	583	720	731	118

From the above Table 1, it can be seen that there are indeed some sports that require athletes to be young enough to have fewer years of service compared to other sports, which requires additional treatment when fed into the model for training.

3.2.4. Data standardization

In order to avoid the huge interference of sudden extreme

values for model training and the inconsistency of data scales, normalization was used to improve the accuracy of the model. Standardized specific processes.

Calculate the mean and standard deviation of.
Mean.

$$\mu_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (6)$$

Standard deviation.

$$\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \mu_j)^2} \quad (7)$$

Among them x_{ij} is the j th feature of the i th sample, n is the total number of samples[3].

Standardized features.

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (8)$$

Among them, x'_{ij} is the normalized eigenvalue, the μ_j and σ_j are the mean and standard deviation of the feature.

3.3. Model building

3.3.1. Modeling

After completing the basic data preprocessing and feature engineering parts mentioned above, the model will be built formally. This study will use a layered prediction strategy combined with a BP neural network to predict the number of medals won by each country in the Olympic Games in 2028. Firstly, the first layer of neural network aims to categorize whether countries can win medals in the next Olympic Games based on the characteristics of the dataset, especially the historical data of each country. Subsequently, the second layer of neural network will receive the output of the first layer of

the network as the input of the countries with the potential ability to win medals, and further make predictions in these data, so as to realize the accurate prediction of the number of specific medals of these countries[4].

The advantage of the stepwise prediction method is that it is able to deal effectively with large data sets, especially those containing a large number of countries that have failed to win medals at the Olympic Games for a long period of time. The likelihood of these countries winning a medal at the next Olympic Games remains low. However, incorporating these data into the final prediction model may reduce the prediction efficiency and may adversely affect the model weights due to the large number of these countries, thus affecting the prediction results. By introducing a layer of BP neural network for preliminary classification, these redundant data can be effectively eliminated to improve the accuracy and robustness of the prediction model[5].

3.3.2. BackPropagation Neural Networks

As a classical feed-forward neural network, BP neural network is widely used in pattern recognition, time series prediction and regression analysis, etc. BP neural network is an adaptive nonlinear modeling tool by adjusting the network weights and biases through the back propagation algorithm. Its powerful learning ability makes it especially outstanding in modeling complex nonlinear relationships. As a kind of multilayer feedforward neural network, it usually contains three main parts: input layer, hidden layer and output layer. Each layer transfers information to each other through weighted connections, and the goal of the network is to fit the nonlinear mapping relationship between input and output by learning the training data.

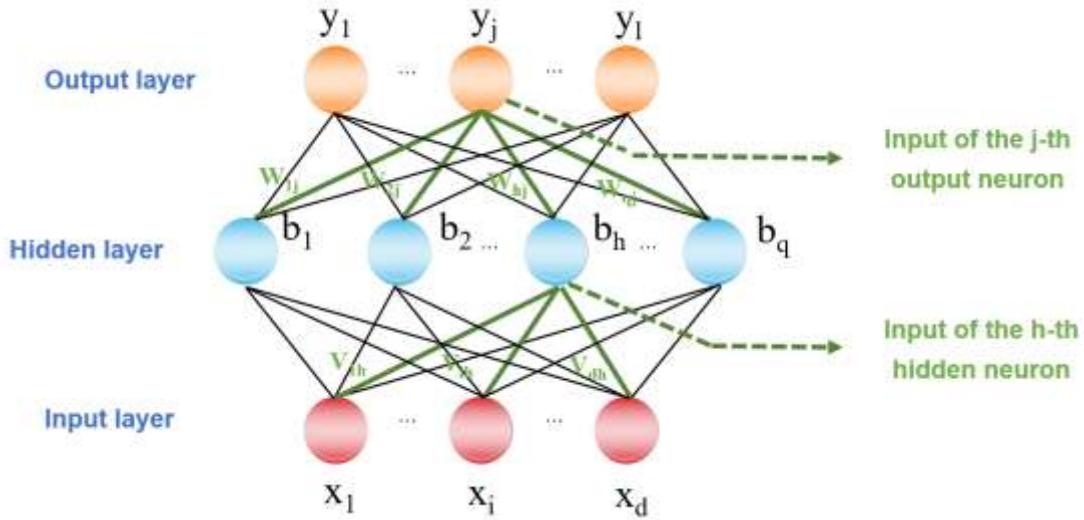


Figure 5. BackPropagation Neural Networks

As for the ontology model shown in Figure 5, the input values corresponding to each neuron in the input layer are x_1, x_2, x_3, x_4, x_5 (representing the feature dataset country, year, number of awards, number of years of service, and number of athletes, respectively), and the weight matrix between the input layer and the hidden layer is W_1 , in which the dimensionality of W_1 is $h \times 5$, i.e., there are h neurons, and each neuron is connected to 5 input features.

To train the BP neural network, this paper used the mean square error (MSE) as the loss function. Setting y^{true} is the

actual output value (0 or 1), then the mean square error loss function is.

$$E = \frac{1}{n} \sum_{k=1}^m (y_k^{true} - y_k^{pred})^2 \quad (9)$$

Among them, y_k^{pred} is the predicted value of the k th sample, the y_k^{true} is its actual label (0 or 1).

For the use of BP neural network the flowchart is shown in Figure 6.

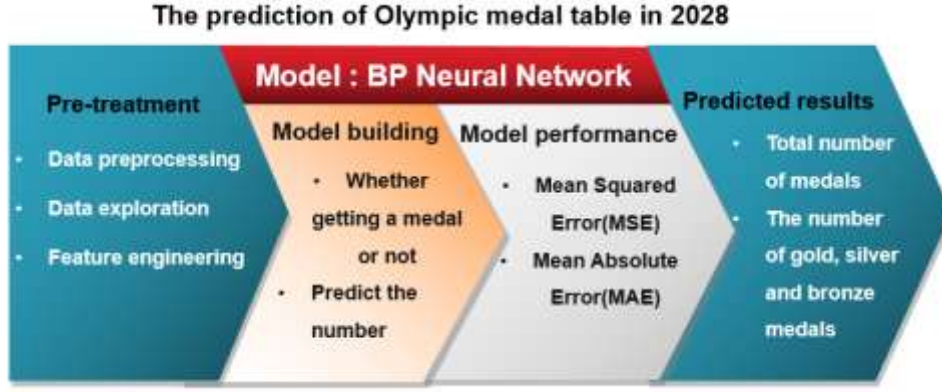


Figure 6. The prediction of Olympic medal table in 2028

3.4. Model performance studies

With the above model based on BackPropagation, it is easy to derive the data required for the problem. However, whether the data trained by this model is reliable and can be used as a prediction for future competitions, this method will need to be evaluated using a test set to assess the performance of the model. In this study, the first classification model will be evaluated using Precision to ensure that the separation of probable winners and non-winning countries is reliable. After that, MSE and MAE were used to test the confidence level of the regression model predictions[6].

Accuracy is the proportion of samples predicted by the classification model to be in the positive category that are actually in the positive category.

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

Mean Square Error (MSE): Used to represent the mean of the squared error between the predicted and true values of the regression model.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11)$$

Mean Absolute Error (MAE): The calculation used is the average of the absolute values of the differences between the predicted and true values of the regression model.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (12)$$

In this method, the model obtained from the training set is tested by the test set, and the test results are combined to obtain the following Table 2.

Table 2. The index and value

Index	Value
MSE	0.125292
MAE	0.083827
Precision	0.922857

According to the analysis of the above indicators, the model shows good prediction performance. Specifically, the accuracy of the previous classification model is as high as 0.92, which basically does not misclassify countries with a high probability of winning and countries with a high probability of not being able to win the award, and provides a reliable data source for the subsequent regression model.

The mean square error (MSE) and mean absolute error (MAE) values of the regression model are low, which indicates that the prediction results of the model have a high degree of proximity to the real values, i.e., its prediction results are very reliable.

3.5. Display of forecast results

The prediction results are plotted in Figure 7.

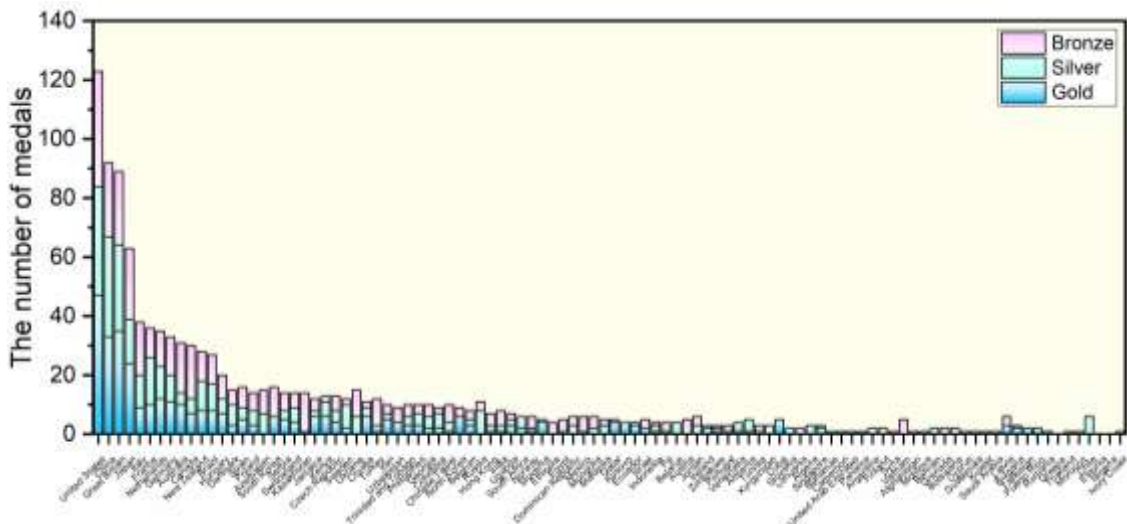


Figure 7. The prediction of Olympic table in 2028

4. Result and Discussion

This is the result obtained through the training of the above model, it can be found that in the Olympic medal table, the top ten countries are the United States, China, Britain, Japan, Italy, France, the Netherlands, Germany, Australia and Canada. The prediction interval is the current value ± 1 block, because when the model is established, the prediction is correct within plus or minus one, and the subsequent prediction interval is the same as above.

In the gold medal list, the top ten countries are the United States, the United Kingdom, China, Japan, the Netherlands, Germany, France, Australia and Italy.

The analysis shows that there is a high degree of overlap between the top 10 countries in the gold medal table and the medal table, although there are slight differences in the order of ranking. This suggests that the gold medal standings largely reflect the medal standings and vice versa, and together they reveal the athletic performance of countries in the Olympic Games. The United States, in particular, tops both the gold and medal lists, which, combined with historical data, suggests that the United States is a leader in the international sports arena.

It can be observed from the medal distribution chart that developed countries generally have a higher number of medals, while developing countries have relatively fewer medals, a trend that reflects a certain positive correlation between gross domestic product (GDP) and a country's comprehensive national power and level of sports competitiveness.

Analyzing the data, it can be noted that certain countries, such as Cuba and Kosovo, have a significantly higher proportion of gold medals in the total number of medals than other countries, despite the fact that their total number of medals is small. This indicates that these countries have significant international competitiveness and great potential in specific sports. According to the comparative analysis with the previous medal table, Syria, South Korea, Luxembourg and other countries will perform worse because of the decline in the award rate, and Dominica and Hungary will perform better.

5. Conclusion

This study centered on the 2028 Los Angeles Olympics and

successfully built a medal prediction model. It entailed extensive work, like gathering historical Olympic data, athlete stats, and sports - related data from different countries. Advanced data processing and algorithms were used to analyze this data, enabling the model to effectively explain medal - influencing factors and predict medal distribution. This offers a solid theoretical base for countries to devise sports development strategies.

Nevertheless, the study has limitations. The model doesn't cover all Olympic events for prediction, and issues like underfitting and local optimality exist, affecting prediction accuracy.

For future research, broadening the dataset, including data on emerging sports, young athletes, and new training methods, is crucial. Employing more advanced machine - learning algorithms can address the current problems and boost prediction accuracy. The improved model has high application potential, applicable in sports management, federations, and teams for better planning. Overall, while the study has achieved results, continuous efforts are needed to refine the model and widen its application in sports.

References

- [1] H. Li, Z. Shao. A review of spatial interpolation analysis algorithms[J]. Computer System Applications, 2019, 28(07): 1-8.
- [2] Liu P, Wang S, Zhao P. Robust estimation and test for Pearson's correlation coefficient[J]. Random Matrices: Theory and Applications, 2024.
- [3] Jia Xiaoyong, XU Chuansheng, BAI Xin. The creation of least squares method and its ideological approach[J]. Journal of Northwestern University (Natural Science Edition), 2006, (03): 507-511.
- [4] Li Xiaofeng, Liu Guangzhong. Improvement of BP algorithm for artificial neural networks and its application[J]. Journal of Sichuan University (Engineering Science Edition), 2000, (02): 105-109.
- [5] Ines F O, Marta S, M. N V. Explainable generalized additive neural networks with independent neural network training[J]. Statistics and Computing, 2023, 34(1).
- [6] Liu Jingjuan. Multiple linear regression model test method[J]. Journal of Hunan Taxation College, 2005, (05): 49-50+60.