Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 5, No. 2, 2025

Research on Prediction Methods Based on Multiple Linear Regression - Artificial Neural Network Model

Jiayan Zhang[1, †, *], Sihan Yang[2, †], Zitong Wang[3, †]

[1]School of Economics, Shanghai University, Shanghai, China

[2]School of Computer Engineering and Science, Shanghai University, Shanghai, China

[3]SILC Business School, Shanghai University, Shanghai, China

[*] **Corresponding author**: Jiayan Zhang (Email: JennieYan@163.com)

[†]**These authors also contributed equally to this work.**

Abstract: The study employs a predictive framework based on a multiple linear regression-artificial neural network (MLR-ANN) model to achieve high-precision predictions for specific targets. First, features are extracted from both overall and individual dimensions, followed by data processing. An artificial neural network model is constructed to uncover non-linear features within the data. Second, a multiple linear regression model is used to fit linear relationships, while incorporating probability distribution characteristics to quantify uncertainty in the prediction process. A result aggregation strategy is further employed to integrate the two models, forming a comprehensive prediction system. Additionally, the study uses a logistic regression model to analyse the inter-variable association mechanisms and evaluates the importance of specific factors through the EWM-TOPSIS method. This framework enhances prediction accuracy and reliability through multi-dimensional feature analysis and model optimisation, providing an analytical framework and methodological reference for similar problems.

Keywords: MLR-ANN Model, Logistic Regression Model, EWM-TOPSIS Method, Time-Series Exclusive Cumulative Algorithm.

1. Introduction

As data structures become increasingly diverse and the mechanisms governing inter-variable relationships grow more complex, the challenge of constructing a predictive framework that balances both feature analysis depth and model generalisation capability has emerged as a key research focus [1]. This study proposes a multiple linear regression-artificial neural network (MLR-ANN) ensemble model [2-3]. By hierarchically extracting data features, the model leverages the non-linear mapping capabilities of artificial neural networks to uncover deep-level patterns in the data, employs multiple linear regression to fit linear relationships, and incorporates probability distribution theory to quantify the uncertainty in the prediction process [4]. Through a result integration strategy, the outputs of the two models are fused to form a comprehensive prediction system that encompasses feature engineering, multi-model collaboration, and uncertainty analysis. Additionally, the study introduces a logistic regression model to analyse the association mechanisms between variables [5], employs the EWM-TOPSIS method to assess feature importance [6], and constructs a multi-dimensional model optimisation framework. This framework systematically integrates the advantages of different modelling methods, providing a new approach for high-precision prediction in complex data environments, with its methodology scalable to the analysis and resolution of similar prediction problems.

2. Medal prediction model construction

2.1. Model choosing

Considering features selected and the characters of data, this paper first adopts Artificial Neural Network, which has excellent fault tolerance and good generalization ability. However, after testing samples, this paper find that it takes too long time for ANN algorithm to process data. Meanwhile, for small-scale data, the results of multiple linear regression (MLR) show higher prediction accuracy than those of ANN. Therefore, this paper finally adopts the MLR-ANN (Multiple Linear Regression-Artificial Neural Network) model.

2.2. Model design

(1) Predicting model counts for each country

This model is a deep neural network implemented using the Sequential model from the Keras library. The structure of the model includes an input layer, two hidden layers, and an output layer.

Step 1: Medal prediction for each athlete

Once the model predicts the expected type of medal that each athlete will win, denoted as $\hat{y}_i$, this paper can calculate the 95% confidence interval for each athlete's predicted medal. The confidence interval provides the range within which the true type of the predicted medal is likely to fall.

For each athlete i, the 95% confidence interval is given by:

$$CI_i = \left[\hat{y}_i - 1.96 \cdot \sigma_{\hat{y}_i}, \hat{y}_i + 1.96 \cdot \sigma_{\hat{y}_i} \right] \tag{1}$$

Where $\hat{y}_i$ is the predicted type of medal for athlete i, $\sigma_{\hat{y}_i}$ is the standard deviation of the prediction for athlete i, 1.96 is the critical value for a 95% confidence interval from the standard normal distribution.

Step 2: Total medal count calculation for the country

Once this paper obtained the predicted number of medals for each athlete, it summarized these results to determine the total number of medals expected for that country.

(2) Features selecting

To enhance the model's predictive accuracy, this paper categorized the factors influencing a country's final medal

count at the Olympic Games into two layers: the evaluation of the country's overall ability in a particular sport and the evaluation of the overall ability of individual athletes.

a) Features of individual athlete

Host country effect: This paper incorporated the participation or non-participation of athletes from the host countries into the evaluation model's consideration factors, and use formula 1 to quantify this kind of effect.

$$Is_H = \sum_{t=1}^{n} \frac{\left(X_0 - \bar{X}\right)}{n} \times 100\% \tag{2}$$

Where X_0 is the percentage of gold medals won by the host in the current session to the total number of gold medals won in that session; $\bar{X}$ is the average of the percentage of gold medals won by the host in other sessions to the total number of gold medals won in that session; n is the total number of sessions in which the country has won gold medals at the Olympics.

Athlete's performance: This paper constructed two indicators: one representing the probability that a specific athlete will win a medal in event t, and the other reflecting the athlete's performance in event. These indicators are used to quantify the likelihood of winning a medal for a particular athlete, as described by the following formulas:

$$P_{i,t} = \frac{n_1}{n_2} \tag{3}$$

$$S_{i,t} = \frac{\sum p_i}{n_2} \tag{4}$$

Where n_1 represents the total number of medals an athlete has won (regardless of the type), n_2 is the number of Olympic Games in which the athlete has participated, and p_i denotes the total points earned from the medals won by the athlete. This paper assigned six points for a gold medal, three for a silver, two for a bronze, and zero for no medal.

In addition to the above features, this paper also considers the athlete's gender, participation events and interval time.

By integrating the above indicators, this paper established the evaluation equation of athlete to evaluate an athlete's performance in a particular sport, quantifying their achievements through the number of medals they have earned:

$$A_s = a_1 + b_1 Is_H + b_2 P_{i,t} + b_3 S_{i,t} + b_4 Period \tag{5}$$

b) Features of Country

The features of country include athletes participation counts, previous national winning probability

current national winning probability and sport. This paper defined S in the model as: Which sport has the highest number of medals won by the country.

Before modeling the state's assessment on a particular program, this paper first made definitions of the following indicators:

$$P_{\text{win}} = \frac{\text{count}\left(A_s \geq 2\right)}{\text{count}\left(A_s\right)} \tag{6}$$

$$P_{\text{past}} = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\text{Medal}_i}{N_i} \tag{7}$$

$$\text{Gold} = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\text{GoldCount}_i}{N_i} \tag{8}$$

$$\text{Silver} = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\text{SilverCount}_i}{N_i} \tag{9}$$

Where n is the predicted number of editions of the Olympic Games, $Medal_i$ refers to the number of medals won by the country in year i. same to $GoldCount_i$,$SilverCount_i$, $BronzeCount_i$.

Combining all the factors above:

$$MC_y = \alpha_1 + \beta_1 N_i + \beta_2 \text{ Gold} + \beta_3 \text{ Silver} + \beta_4 \text{ Bronze} + \beta_5 P_{\text{past}} + \beta_6 P_{\text{win}} + \beta_7 I_{\text{bestsport}} \tag{10}$$

Where β denote the weights for each factor.

This model incorporates both the national level and individual athlete factors to predict the number of medals a country will win at the Olympics. It provides an objective and clear estimate by considering the performance of athletes and the country's selection of sports, while excluding the influence of excessive historical medal counts on future outcomes.

The output of the model is a 1*4 matrix, which represents the number of people who did not win a medal for a particular sport, and the number of people who will win gold, silver, and bronze medals.

In order to further confirm the accuracy of the model, this paper use three indicators, Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-Squared (R^2) to measure.

People all have common sense that it is difficult to achieve sustained success in competitive sports. To account for this, this paper used the promotion rate formula to analyze the relationship between the number of medals of a country wins in 2028 and 2024. This allows us to observe changes in medal counts over time. A promotion greater than 0 means that the country's performance at the2028 Olympics improve compared to previous years, signaling a positive trend, whereas a negative rate suggests that the country's athletes will experience a decline in performance. The Promotion Rate is defined as:

$$P_{ri} = \frac{MC_{2028} - MC_{2024}}{MC_{2024}} \tag{11}$$

Where MC_{2024}, MC_{2028} are country i's medal counts in 2024 and 2028 respectively.

3. Model for predicting the probability of breakthroughs by countries without medals

It is well-established that certain countries have yet to secure an Olympic medal. There is a persistent curiosity regarding the performance of these lagging nations. To explore this phenomenon, it is essential to first identify those countries that have not won any medals to date. In this context, this paper defines two sets:

$$A = \left\{ \text{country}_{p1}, \text{country}_{p2}, \ldots, \text{country}_{pk} \right\} \tag{12}$$

$$B = \left\{ \text{country}_{\text{win}1}, \text{country}_{\text{win}2}, \ldots, \text{country}_{\text{win}j} \right\} \tag{13}$$

$$A-B=C=\{\ \text{country}_{\text{no1}},\ \text{country}_{\text{no2}},\ldots,\ \text{countr}_{\text{nos}}\ \} \tag{14}$$

Where $country_p$, is the country participated in Olympic Games, $country_{win}$ is the country that wins at least one medal in the past games and $country_{no}$ is the country which has never won medals. There must be:

$$s \le j \le k \tag{15}$$

Combined with the promotion rate defined before, this paper predicted in our model whether they would win a medal at the next Olympics, and provided confidence intervals to measure the probability.

In light of the fact that the number of medals won by these countries that have never secured a medal, calculated based on their winning probabilities, is uniformly zero, this study resorts to estimating solely through the number of participants in various sports. This paper counted the number of participators in each sport by these non-medal winning countries in the last two Olympic Games (2020 and 2024). Since most Olympic athletes retire after more than two Games and have no impact on the 2028 Olympics, this number is denoted as T_1. At the same time, based on the given data, this paper calculated the total number of participations in these sports by all countries in these two Olympic Games, denoted as T_2. Since most participating countries did not win medals, this paper roughly calculated:

$$P=\frac{T_1}{T_2} \tag{16}$$

Meanwhile, this paper defined SS (the Sum Score) for each country as:

$$SS=\sum_{k=1}^{n} T_1 P \tag{17}$$

It is postulated that the larger P, the higher the probability of a country winning a medal in a specific sport. The larger the SS, the greater the probability that the country will win a medal in the next Olympic Games.

To predict the number of countries that will win medals in the next Olympics, this paper used the Time-Series Exclusive Cumulative Algorithm. First, this paper recorded the NOC data in 1896 as the initial dataset. Subsequently, for the NOC records in 1900, this paper systematically compared them with the 1896 data, and only retain the NOCs in the 1900 records that did not appear in the 1896 data, and update the dataset. Similarly, this paper performed the same operation on the NOC records of subsequent years (up to 2024) year by year. In this way, this paper obtained the number of countries that have never won a medal in each Olympic Games since the late 19th century. This paper used linear regression fitting based on the least-squares method:

$$y=mx+b \tag{18}$$

$$m=\frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right)-\left(\sum x\right)^2} \tag{19}$$

$$b=\frac{\sum y-m\left(\sum x\right)}{n} \tag{20}$$

To calculate the sort of odds to this estimate, this paper used a logistic regression model:

This paper performed a linear combination of the input features:

$$z=\beta_0+\beta_1 SS \tag{21}$$

Where, z is the result of the linear combination, β_0 is the intercept (bias term), and β_1 is the coefficient corresponding to the SS feature. These parameters need to be learned during the model training process.

This paper used the Sigmoid function to convert the result z of the linear combination into a probability value p:

$$p=\frac{1}{1+e^{-z}} \tag{22}$$

The log-loss function is usually used to measure the difference between the predicted probability and the true label:

$$L(\beta)=-\frac{1}{m}\sum_{i=1}^{m}\left[y_i\log\left(p_i\right)+\left(1-y_i\right)\log\left(1-p_i\right)\right] \tag{23}$$

Accuracy (AC) is one of the most prevalently used evaluation metrics in classification models. It is defined as the proportion of correctly predicted samples within the total number of samples and its formula is presented as:

$$AC=\frac{TP+TN}{TP+TN+FP+FN} \tag{24}$$

Where, TP (True Positives) represents the number of true positives, that is, the number of samples predicted as positive and actually positive; TN (True Negatives) represents the number of true negatives, that is, the number of samples predicted as negative and actually negative; FP (FalsePositives) represents the number of false positives, that is, the number of samples predicted as positive but actually negative; FN (False Negatives) represents the number of false negatives, that is, the number of samples predicted as negative but actually positive.

4. Correlation analysis between sports events and medals

One thing is certain that events and sport chosen by a country will definitely affect their final medal achievement, but usually it is hard to say what sport to choose, because each country's sports team has a different way of training, the emphasis on the sport may be skewed. To evaluate the importance of each sport, this paper developed the Entropy Weight Method-TOPSIS Method. This paper choosed three indicators

$$I_1=c_1, I_2=\frac{c_2}{c_1}, I_3=6n_{\text{Gold}}+3n_{\text{Silver}}+2n_{\text{Bronze}} \tag{25}$$

In these formulas, c_1 is the number of times the country participates in the sport, c_2 is the number of times the country wins a medal, n_{Gold} is the number of gold medal the country wins, n_{Silver} is the number of silver medal the country wins, and n_{Bronze} is the number of bronze medal the country wins.

After acquiring the requisite data from the dataset, it is necessary to standardize these three datasets in order to eliminate the impact of varying indicator dimensions.

Suppose there are n samples and m indicators, and the original data matrix is $X = (x_{ij})_{25083}$.

For each of these indices x_i, y_i is obtained after normalization.

$$y_{ij} = \frac{x_{ij} - x_{ij\min}}{x_{ij\max} - x_{ij\min}} \tag{26}$$

This paper defines the proportion of the sample i in the j indicator p_{ij} as:

$$p_{ij} = \frac{y_{ij}}{\sum_{i=1}^{n} y_{ij}} \tag{27}$$

The entropy value e_j of the j indicator:

$$e_j = -k\sum_{i=1}^{n} p_{ij} \ln\left(p_{ij}\right), k = \frac{1}{\ln n} \tag{28}$$

When $p_{ij} = 0$, let $p_{ij}\ln(p_{ij}) = 0$.

Calculate the difference coefficient g_j, of the j indicator:

$$g_j = 1 - e_j \tag{29}$$

the weight w_j of the j indicator:

$$w_j = \frac{g_j}{\sum_{i=1}^{3} g_j} \tag{30}$$

After determining the weights of each indicator using the entropy-weight method, this paper calculated a score for each sport in a specific country:

$$Score = \sum_{j=1}^{3} w_j I_j \tag{31}$$

Where $w_1 = 0.21238$, $w_2 = 0.32978$, and $w_3 = 0.45785$ are the weights of the three indicators respectively.

After obtaining the final score for each sport in different countries, for each country, this paper only retained one score. The sport corresponding to this retained score is regarded as the most important sport for that country.

By the way, for countries that have never won a medal, the larger the SS, the greater the probability of winning a medal. The variables in SS are the number of athletes participating in a sport (T_1) and the total number of people participating in the sport in each country (T_2). When T_1 is constant, the smaller T_2 is, the larger SS is, which means that when the distribution of athletes is stable, the fewer people participate in the event, the greater the probability of winning a medal; when T_2 is constant, the larger T_1 (the number of athletes involved) is, the greater the probability of winning a medal.

5. Model results and validation

5.1. Medal predictions results

After data compilation, analysis and measurement, the athlete evaluation equation turns out to be:

$$A_s = 1.2589 + 0.057Is_H + 0.9834P_{i,t} - 0.0107S_{i,t} - 0.0078 \text{ Period} \tag{32}$$

The United States, China and France ranked the first three places, with 148, 117, 98 medals won respectively. The first place of gold medal comes to the United States.

Table 1. Confidence intervals for results

Nation	Gold Medal	Silver Medal	Bronze Medal	Total Medal
United States	[35.16,50.84]	[24.16,39.84]	[53.16,68.84]	[140.16,155.84]
China	[25.16,40.84]	[45.16,60.84]	[23.16,38.84]	[109.16,124.84]
France	[21.16,36.84]	[30.16,45.84]	[23.16,38.84]	[90.16,105.84]
Great Britain	[15.16,30.84]	[24.16,39.84]	[15.16,30.84]	[70.16,85.84]
Australia	[11.16,26.84]	[19.16,34.84]	[18.16,33.84]	[64.16,79.84]
Italy	[7.16,22.84]	[15.16,30.84]	[6.16,21.84]	[42.16,57.84]
Japan	[8.16,23.84]	[8.16,23.84]	[6.16,21.84]	[38.16,53.84]
Germany	[3.16,8.84]	[9.16,24.84]	[8.16,23.84]	[36.16,51.84]

The Table 1 presents the 95% confidence intervals for the top eight countries in the medal table, based on the number of gold, silver, and bronze medals, as well as the total medal count. It is evident that the predicted medal counts fall within the corresponding confidence intervals. Additionally, the model demonstrates high accuracy and reliability, with the following performance metrics: $MSE = 1.06, MAE = 1.06, R^2 = 0.69$.

5.2. Predicted results for countries without medals

In addition, this paper also calculates the promotion rate of countries participating in the Olympics Games. Analyzing the positivity and negativity of the improvement rate, it could be clearly learned that 34 countries including Thailand will underperform in the next Olympics, while 47 countries including the United States, Japan, and Canada will improve and win more medals in the next Olympics. There are 88 countries, including Vanuatu and Malawi, that have never won a medal in previous Olympic Games, excluding those that did not participate or win medals due to name changes or historical dissolution.

After calculation, $\hat{y}_{2028}$ is 3. This paper selected the three countries with the top 3 SS, namely ElSalvador, Lebanon, and Samoa, and believe that these three countries would win medals in the next Olympic Games.

Based on the execution results, this paper evaluated the certainty of the prediction through the logistic regression model. The sort of odds this paper give to this estimate is approximately 0.92. This means that approximately 92% of the prediction results of the model on the test set are correct. Based on the current calculation method and model settings,

this paper has a relatively high degree of certainty in predicting that the top 3 countries in terms of SS can win medals.

5.3. Priority of sports events

By selecting the highest-scoring sport for a country, this paper can identify which sports the country's Olympic Committee should prioritize. For example, swimming should be emphasized for Australia, diving for China, and handball for France.

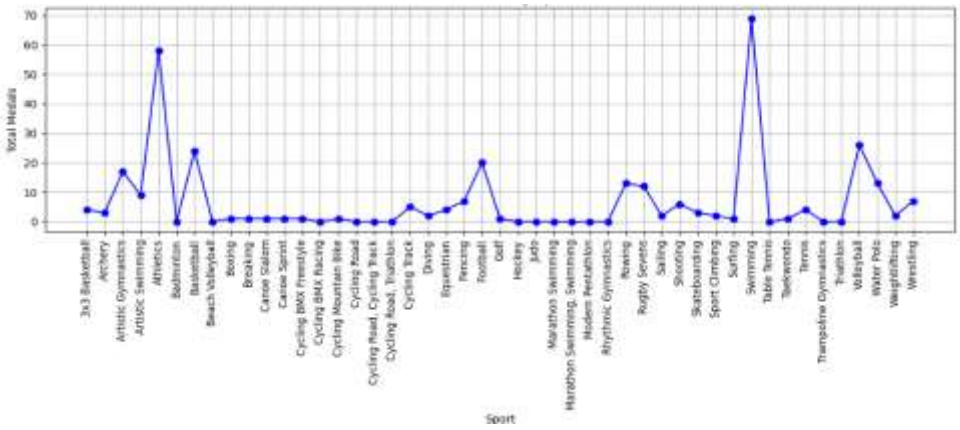


Figure 1. Important sport of countries selected

Figure 1 is a line graph illustrating the relationship between the total number of gold medals won by the United States and the various types of sports in which these medals were achieved. It is evident that the U. S. excels in multiple sports. For example, the U. S. has a clear advantage in swimming, volleyball, and soccer. In this study, this paper also calculated the overall score for each country's sports through the EWM method. Considering factors such as athletes' individual capabilities and other relevant variables, it is essential for national sports teams to base their sport selection on an analysis of historical data. Before evaluating the value of a sport, athletes' individual physical fitness and past winning percentage should be assessed and combined with the team's overall historical winning percentage to determine the most advantageous sport for the home country.

6. Conclusions

The study employs a predictive framework based on MLR-ANN model, which provides an innovative solution for high-precision prediction of complex systems through hierarchical processing of data features and multi-model collaboration mechanisms. The study first uses artificial neural networks to capture non-linear features of the data, employs multiple linear regression to fit linear relationships, and combines probability distributions to quantify prediction uncertainty. The model integration strategy effectively combines the outputs of the two models, forming a comprehensive prediction system covering feature analysis, relationship modelling, and result verification. Additionally, the study employs logistic regression and the EWM-TOPSIS method to further reveal the inter-variable association mechanisms and quantify feature importance, thereby enhancing the model's interpretability and optimisation dimensions.

Experimental validation demonstrates the framework's effectiveness in predictive performance metrics and classification tasks. Its multi-model collaborative mechanism and systematic analysis process provide a reusable methodological reference for similar predictive problems. Future research may further explore dynamic feature update mechanisms and model lightweight optimisation to expand the framework's applicability in real-time predictive scenarios.

References

[1] Zhang Chi, Guo Yuan, Li Ming. Review of the Development and Application of Artificial Neural Network Models [J]. Computer Engineering and Applications, 2021, 57 (11): 57-69.

[2] Torkashvand M A, Ahmadi A, Nikravesh L N . Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network(ANN) and multiple linear regressions(MLR) [J]. Journal of Integrative Agriculture, 2017, 16 (07): 1634-1644.

[3] Zhang Bohu, Hu Yao, Wang Yan, et al. Inversion of in-situ stress field and fracture prediction based on the MLR–ANN algorithm [J]. Journal of Southwest Petroleum University (Natural Science Edition), 2024, 46 (03): 1-12.

[4] Zhang Lina, Jiang Zhicheng, Liu Dayong, et al. A Method for Processing Correlated Flow Data Based on K-means Cluster Analysis and Multiple Linear Regression [J]. Petroleum Tubulars and Instruments, 2024, 10 (01): 52-56+62.

[5] Zhang, Hanxia. Analysis of scenarios suitable for linear regression and logistic regression [J]. Automation and Instrumentation, 2022, (10): 1-4+8.

[6] Luo Ziyuan, Tian Jian, Ding Shiyuan, et al. Rainfall-induced flood risk assessment in Xiamen City based on an entropy-weighted TOPSIS and neural network composite method [J]. Journal of Disaster Science, 2022, 37 (04): 184-192.