Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 10
ISET 2025

Multi-objective Prediction Model based on GA-BP Neural Network and Logistic Regression

Jingzhi Zhou [a], Yang Gao, Junchao Su

Nanjing University of Information Science and Technology Nanjing, China

[a]202383290362@nuist.edu.cn

Abstract. This study develops a multi-objective prediction model integrating two complementary approaches to address complex prediction tasks in hierarchical data structures. The first is a hybrid Genetic Algorithm-Back Propagation Neural Network (GA-BP), which utilizes advanced feature selection techniques, including Lasso regression and XGBoost, to identify key predictors while addressing nonlinear dependencies and convergence issues. The GA-BP model achieves enhanced robustness, effectively modeling complex relationships through global optimization of initial weights and biases. The second approach is a logistic regression submodel, designed to estimate probabilities of binary classification events with a focus on refined feature adjustments and regularization techniques. This framework highlights the significance of feature engineering, integrating both objective and subjective feature weightings to improve accuracy across multivariable datasets. By combining machine learning methodologies with statistical rigor, this integrated model enhances prediction performance and provides actionable insights for diverse use cases in complex systems.

Keywords: Genetic Algorithm; Back Propagation Neural Network; Logistic Regression; Multi-objective Prediction; Feature Engineering.

1. Introduction

Predicting multivariate outcomes in complex systems is a challenging task due to the inherent complexity of nonlinear relationships, high multicollinearity, and diverse data structures. The ability to create accurate and robust models requires addressing these challenges systematically.Traditional statistical modeling approaches, such as grey prediction [1] and linear regression [2], have been widely used in areas like economic performance forecasting and time series analysis.Similarly, machine learning methods, like feedforward neural networks [3] , have provided alternative solutions for handling nonlinear relationships and time-dependent data. However, these techniques often exhibit significant limitations, including insufficient feature selection capabilities, susceptibility to local optima, and poor generalizability when applied across datasets with varying distributions and contextual heterogeneity.

To address these shortcomings, this study proposes a multi-objective prediction model that integrates the strengths of Genetic Algorithm-Back Propagation (GA-BP) neural networks and logistic regression methods. The framework is specifically designed to leverage advanced machine learning and statistical optimization techniques, thereby enhancing the predictive accuracy and robustness needed for multi-variable systems. This approach adopts a two-stage architecture that targets distinct prediction objectives, providing a comprehensive solution to multivariate forecasting challenges.

The first component of the framework, the GA-BP neural network, is devised for predicting continuous target variables [4]. This submodel employs advanced feature selection algorithms, such as Lasso regression and XGBoost, to eliminate redundant or irrelevant variables, ultimately optimizing feature interactions and improving the overall interpretability of the model. A common issue in traditional neural networks—slow convergence rates and sensitivity to local optima—is effectively addressed by employing genetic algorithms [5] to globally optimize the network's initial weights and biases. This optimization process improves the efficiency and reliability of the model in capturing complex, nonlinear patterns in the data .

In addition, the second component, the logistic regression [6] submodel focuses on binary classification tasks by predicting categorical outcomes with associated probabilities. [7] It utilizes refined feature selection and regularization techniques, including L1 and L2 penalties, to ensure robust generalization while minimizing overfitting. These methods enable the logistic regression submodel to function effectively under conditions of spatial and temporal heterogeneity, offering high interpretability and consistent performance across diverse datasets. The work can be seen as Fig.1. By combining these two complementary approaches, the proposed integrated framework offers several key advantages:

(1) Robust Feature Engineering: The model incorporates advanced feature selection methods to efficiently process high-dimensional, multicollinear data.

(2) Dual-Task Optimization: Separate submodels address continuous and binary outcomes, enhancing the adaptability of the framework to various prediction tasks.

(3) Global Parameter Optimization: By employing genetic algorithms, the model ensures faster convergence and greater accuracy, effectively overcoming local minima in neural network training.

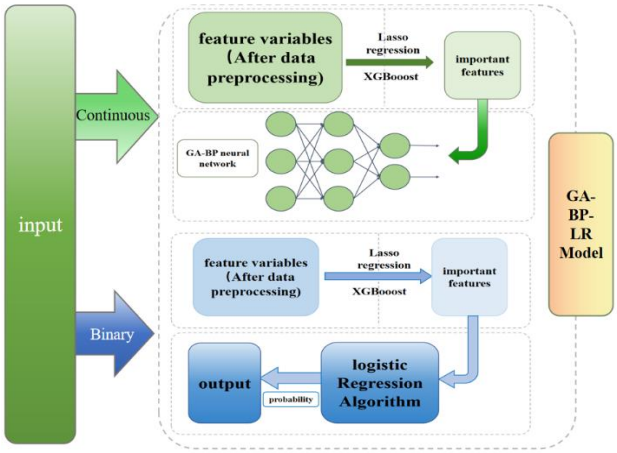


Figure 1. Flow chart of the work

2. Related Work

2.1. GA-BP Neural Network

The GA-BP neural network integrates Genetic Algorithms (GA) with Back Propagation (BP) networks to optimize performance. BP networks, comprising input, hidden, and output layers, use gradient descent to minimize prediction errors by iteratively updating weights and biases. However, they struggle with slow convergence and local minima due to initial parameter sensitivity. GA addresses this by globally optimizing initial weights and biases through selection, crossover, and mutation, guided by a fitness function. This hybrid framework combines GA's global optimization with BP's nonlinear modeling, enhancing convergence speed and predictive accuracy, particularly for complex Olympic data like nonlinear and multicollinear patterns, enabling reliable medal count predictions.

2.2. Logistic Regression

Logistic regression is a binary classification method that predicts event probabilities (0 to 1) using the logistic function on a linear combination of input features. Parameters are optimized via Maximum Likelihood Estimation (MLE), with regularization (L1 for feature selection, L2 for smoother estimates) to prevent overfitting and improve generalization. Training involves coefficient initialization, iterative optimization through gradient methods, and hyperparameter tuning via cross-validation. Known for simplicity, interpretability, and robustness, logistic regression delivers probabilistic outputs, aiding nuanced decisions and effective predictor selection.

3. Methodology

Predicting multivariate outcomes requires addressing challenges such as nonlinear relationships, multicollinearity, and data heterogeneity, and this study introduces a multi-objective prediction framework that combines the advantages of GA-BP neural networks for continuous variables and logistic regression for binary classification. This dual-purpose structure ensures robustness and adaptability to various prediction tasks.Medal-winning countries: Include countries that have won at least one Olympic medal in history. These countries are concerned about the number of gold MEDALS or total MEDALS at future Olympics.

Continuous data prediction: GA-BP neural networks capture complex nonlinear interactions and improve accuracy and interpretability using feature selection techniques like Lasso regression and XGBoost. Genetic Algorithm (GA) optimizes the initial parameters to solve the problem of slow convergence and local optimum of BP network .

Binary Classification: The logistic regression model predicts the probability of the classification result. It uses advanced regularization techniques (L1, L2) to minimize overfitting, ensuring consistent performance across heterogeneous datasets.

Based on the above division, this study established two sub-models, and finally constructed Multi-objective prediction model in an integrated way to achieve the goal of prediction. The following describes the modeling idea of the two submodels in turn.

3.1. Preliminaries

We denote the intercept of the regression model as β_0, the coefficient of the i-th feature in the regression model as β_i,the regularization parameter in Lasso and Ridge regression as λ , the learning rate in the GA-BP neural network as η, the observed outcome for the i-th sample as y_i,the predicted outcome for the i-th sample as $\hat{y}_i$, and the area under the ROC curve, measuring model performance as AUC.

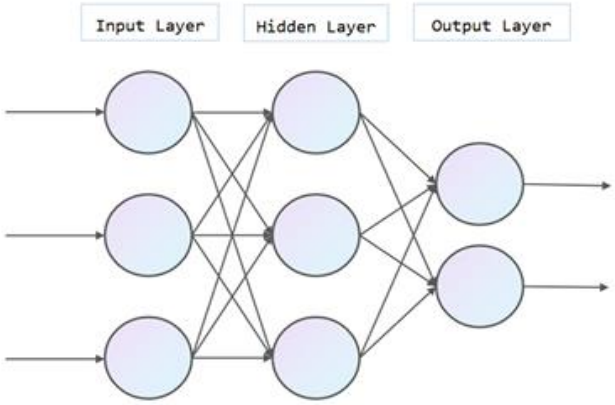


Figure 2. BP neural network

While Notations can be seen as above , to conduct the experiment, we use the medals of Olympics as the experimental subjects to predict the numbers of medels and gold medals of some main countries and the probability of a country which has not won a medal before to win the first one. The assumptions are made as below to make the model more reasonable according to the data we find.

(1) The proportion of male and female athletes in the competition will remain unchanged from 2024 to 2028: The athlete lineup will not change significantly in four years, so the proportion of male and female will not change significantly, and the actual situation is difficult to predict, so the calculation is simplified.

(2) The rate of change of the athletes' ratings is the same as that of the number of athletes from 2024 to 2028: The score of athletes is related to many reasons, the sum of the score of athletes has a large relationship with the number of athletes, and the average age of athletes changes little (new athletes

join and old athletes retire). The state of ordinary athletes is unpredictable, the number of medals is not stable; the actual situation is difficult to quantify and estimate.

(3) The age level of athletes is measured by the time since their first participation in the Olympic Games: The age of athletes when they first participate in the Olympic Games has little difference in general (the selection and training system of athletes is gradually mature).

3.2. Submodel I: Continuous Data Prediction Model based on GA-BP Neural Network

The prediction model based on GA-BP neural network aims to analyze and model the complex nonlinear relationship that affects the accuracy of multivariate outcome prediction. The model integrates data from multiple levels and effectively handles key predictor variables. In order to ensure the accuracy of feature selection, two methods are successively used for feature selection. Lasso regression [8] was first deployed followed by XGBoost [9]. After applying these two methods, the selected features are used as input to the GA-BP neural network. In order to ensure the accuracy of the selected variables, we will successively use two methods for screening. The first is Lasso regression, and the second is XGBoost. After two rounds of screening, the features that still exist are the selected feature variables.

When selecting the features, Lasso regression bounds the regression coefficients by adding the L1 norm (regularization term) to the loss function of ordinary least squares regression (OLS), which is formulated as follows.

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (1)$$

Here, the role of the regularization parameter λ is to make a trade-off between the model complexity and the fit. The purpose of introducing L1 regularization is to shrink the coefficients of unimportant variables or even shrink them to zero for variable selection and model simplification. In the analysis, in order to determine the optimal λ value, the cross-validation method is used to integrate multiple subsets by splitting the data for training and validation respectively, and the corresponding mean square error (MSE) is calculated. By testing multiple sets of candidate λ values, the value of λ that minimizes the average MSE is selected to ensure the effectiveness and robustness of the model. In this process, we eliminate the features with weak relationship with the output variable according to the regularization property of L1, so as to retain the features that contribute significantly to the model.

After that, we will perform XGBoost on the remaining variables to further screen the key features.

XGBoost regression integrates gradient boosting and decision tree algorithms to improve prediction performance, where the optimization objective includes both a loss function and a regularization term to control model complexity. The mathematical objective for a single round of XGBoost can be expressed as:

$$\mathrm{Obj}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

XGBoost progressively optimizes the loss function by adding new trees in forward stepwise. In each round, the algorithm will build a new decision tree based on the prediction residual of the current model to reduce the prediction error. During the tree building process, XGBoost calculates the importance of each feature in the model. The importance of a feature can be measured in many ways, such as based on the number of times it splits in the tree or based on its contribution to the target variable. As a result, less influential features will be selected less frequently during the tree splitting process. Thus, the features with less influence are filtered out.

After the 2 methods above ,we finally get the features that we need. Then we will put the relative data into GA-BP Neural Network, which can be seen clearly in Fig. 2.

GA-BP neural network combined with Genetic Algorithm (GA) to optimize the initialization parameters effectively solved the limitations of the traditional BP network, and provided a method

for prediction. The neural network architecture consists of two main stages: parameter initialization based on GA and iterative optimization based on BP. This hybrid framework handles the features just filtered. The network architecture shown in Fig. 3 consists of an input layer with strictly selected features, followed by a hidden layer where weighting calculations and activation functions capture nonlinear relationships that ultimately generate the predicted output. Initially, GA optimizes the weight W and bias b parameters of the network through processes inspired by natural selection, such as selection, crossover, and mutation. The optimization is guided by a fitness function that aims to minimize the mean squared error (MSE) over the training set:

$$F = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_{GA-BP}(X_i; W, b)\right)^2 \tag{3}$$

Where $f\_(GA-BP)$ is the predicted output, y_i is the actual target, X_i represents the input features, and n denotes the number of training samples.

With the initialized parameters $W\text{^}init$ and $b\text{^}init$, the BP training phase refines these values using gradient descent to minimize the following loss function:

$$L(W, b) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_{BP}(X_i; W, b)\right)^2 \tag{4}$$

Where the gradients with respect to weights and biases are calculated as:

$$\nabla_W L = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - f_{BP}(X_i; W, b)\right) \cdot \frac{\partial f_{BP}}{\partial W} \tag{5}$$

$$\nabla_b L = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - f_{BP}(X_i; W, b)\right) \cdot \frac{\partial f_{BP}}{\partial b} \tag{6}$$

The weights and biases are iteratively updated following the rule:

$$W^{(t+1)} = W^{(t)} - \eta\nabla_W L, \quad b^{(t+1)} = b^{(t)} - \eta\nabla_b L \tag{7}$$

Where t is the iteration index and η is the learning rate. This iterative process ensures the net work converges to an optimal solution, effectively capturing the nonlinear relationships governing data distributions. The model's predictive output can be expressed as:

$$\hat{y} = f_{GA-BP}(X; W^*, b^*) \tag{8}$$

which tells us about the results of prediction.

3.3. Submodel II: Binary Classification Prediction Model based on Logistic Regression

The logistic regression model is designed to analyze and predict binary outcomes by effectively capturing nonlinear relationships among critical factors impacting the target variable. By integrating data from multiple levels, the model generates probabilistic estimates for classification tasks. To ensure precise feature selection, Lasso regression is first applied to remove irrelevant features by reducing their coefficients to zero, thereby simplifying the model and retaining variables with substantial contributions. Following this, XGBoost is used to further optimize the feature set through gradient boosting and decision tree algorithms, highlighting the most impactful predictors. The refined features obtained from these processes are then utilized as inputs to the logistic regression model, ensuring robustness and high accuracy in binary classification predictions.

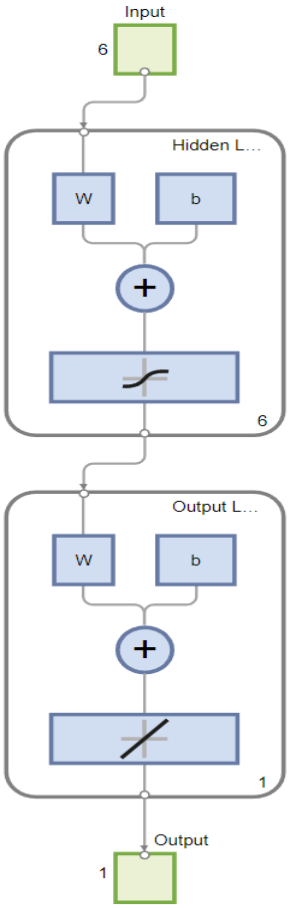


Figure 3. neural network hidden layer

Lasso regression, XGBoost and other methods are used for feature selection to eliminate redundant features, and the final results are used as input to the logistic regression framework. This approach provides not only a binary classification (whether to win a medal or not) but also a probabilistic output highlighting the likelihood of winning a medal for the first time. Submodel II uses logistic regression to predict the probability of not obtaining 0 and 1 in the binary classification problem. As a widely used binary classification model, logistic regression maps a linear combination of input features to a probability value between 0 and 1 through the logistic function.

The probability $P(y\_i = 1 \mid x\_i)$, where $y\_i = 1$ is calculated as:

$$P(y_i = 1|x_i) = h(x_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \tag{9}$$

Where β_0 is the intercept, β_1 is a vector of coefficients, and $x\_i = [x\_i1, x\_i2, \dots, x\_ik]$ represents the feature vector for the i-th observation.

Parameters β_0 and β_1 are estimated by maximizing the likelihood of the observed data. The likelihood function is given as:

$$L(\beta) = \prod_{i=1}^{n} \left[h(x_i)^{y_i} \cdot \left(1 - h(x_i)\right)^{1-y_i}\right] \tag{10}$$

and its corresponding log-likelihood function becomes:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \log h\,(x_i) + (1 - y_i) \log\left(1 - h(x_i)\right)\right] \tag{11}$$

The gradients of the log-likelihood with respect to the coefficients are derived as:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left(y_i - h(x_i)\right) x_{ij}\,, \quad j = 0,1, \dots, k \tag{12}$$

Where x_ij is the j-th feature value for the i-th observation. These gradients are used to iteratively update the parameters via gradient ascent.

To address overfitting, the logistic regression model incorporates regularization terms. This study applies both ℓ_1-norm (Lasso) and ℓ_2-norm (Ridge) penalties. The regularized objective function is expressed as:

$$\mathcal{L}_{\text{regularized}}(\beta) = -\ell(\beta) + \lambda_1 \sum_{j=1}^{k} |\beta_j| + \frac{\lambda_2}{2} \sum_{j=1}^{k} \beta_j^2 \quad (13)$$

Where λ_1 and λ_2 are the regularization parameters controlling the strength of the penalties. Feature selection is facilitated by the ℓ_1-norm, which reduces insignificant coefficients to zero.

Based on the feature selection, variables are included in the logistic regression model. The logistic regression model training involves:

(1) initialization of coefficients β_0 and β.

(2) Iterative optimization based on gradient ascent and regularization to maximize the objective function.

(3) Validation and tuning of hyperparameters (e.g., regularization strength λ_1, λ_2) using cross-validation.

The trained model outputs the probability of each country winning its first medal as:

$$\widehat{y_i} = P(y_i = 1|x_i) = h(x_i) \quad (14)$$

Examples with(〚 y〛_i)^ > 0.5 are classified as 1.

4. Experiment

4.1. Experimental Settings

4.1.1. Dataset

We use 4 datasets listed below:

(1) Data_Dictionary dataset —— database descriptions with examples.

(2) SummerOly_Athletes dataset —— all competitors with their sport, year, and result (medal type ornone).

(3) SummerOly_Medal_Counts dataset —— complete country medal count tables for all summerOlympics from 1896 to 2024.

(4) SummerOly_Hosts dataset —— list of host country for all summer Olympics from 1896 to 2032.

4.1.2. Hyper-parameters

We use GA to optimize the hyperparameters of the BP neural network, get the optimal initial threshold weight: the weight parameter between the input and hidden layer, the bias of the hidden layer neurons, the bias between the hidden layer and the output layer and the bias of the output layer neurons.

4.2. Empirical Results

4.2.1. Results of Submodel I

Continuous Data Prediction Model based on GA-BP neural network aims to predict the total number of MEDALS and gold MEDALS of the medal-winning countries in the 2028 Olympic Games. The model integrates athlete-level and country-level data to analyze and model the complex nonlinear relationships that affect Olympic performance. The key features of the input include gender ratio,

duration of IOC membership, number of athletes, host country effect, number of dominant sports, and athlete rating. First, in order to get country-level data, we need to rate the athletes in each country. We used a combination of objective and subjective weighting methods to calculate the athlete score, where the CRITIC method of objective weighting considers the contrast and correlation between indicators, and the subjective weighting is the expert evaluation method. These two methods are averaged to generate balanced and reliable athlete scores.

Based on the existing data, we will score the comprehensive strength of the athletes based on three indicators: the service time of the athletes, the historical number of cards won by the athletes in the project, and the number of participation times of the athletes. In the expert weighting method, we set the weights of the above three indicators as 0.4, 0.3 and 0.3 respectively based on the existing research results [10]. Based on the weights above, we will calculate the athlete's score subjectively.

By using SPSSPRO, we calculate the results of weights of the 3 key features we selected. The results are shown as Table I:

Table 1. Results of CRITIC weight method

Formula	Weight
Time-len	0.594
Medal-count	0.204
Participation-count	0.202

We then aggregate athlete competitive status scores to countries to calculate an overall "strength indicator" for the country. Then we will select the key features to evaluate country using Lasso Regression and XGBoost .

As is Fig. 4.a) shown, after testing multiple sets of candidate λ values, the λ value that minimizes the average MSE, which is determined to be 0.02 in this analysis, is chosen, thus ensuring the validity of the model.

Combining Fig. 4 b) and Table II, we retained six key features after Lasso regression and cross-validation: sex ratio, duration of IOC membership, number of athletes, host country effect, number of dominant sports, and athlete score. Fig. 4 b) shows that these variables do not shrink to zero, indicating their significant explanatory power. Table II shows the standardized coefficients (-0.757,-0.134, 0.343, 7.572,-0.113, 0.006) and non-standardized coefficients (-0.130,-0.124, 0.336, 10.014,-0.094, 0.004), reflecting their different influence degrees. The unstandardized coefficient of the host country is the largest (10.014), indicating that the home field effect is significant, while other variables, such as athlete performance and dominant events, have less influence. The retained variables form the basis of the model normalization formula: $y = 1.686 - 0.757 \cdot x\_1 - 0.134 \cdot x\_2 + 0.343 \cdot x\_3 + 7.572 \cdot x\_4 - 0.113 \cdot [\![x]\!]\_5 + 0.006 \cdot x\_6$.

The x_1 equals with data of sex ratio, others are alike.

Table 2. Regreesion Coefficents and R^2 Value

Variable Name	Standardized Coefficient	Unstand-ardized Coefficient	**R^2**
Distance	1.686	0.763	0.749
sex ratio	-0.757	-0.130	
time of being IOC membership	-0.134	-0.124	
number of athletes	0.343	0.336	
host country effect	7.572	10.014	
number of dominant sports	-0.113	-0.094	
athlete rating	0.006	0.004	

These results not only verify the effectiveness of Lasso regression for high-dimensional feature selection, but also further show the important role of the retained variables in the model for the prediction of the target value.

Then ,we will use XGBoost to do the second selection. After we set parameters, we then have the results of screening, shown as Fig. 5:

To address the lack of direct data for the 2028 Olympic Games, some adjustments were made to ensure the accuracy of the model. Firstly, the LSTM network is used to predict the number of athletes, and historical trends are used to make reliable predictions. Athlete scores are then adjusted equally to the projected number of athletes, a process that assumes that an increase or decrease in the

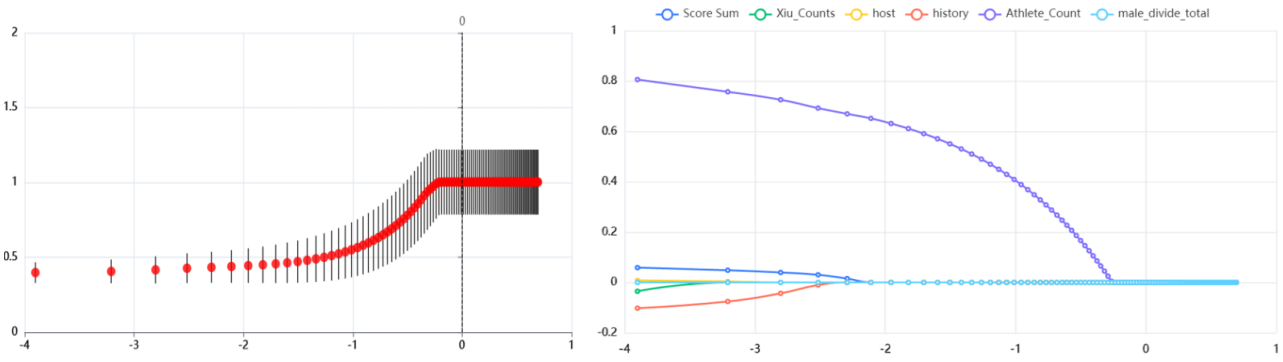
total number of athletes will directly affect a country's competitive strength, thus aligning the ranking with the latest participation level.

The host country indicator assigns a value of 1 to the United States, the 2028 host country ,and 0 to other countries, taking into account host country strengths such as local support and infrastructure investment. In addition, the membership period of all countries in the IOC will be increased by four years, which is in line with the objective time logic.

The sex ratio and the number of major sports were kept constant because historical data showed that these metrics were generally stable, avoiding random changes that could reduce the reliability of the predictions. These adjustments ensure that projections for 2028 are realistic and robust.

With logical adjustments and robust prediction techniques to handle missing data, the model accurately integrates the dynamic interactions between variables that influence Olympic success. Based on the results of the Multi-objective prediction model, we derive the gold and total MEDALS forecast values for each country in the 2028 Olympic Games and provide forecast intervals to quantify the uncertainty. This not only helps analyze model performance, but also identifies countries that are likely to move significantly in the medal table, which can be seen as Table III and Table IV.

Based on the updated results, the United States are expected to maintain their dominance in the MEDAL table, with projected total MEDALS ranging from 129 to 158 respectively. Countries likely to improve include:United States, Australia, Italy, France and German. Countries likely to decline are:China , Great Britain, South Korea, Japan and Netherlands.



a) Lasso Cross regression validation plot b) Plot of model regression coefficients

Figure 4. Lasso Regression Results

This analysis relies on forecast intervals, which highlight the variability and uncertainty as- sociated with each country's key campaigns, strategic planning, and external factors such as the advantage of the host country.

4.2.2. Results of Submodel II

We adopted the same feature screening method as the submodel, and learned that all four variables were preserved.

Then, Based on the provided data and the output of the logistic regression model, we analyzed the probability of countries winning their first Olympic MEDALS. The predictions are structured to assess the probability of each country, integrating the predicted probabilities and threshold based classification to gain meaningful insights.

As shown in Fig. 6, the graph visualizes the predicted probabilities for all samples belonging to the positive class (y = 1, representing the winning country) with respect to a given classification threshold. Countries like Guinea (GUI), above the decision threshold (usually set at 0.5), are expected to have a high probability of winning their first medal. GUI, labeled y = 1 in the dataset, verifies that it has won its first medal with historical data. Conversely, other countries such as Bhutan (BHU), Antigua and Barbuda (ANT), and Rwanda (RWA) are well below the threshold, indicating their low potential in the upcoming Olympic cycle.

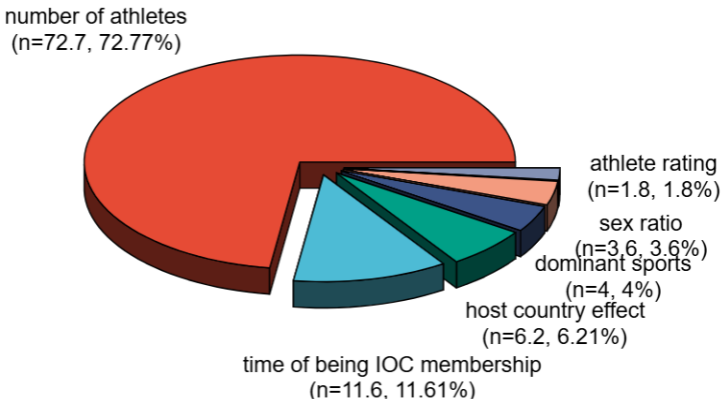


Figure 5. plot of XGBoost feature importance

Table 3. Predicted Total Medal Counts for the 2028 Olympic Games

NOC	2024 Total	min	max
United States	126	129	158
China	91	56	91
Japan	45	17	51
Australia	53	63	98
France	64	95	129
Netherlands	34	7	41
Great Britain	65	6	41
South Korea	32	0	25
Italy	40	46	80
Germany	33	51	86

Table V complements this by showing predicted (y_{pred}) and actual (y) outcomes. The model accurately identifies high-potential countries like GUI and provides actionable insights for marginal cases near the threshold, suggesting opportunities for strategic investments in infrastructure or training.

While the model demonstrates strong predictive power, misclassifications (e.g. red points above or blue points below the threshold) highlight areas for improvement. Incorporating geopolitical or economic indicators could enhance predictions for low-rated countries like Eswatini (SWZ) and Solomon Islands (SOL).

4.3. Sensitivity and Evaluation

4.3.1. Submodel I

In our sensitivity analysis, we defined a 20% perturbation for the independent variables to evaluate the robustness of our model. Due to the nature of certain variables, such as the binary nature of the

host country indicator and the constrained variability of the sex ratio, these variables were excluded from the analysis. The remaining key predictors—number of athletes, athlete rating, duration of IOC membership, and number of dominant sports—were subjected to 20% perturbations, and their impacts on the dependent variable y were analyzed.

As is shown in Table VI ,the baseline value of y (origin) was recorded as 7.1694. A 20% increase in the number of athletes resulted in a significant change in y, increasing it to 9.2171, corresponding to a rate of change of approximately 28.56%, highlighting its strong influence on the model's output.

Table 4. Predicted Gold Medal Counts for the 2028 Olympic Games

NOC	2024 Total gold	min gold	max gold
United States	40	50	63
China	40	17	30
Japan	20	3	16
Australia	18	23	36
France	16	47	61
Netherlands	15	0	14
Great Britain	14	0	13
South Korea	13	0	9
Italy	12	12	25
Germany	12	15	28

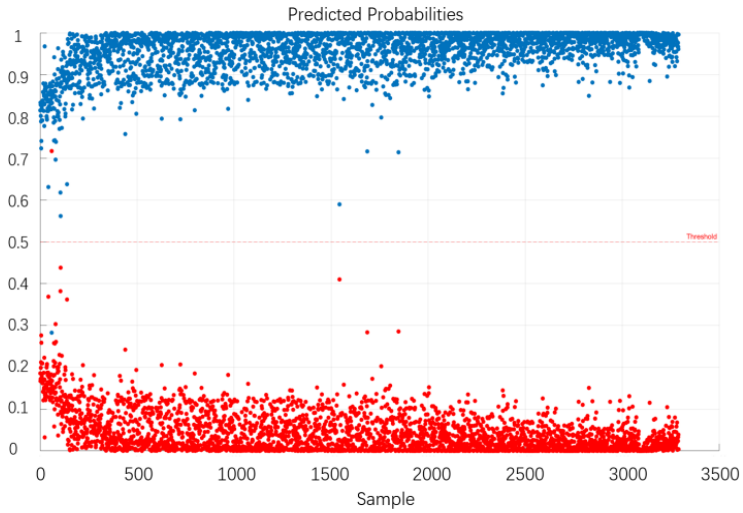


Figure 6. Visualization of Predicted Probabilities

Table 5. Predicted and Actual Medal Outcomes

Country	Predicted (ypred)	Probability(P)
ANT	0	0.2093
BHU	0	0.3688
GUI	1	0.7173
IVB	0	0.2608
MON	0	0.3621
RWA	0	0.1236
SOL	0	0.0333
SWZ	0	0.0215
VAN	0	0.0061

Athlete rating caused a moderate variation in y, increasing it to 7.5218 with a rate of change of 4.91%. The duration of IOC membership exhibited minimal sensitivity, with a resulting y of 7.1963 and a negligible rate of change of 0.38%. Similarly, dominant sports showed a noticeable effect, increasing y to 8.0173 and yielding a rate of change of 11.83. The performance curve can be seen as Fig 7.

These results confirm that the number of athletes and the number of dominant sports exert the most substantial influence, underscoring their critical roles in determining medal outcomes. Hence, model predictions are particularly sensitive to these inputs.

4.3.2. Submodel II

Receiver Operating Characteristic (ROC) curves evaluate logistic regression models by plot- ting the true positive rate (TPR) versus false positive rate (FPR) across thresholds, highlighting the sensitivity-specificity tradeoff. The Area Under The Curve (AUC) summarizes the discriminative power of the model.

Table 6. Sensitivity Analysis Results

Variable	y	Rate of Change
Origin	7.1694	–
Number of Athletes	9.2171	0.2856
Athlete Rating	7.5218	0.0491
IOC Membership Time	7.1963	0.0038
Dominant Sports	8.0173	0.1183

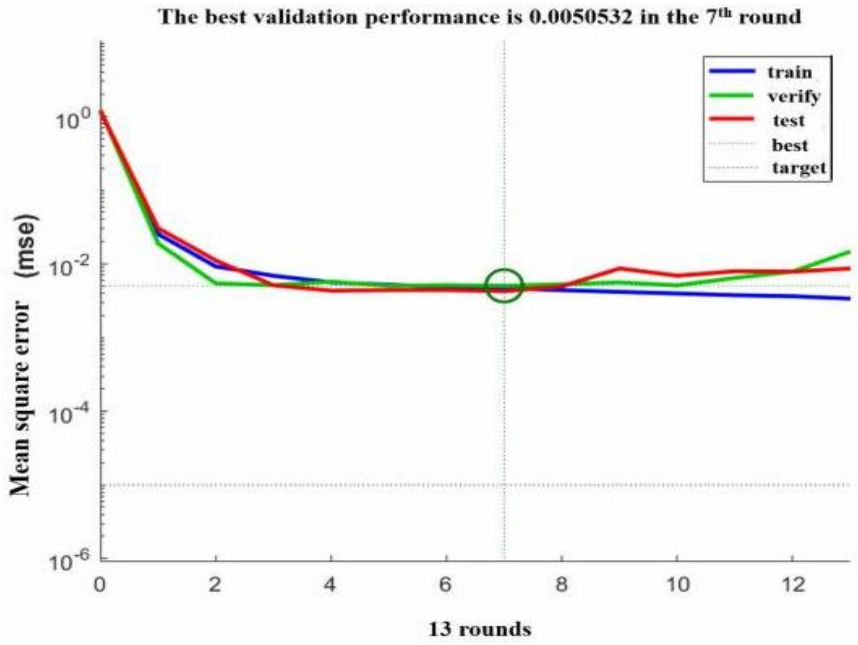


Figure 7. Sensitivity Analysis Results

As is seen in Fig 8, the AUC of this graph is 0.78558 (higher than 0.75), showing a strong performance, which is significantly better than random guessing, and can effectively capture the data pattern. The curves show an improvement in TPR with little increase in FPR, which indicates robustness on imbalanced datasets (e.g., predicting the country likely to win the first medal). However, the curve flattens slightly at higher FPRs, indicating reduced sensitivity to some samples.

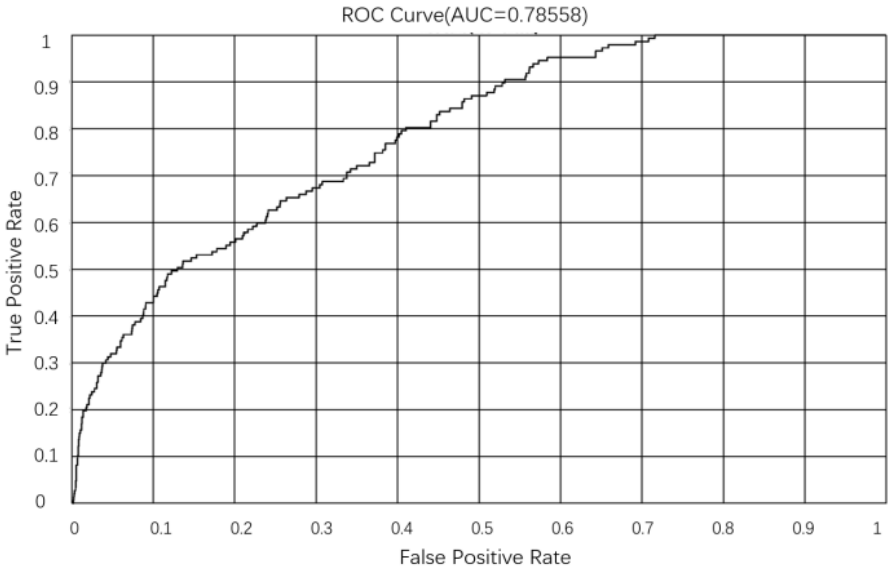


Figure 8. ROC curve

5. Conclusion

This study presents a multi-objective prediction model that successfully integrates GA-BP neural network and logistic regression approaches to address complex prediction challenges. First, the study establishes a comprehensive feature selection process by combining Lasso regression and XGBoost methods, which effectively identifies key predictive variables while eliminating redundant features, thereby improving model robustness and adaptability to high-dimensional data. Second, the GA-BP neural network submodel incorporates genetic algorithms to optimize global parameters, addressing traditional BP network limitations such as slow convergence and susceptibility to local optima, thereby significantly enhancing its capacity for continuous predictions. Third, the logistic regression submodel employs advanced regularization techniques (L1 and L2), ensuring reliable binary classification predictions and robust generalization across diverse datasets.

The integrated framework demonstrates its value through effective feature engineering, dual-task optimization for continuous and categorical outcomes, and global parameter optimization utilizing evolutionary algorithms. Comprehensive sensitivity analyses and ROC curve evaluations have validated the model's practicality and predictive performance. Despite its achievements, future research could explore incorporating additional dynamic indicators, leveraging deep learning techniques, and enhancing methods for handling heterogeneous and imbalanced datasets to further augment prediction accuracy and expand the framework's applications across other complex systems requiring multi-objective predictions.

References

[1] X. Xie. Research on Olympic Performance Prediction Using Grey Prediction Model [J]. Electronic World, 2018, (02): 48–49.

[2] Anthony S. Leicht, Miguel A. Gomez, Carl T. Woods. Team Performance Indicators Explain Outcome During Women's Basketball Matches at the Olympic Games. [J], Sports (Basel, Switzerland), 2017, 5 (4): 96-96.

[3] Sadeq D. Al-Majidi, Maysam F. Abbod, Hamed S. Al-Raweshidy. A Particle Swarm Optimisation-Trained Feedforward Neural Network for Predicting the Maximum Power Point of a Photovoltaic Array [J], Engineering Applications of Artificial Intelligence, 2020, 92: 103688-103688.

[4] Weihua W., Rui C., Yuantong L., Ayodele D. A., Changyong Y., Zengtao C., et al. Prediction of Tool Wear Based on GA-BP Neural Network [J]. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 2022, 236 (12): 1564–1573.

[5] S. Zhang, Z. Huo, C. Zhai. Building Carbon Emission Scenario Prediction Using STIRPAT and GA-BP Neural Network Model [J]. Sustainability, 2022, 14 (15): 9369-9369.

[6] F. C. Pampel. Logistic Regression: A Primer [M]. msra, 2021.

[7] Emily C Z, Chandana A R, Rahul D T, Sujata P, et al. Logistic Regression in Clinical Studies [J], International Journal of Radiation Oncology Biology Physics, 2021, 112 (2): 271-277.

[8] Archana J. M., Valerie P., Rishabh S., Anand P., et al. Logistic LASSO Regression for Dietary Intakes and Breast Cancer [J]. Nutrients, 2020, 12 (9): 2652-2652.

[9] Ahmedbahaaaldin I. A. O., Ali N. A., Ming F. C., Yuk F. H., Ahmed E. Extreme Gradient Boosting (xgboost) Model to Predict the Groundwater Levels in Selangor Malaysia [J]. Ain Shams Engineering Journal, 2021, 12 (2): 1545–1556.

[10] A. Perperoglou, M. Huebner. Quantile Foliation for Modelling Performance Across Body Mass and Age in Olympic Weightlifting [J]. Statistical Modelling, 2020, 21 (6): 546–563.