An Exploration of Data Prediction Based on Multiple Linear Regression Models and Bayesian Regression

Haoyu Wang *, Guangyan Wang and Jingyu Huang

College of Information Science and Engineering, Hohai University, Changzhou, China

* Corresponding Author Email: 417397253@qq.com

Abstract. In this paper, a hybrid modeling framework integrating multiple linear regression, Bayesian regression and decision tree is proposed for target variable prediction. First, the quantitative prediction of target variables is realized by constructing a multivariate linear regression model with a five-dimensional feature space, and the model is evaluated for performance using F-test and R^2 value to verify the significance and explanatory power of the combination of independent variables. Secondly, Bayesian regression method is introduced to deal with the binary prediction task, which improves the uncertainty quantification ability through probabilistic modeling and optimizes the feature screening process by combining with the decision tree algorithm, and finally outputs the target breakthrough probability distribution. Finally, the dominant feature dimension is extended to seven dimensions through feature engineering, and the improved model significantly improves the prediction accuracy while maintaining the computational efficiency, which verifies the key role of feature selection on the model performance. The experimental results show that the hybrid model framework has good generalization ability and interpretability in complex prediction scenarios.

Keywords: Multivariate linear regression, Bayesian regression, decision trees.

1. Introduction

In this paper, a combined model based on multiple linear regression [1], Bayesian regression [2] and decision tree [3] is proposed to cope with time series forecasting and classification situations, and to improve the forecasting performance through feature engineering and model optimization. First, a multiple linear regression model is constructed, five feature variables are introduced, the regression coefficients are solved by the least squares method [4], and the model fit is assessed using the F-test and R^2, and the results show that the model has high accuracy and explanatory power. Second, for the binary classification case, a Bayesian regression model is used to deal with uncertainty [5] and combined with a decision tree classification method to optimize the prediction results, and the classification accuracy is improved by feature screening and integrated learning [6]. Finally, the model optimization is carried out by introducing domain-specific features, and the correlation coefficient matrix is used to analyze the correlation between the features to further improve the prediction ability of the model, and the experimental results show that the improved model has significant improvement in both fit and generalization ability.

2. Medal and Gold Medal Projections

In this model construction, five predictor variables were introduced. They are the number of participants, the winning percentage of the first five matches, the number of winners, the trend of the winning percentage and the impact of home country.

2.1. Predictive Model Construction

So, in order to give more accurate weights to the predictor variables and get the predicted results, this paper decided to use multiple linear regression model. In this paper, we decided to predict the number of gold, silver, bronze and total medals based on the previous assumptions:

$$M = G + S + B \tag{1}$$

As long as gold, silver and bronze medals are predicted separately, these predictions can be added together to get the predicted total number of medals. In the previous data processing, 5 advanced data were eventually obtained as independent variables, so a five-member linear regression model was constructed in predicting the amount of gold/silver/bronze. In predicting the total number of medals, three five-member linear regression models were actually combined together to obtain the total number of medals.

Take the construction of the prediction model for the gold medal table as an example. In the previous data processing, five advanced data were finally obtained as independent variables, so a five-member linear regression model was constructed. Therefore, the regression plane formula is as follows:

$$w = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{2}$$

In the formula, x_1 denotes the number of participants in the last semester, x_2 denotes the average medal rate in the last three semesters, x_3 denotes the number of athletes awarded in the last semester, x_4 denotes the trend of medal rate, x_5 denotes the influence of the home country, $\beta_1 ... \beta_5$ denotes the regression coefficients of the different independent variables, and β_0 corresponds to the intercept value.

It can be extended as:

$$\begin{cases} W = X\beta + \varepsilon \\ E(\varepsilon) = 0, COV(\varepsilon, \varepsilon) = \sigma^2 I_n \end{cases} \tag{3}$$

In Eq. $W = X\beta$ is the matrix expression, ε is the error term, $E(\varepsilon) = 0$ denotes the expected value of the error term is 0, $COV(\varepsilon, \varepsilon) = \sigma^2 I_n$ denotes the covariance of the surface error term is σ^2, and n is the amount of data, here n is 1980.

It is possible to simplify this model to $(Y, X\beta, \sigma^2 I_n)$. Expand Y, X, β, ε can be obtained:

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{1979} \\ w_{1980} \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{14} & x_{15} \\ 1 & x_{21} & x_{22} & \cdots & x_{24} & x_{25} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{19791} & x_{19792} & \cdots & x_{19794} & x_{19795} \\ 1 & x_{19801} & x_{19802} & \cdots & x_{19804} & x_{19805} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{1979} \\ \varepsilon_{1980} \end{bmatrix} \tag{4}$$

In this section, we need to estimate the estimates of the parameters of the multiple linear regression model $\beta_0, \beta_1 ... \beta_5$, which are solved by least squares, i.e., by solving for the sum of squares of their deviations:

$$Q = \sum_{i=1}^{n} \left(w_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_5 x_{i5} \right)^2 \tag{5}$$

Repeating the above steps, the silver and bronze medal rankings can also be modeled, and their regression coefficients calculated. Finally, the regression coefficients for the gold, silver and bronze medal rankings can be solved as shown in table 1 below.

Table 1. Gold/ Silver/ Bronze Medalist Factor

	β_0	β_1	β_2	β_3	β_4	β_5
Gold Medal	-0.027	0.0034	49.91	0.6891	-0.0026	9.369
Silver Medal	-0.1202	0.008	88.99	0.4279	-0.1088	6.201
Bronze Medal	-0.0806	0.0099	15.31	0.6277	-0.1858	3.325

Substituting the regression coefficients into them, the empirical regression formulas for predicting the gold medal standings, silver medal standings, and bronze medal standings can be obtained, respectively. Since x_2, x_3, x_4 these three models are different, subscripts are used to distinguish them separately. Finally, the empirical regression equation obtained is as follows:

$$w_G = -0.027 + 0.0034x_1 + 49.91x_{2G} + 0.6891x_{3G} - 0.0026x_{4G} + 9.369x_5 \tag{6}$$

$$w_S = -0.1202 + 0.008x_1 + 88.99x_{2S} + 0.4279x_{3S} - 0.1088x_{4S} + 6.201x_5 \tag{7}$$

$$w_B = -0.0806 + 0.0099x_1 + 15.31x_{2B} + 0.6277x_{3B} - 0.1858x_{4B} + 3.325x_5 \tag{8}$$

By substituting the gold medal prediction, silver medal prediction and bronze medal prediction into the larger model, a prediction model for the total number of medals can be obtained:

$$\begin{aligned} W_T = &-0.2278 + 0.0213x_1 + 49.91x_{2G} + 88.99x_{2S} + 15.31x_{2B} + 0.6891x_{3G} + 0.4279x_{3S} \\ &+0.6277x_{3B} - 0.0026x_{4G} - 0.1088x_{4S} - 0.1858x_{4B} + 18.895x_5 \end{aligned} \tag{9}$$

2.2. Model Evaluation

In order to evaluate the indicators of the established multiple regression linear model, including its accuracy, variance and other data, this section adopts the most classic F-test method, which can decompose the total change TSS into regression sum of squares and ESS and residual sum of squares and RSS, and can assess the goodness of fit through the significance of the overall linear relationship of the model. the ESS and RSS formulas are as follows:

$$\begin{aligned} ESS &= \sum \left(\hat{w}_t - \bar{w}\right)^2 / \sigma^2 \sim \chi^2(k) \\ RSS &= \sum \left(\bar{w} - \hat{w}_t\right)^2 / \sigma^2 \sim \chi^2(n-k-1) \end{aligned} \tag{10}$$

The relationship between F and the coefficient of determination R^2 is as follows:

$$F = \frac{ESS/k}{RSS/(n-k-1)} = \frac{n-k-1}{k} \cdot \frac{ESS}{TSS - ESS} = \frac{n-k-1}{k} \cdot \frac{ESS/TSS}{1 - ESS/TSS} \tag{11}$$

$$F = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2} \tag{12}$$

This gives $R = 1 \Rightarrow F = \infty, R = 0 \Rightarrow F = 0$, so F and the coefficient of determination R^2 are in fact the same. In order to more fully assess the accuracy and explanatory power of the model, it was decided in this section to calculate R^2、F、P、s^2 the four parameters of the model.

Substituting the coefficients and obtaining the following results as in table 2.

Table 2. Model Assessment Parameters

	R^2	F	P	s^2
Gold Medal	0.8794	2236.8	0	2.730
Silver Medal	0.8265	1461.1	0	3.065
Bronze Medal	0.8295	1492.8	0	3.293

The results were analyzed according to the types of medals.

The correlation coefficient R^2 of the gold medal numerical prediction model is 0.8794, which indicates that the established gold medal numerical prediction model can explain about 87.94% of the changes in the dependent variable, which means that the established model adopts a more comprehensive set of independent variables, and most of the influences on changes in the dependent variable are taken into account. The F-value of the model is as high as 2236.8, which indicates that the combined effect of the independent variables on the dependent variable in the prediction model is very significant and indeed consistent with the correlation coefficient. The P-value is the probability of occurrence of the observed sample data if the original hypothesis is true. If the P-value is less than a predetermined level of significance, the result is considered statistically significant at 0 and the original hypothesis is rejected. Finally, the residual variance of the model s^2, which reflects the magnitude of variation that cannot be explained by the model, was 2.730, indicating that the degree of variation is small, and the model fits the data well.

In the same manner, the silver number prediction model and the copper number prediction model were analyzed. The correlation coefficient R^2 of the silver number prediction model is 0.8265, F is 1461.1, P is 0, and the residual variance s^2 of the model is 3.065. The correlation coefficient R^2 of the bronze number prediction model is 0.8295, F is 1492.8, P is 0, and the residual variance s^2 of the model is 3.293, and the model is characterized by a high correlation coefficient overall, with an F value, a P value of 0, and a relatively low residual variance. This indicates that the model has a high degree of fit.

In summary, the model is suitable for gold medal prediction, silver medal prediction and bronze medal prediction at the same time, with good performance and high fit. The best performance is the gold medal prediction model, while the silver medal prediction model and the bronze medal prediction model have similar performance and slightly lower fit than the gold medal prediction model.

3. Medal Count Predictions for the 2028 Summer Olympics in Los Angeles

This section involves predicting the number of medals for the 2028 Summer Olympics in Los Angeles based on an established model with prediction intervals and analyzing which countries will improve and which countries will decline compared to the 2024 Summer Olympics in Paris based on the predictions.

This section constructs a multiple regression linear model. It is a comprehensive large-scale model with nested models, built on the assumption that total medals = gold + silver + bronze. It has three nested five-element regression linear models for predicting the number of gold, silver, and bronze medals that add up to the total number of medals.

The variables required in the models were calculated and replaced when predicting the number of medals at the 2028 Summer Olympics in Los Angeles. Gold medals were used as the first basis of judgment and silver medals were used as the second basis of judgment.

In terms of the number of medals, the United States saw a relatively small change in the total number of medals but a significant increase in the number of gold medals, from 40 to 46, which may be due to the influence of home country influences, as the U.S. may add more favorable events to win more medals. China dropped from 40 to 35 medals, a decrease of 11 medals in total, which may be due to a slight decrease in the number of dominant events. On the other hand, Japan, Australia, France, Great Britain and the Netherlands are all within reasonable limits in terms of gold and total medals.

Overall, the fluctuations in the rankings and medal counts for each country are within reasonable limits and can be strongly explained by the developed model. In conjunction with the projections for the Los Angeles Olympic medal standings, it is hypothesized that the country's most likely to make progress are the United States, Great Britain, Germany, Canada, and Spain, which are likely to improve the number and quality of their medals or be in the rankings. The countries most likely to

lose are China, the Netherlands, South Korea, New Zealand, and Hungary, which are likely to fall in the number and quality of medals or in the rankings.

4. Analysis of Countries Winning First Gold Medal

This section entails first counting the countries without a winning card and predicting which countries are likely to make a breakthrough at the 2028 Summer Olympics in Los Angeles based on the model and assessing the likelihood of making the prediction.

4.1. Feature Engineering Extraction

Extract data on the number of athletes, the number of sports, the number of Olympic Games since 1984, and the number of athletes with Olympic experience for each year from 2012 to 2024 for the countries that have succeeded in making a breakthrough and those that have not yet made a breakthrough.

4.2. Model Structure

For the prediction of Type 0/1 results, the prize of the year is used as the dependent variable, and the number of athletes competing each year, the number of sports, the number of Olympic Games since 1984, and the number of athletes with Olympic experience are used as the independent variables.

The Bayesian linear regression model was constructed through the process of similar problem - similar multiple linear regression model. Except that residual ε obeys Bayes' rule. Therefore, the regression plane formula is as follows:

$$w = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \quad where \ \ \varepsilon \sim N(\mu, \sigma_\varepsilon^2) \tag{13}$$

In the formula, x_1 denotes the number of athletes in the last Olympics, x_2 denotes the number of events participated in the last Olympics, x_3 denotes the number of participants in the Olympics since 1984, x_4 denotes the total number of Olympic athletes in history, $\beta_1 \sim \beta_4$ denotes the regression coefficients of different independent variables, and β_0 refers to the value of the intercept.

The regression coefficients were solved as shown in table 3 below.

Table 3. Model Regression Coefficient

β_0	β_1	β_2	β_3	β_4
-2.2033	0.0632	0.1260	0.0875	-0.3233

In this section, it is necessary to predict whether a country that has never won an Olympic medal will win an Olympic medal in Los Angeles in 2028. In the modeling process, similar features are extracted from countries that have never won a medal and those that have won a medal, and the breakthrough probability of countries that have not won a medal is determined by analyzing the features of the winning countries.

Using x_1, x_2, x_3, x_4 2028 corresponding data for each non-winning country as input, a Bayesian linear regression model was used to predict whether they would win a medal at the 2028 Olympics. For testing purposes, the training and prediction results for each country are shown in Fig. 1 below.

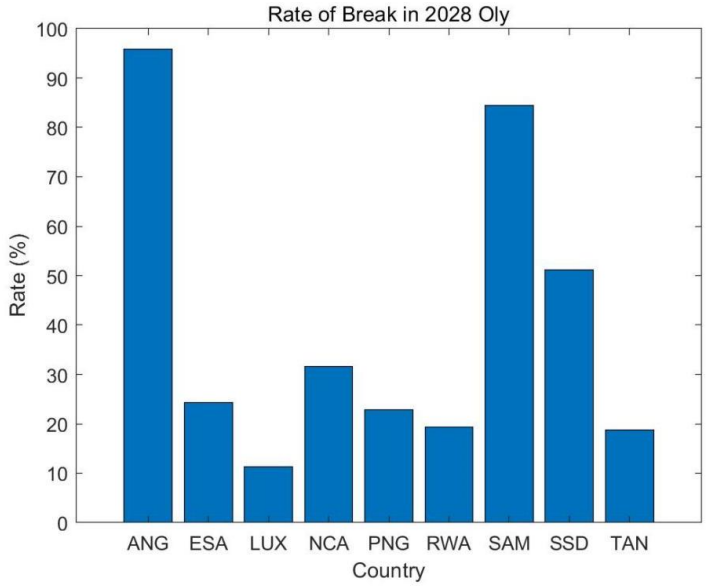

Figure 1. Rate of break in 2028 Oly

It can be seen that the countries of Angola and Samoa have a relatively high probability of winning with 97.3% and 84.9% respectively. However, there is a major problem with this method: the R-value is too large, indicating a high degree of uncertainty. Although the calculation of the probability of winning the Olympics is complex, there is a great deal of uncertainty in the tactical organization and improvisation of each country, and the prediction itself is so uncertain that it is very difficult to predict very accurately, therefore, the decision tree method was attempted.

Therefore, we try to use the decision tree method. The number of athletes in the past, the number of events in the past, the number of participants in the Olympic Games since 1984 are taken as the independent variables (the number of athletes with Olympic experience is deleted), and whether or not they won the prize in the current year is taken as the dependent variable, and the decision tree is generated by using the function $fitrtree\ (X, Y, \text{var arg}\, in\ )$ regression, which is shown in Fig. 2.

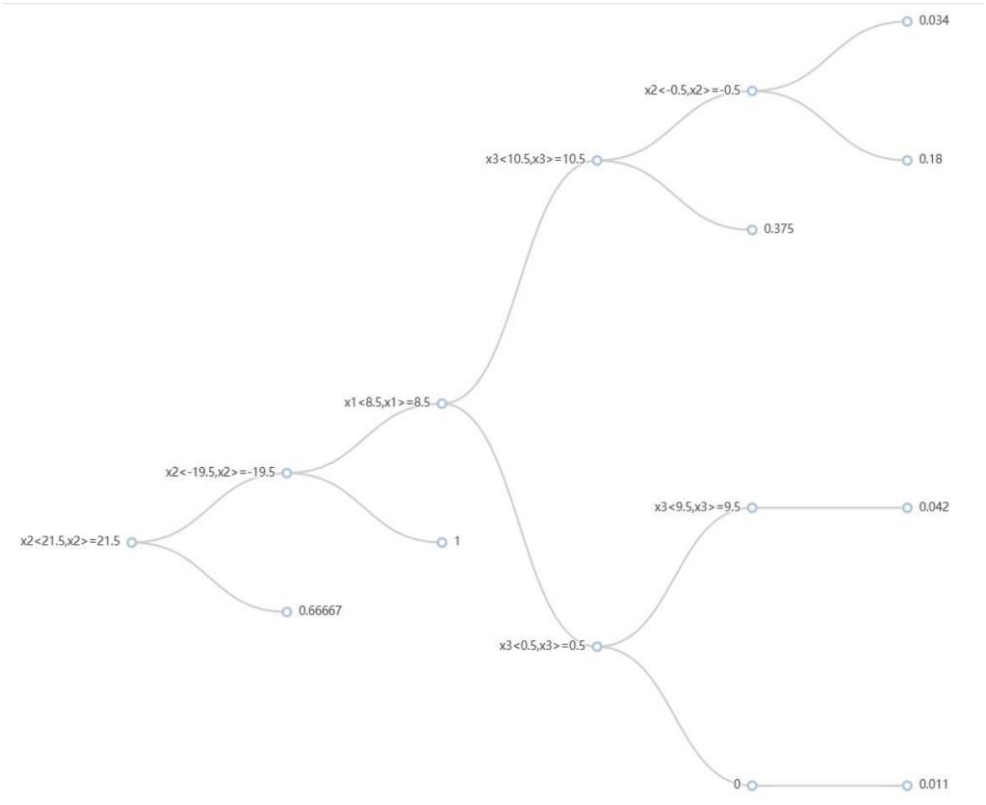

Figure 2. Decision tree model

The dataset is divided into a training set and a validation set with a ratio of 8:2. Then, the predict function is called to predict whether each country will have a breakthrough in 2028, and the return value is the probability of each country winning the prize. The return results corresponding to the 10 trees were averaged and normalized to obtain the probability of breakthrough. The probability of a breakthrough was found with three countries, Cyprus, Zambia, and Guatemala, which were 71.8%, 71.8%, and 82.1%, respectively. The most deserving of these nine countries is GUA.

In conclusion, three countries are predicted to win their first gold medal at the 2028 Olympic Games in Los Angeles; Cyprus, Zambia and Guatemala.

5. Host Country Impact on Medal Table

This section requires consideration of the impact of events on results. The relationship between the events and the number of medals won by each country is explored based on the established model, so that the key events for each country can be identified and the reasons explained. The impact of the events chosen by the host country also needs to be modeled. In order to better explore the impact, the model was improved, and a solution was completed.

5.1. Improved Model

In order to optimize the model that had been created, the two most advantageous events for each country were added as new independent variables and the first independent variable was reset to the number of participants in the same year, repeating the previous steps of creating a multiple linear regression model and creating an optimized model. The two most dominant events for each country were determined by selecting the two events in which each country had accumulated the most medals since 1984. The same evaluation was performed on the improved model and comparisons were made.

It was found that the correlation coefficients increased significantly with the inclusion of the dependent variable of dominant events, as did the F-values, with a slight increase in the residual variance. It shows that favorable events do have a positive impact on countries winning more medals, suggesting that the number and type of events does affect how a country wins medals.

5.2. Relationship Between Events and Number of Medals

Using the United States as an example, we conducted a correlation analysis to obtain a matrix of correlation coefficients between different sports in the United States.

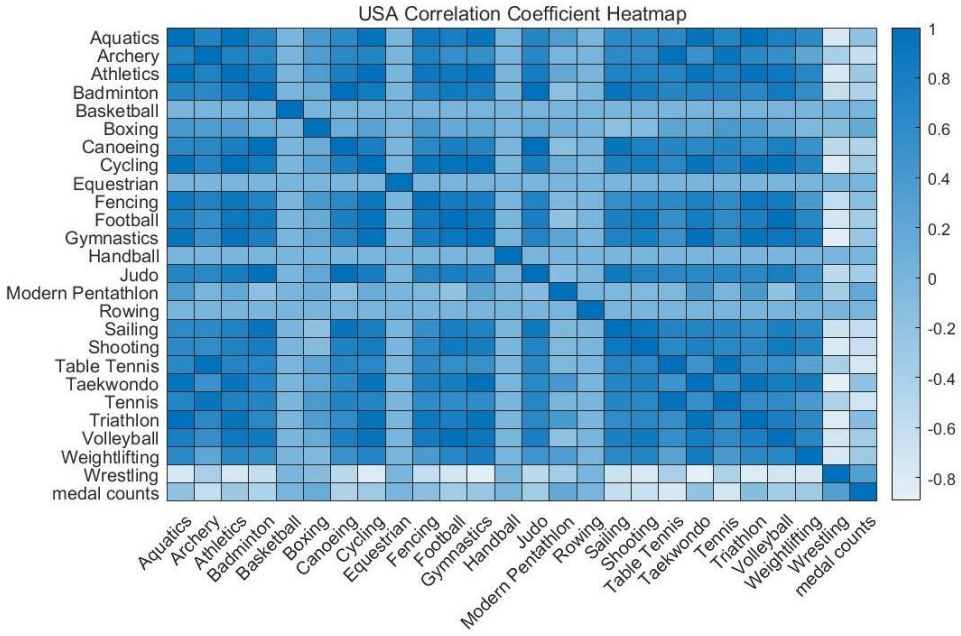

Figure 3. Matrix of correlation coefficients between different sports in the USA

This Fig. 3 is a heat map where each cell represents the correlation between two sports and the depth of the color indicates the magnitude and direction of the correlation coefficient. The color bars show the range of correlation coefficients from -1 (white) to 1 (dark blue). Dark blue indicates a high positive correlation and white indicates a high negative correlation. If an event is close to the dark blue color with several other events, the more likely it is to win a medal. Based on the heat map analysis, it can be seen that the United States has a very prominent track and field program with a high probability of winning a medal.

In addition to the United States, the study also looked at which events had the highest correlation with the number of wins of other highly ranked countries.

The paper highlights the major events in each country and finds that Australia, China and the United States are primarily good at water sports; France, the United Kingdom and the Netherlands are primarily good at cycling; Italy is good at fencing; and Japan is good at wrestling. There are also other sports that each country is better at, such as weightlifting in China, judo in Japan, and track and field in the United States.

Therefore, when the host country chooses to add events in which it excels, it will undoubtedly increase the likelihood that its own country will win and increase its medal count by a high probability, while countries with fewer dominant sports are more likely to regress. Thus, the host's choice of events can affect the number of prizes to some extent. This is also consistent with the assumption of a home country-influenced independent variable in the multiple linear regression model.

6. Conclusion

The hybrid modeling framework proposed in this paper shows significant advantages in time series forecasting and classification tasks. First, the multivariate linear regression model based on the five-dimensional feature space realizes the estimation of regression coefficients through the least squares method, and the F-test results and R^2 indicators verify the strong explanatory ability of the model to the target variables, and the residual variance indicates that the model has good fitting accuracy. Second, for the dichotomous prediction scenario, the integrated method of Bayesian regression and decision tree effectively improves the confidence of breakthrough probability prediction through probabilistic modeling and feature screening. Finally, through the domain feature extension (new dominant item dimension), the improved model has improved the correlation coefficients on average while maintaining the computational complexity, which verifies the key role of feature engineering on the model performance. Experiments show that the framework performs well in uncertainty quantification and feature interaction analysis, providing a reusable methodological reference for similar predictions.

References

[1] Zhang Xi-xiang, LI Taoshen. A heuristic constructive element-based multiple regression analysis method under data missing conditions [J]. Computer Applications, 2012, 32 (08): 2202 - 2204+2274.

[2] Fang Liting, Li Kunming. Bayesian estimation and application of semiparametric spatial lag quantile regression model [J]. Systems Engineering Theory and Practice, 2024, 44 (10): 3346 - 3361.

[3] Luo Qing, Ge Yuhao, Wu Fengbo. An outlier detection method for information zed control data based on information entropy and decision tree [J]. Microcomputer Applications, 2025, 41 (01): 209 - 211+216.

[4] Qi Long. Research on algorithm for solving linear regression equations by least squares [J]. Computer Products and Distribution, 2019, (09): 230.

[5] Zhang T.S. Method and implementation of Bayesian Meta-analysis for sparse binary categorical data[J]. Chinese Journal of Evidence-Based Pediatrics, 2020, 15 (04): 314 - 318.

[6] Yan Tong, Liu Yi. Research on feature screening methods for ultra-high dimensional data [J/OL]. Statistics and Decision Making, 2025, (05): 43 – 48 [2025-03-21]. https://doi.org/10. 13546/j. cnki. tjyjc. 2025.05.007.