

Prediction of Tennis Match Momentum Based on PCA and LGBM Models

Xiyihan Cao*, Yuchen Li and Yongxu Liu

Department of Mathematics, Jilin University, Changchun in Lilin Province, China

* Corresponding author: 2112108203@qq.com

Abstract. This study investigated the impact of momentum on the situation in sports based on data from the 2023 Wimbledon final. Using a BP neural network to interpolate missing values and screening the data by Principal Component Analysis (PCA), the best performing LGBM model was selected to achieve accurate quantification and real-time visualization of momentum using the classic match between Carlos Alcaraz and Novak Djokovic as an example. It is found that our model is highly accurate and performs well. In addition, we summarize the role of momentum in matches and provide professional strategy suggestions, which will provide valuable information for coaches and players, and we hope that this study will provide useful references for the development of sports.

Keywords: Pca, lgbm predictive model, tennis momentum, indicator selection.

1. Introduction

In sports, momentum is a key factor that can quietly influence the outcome of a game. In tennis, every point, mistake or tactical choice made by a player can be a turning point that changes the momentum of a match. However, the quantification of momentum is not intuitive, so the development of accurate and stable mathematical models and algorithms is essential to measure and predict game momentum. The study uses the 2023 Wimbledon men's singles final between 20-year-old Carlos Alcaras and 36-year-old Novak Djokovic as a case study. Our goal was to build a model to capture the flow of momentum during a race and evaluate athlete performance at key race points [1], while providing a visual description of the flow of the race. In order to achieve accurate prediction of game momentum, we adopted the LGBM model, which is an efficient gradient lifting decision tree algorithm with high efficiency and low memory usage. In comparison with other classical machine learning models, we found that LGBM models performed best in accuracy, recall, accuracy, F1 scores, and AUC. Ultimately, our study not only provides a new method for quantifying momentum in tennis matches, but also provides valuable tactical decision support for coaches and players. We hope that this study can promote the development of sports science and provide new ideas and expansion space for future research.

2. Data Preprocessing

2.1. Missing value interpolation based on BP neural network

We choose BP neural network to interpolate the missing values, which is used to interpolate the missing values due to its powerful nonlinear mapping ability, aiming to predict the missing data values using the existing valid data. The basic idea of using BP neural networks for missing value interpolation is to first train the BP neural network using samples containing complete data [2], and then predict the missing data values by the trained network using other variables of the missing data as inputs. This method makes full use of the potential nonlinear relationship between the data and is adaptable to various complexities of the missing data.

We construct a BP neural network model by taking the non-missing values in the dataset as the training set and the features corresponding to the missing values as the target values. Through continuous iterative training, the neural network can learn the complex relationship between the features and thus be able to predict the missing values more accurately.

The following comparison between predicted, true, and mean values is shown in Fig. 1 and Fig.2.

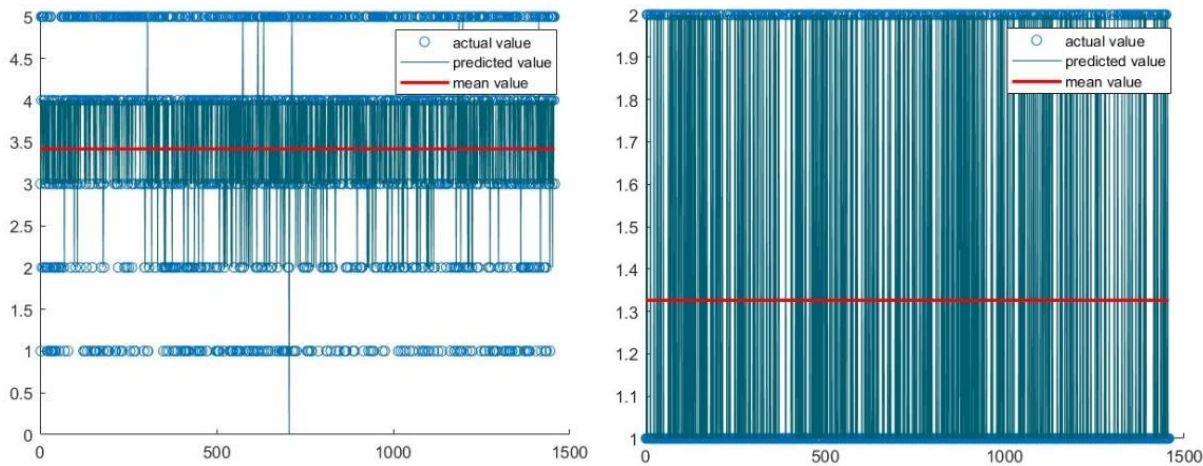


Figure 1. Comparison of serve direction interpolation and serve drop depth interpolation

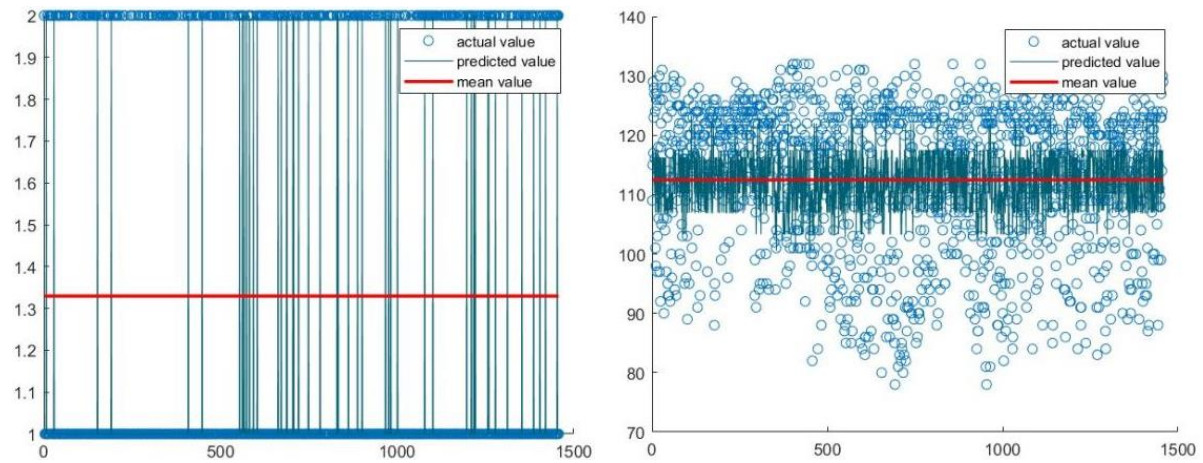


Figure 2. Comparison of return accuracy interpolation and stroke speed interpolation

2.2. Selection of Characterization Factors

We believe that in addition to being on the serving side or not, the reasons that are critically relevant for a player to obtain a victory are also related to the player's level of fatigue, individual technical ability, and real-time mentality of the match. We select 20 characteristic factors for analysis. Due to the limitation of the length of the paper, only part of the characteristic factors is shown in Table 1.

Table 1. Partial characteristic factor

x_1	The number of games won in the current set
x_2	The lead progress in the current game
x_3	first or second serve
x_4	current game leading or not
x_5	The lead progress in the current match
x_6	hitting an untouchable winning serve or not
x_7	hitting an untouchable winning shot or not
x_8	made two consecutive serving errors, losing the match point
x_9	made an unforced error

3. LGBM Model

3.1. Screening data based on Principal Component Analysis (PCA)

First of all, we carry out KMO and Bartlett's test to determine whether principal component analysis can be carried out [3].

Table 2. KMO test and Bartlett's test

KMO test and Bartlett's test		
KMO value		0.595
	approximate chi-square (math.)	44356.205
Bartlett's test of sphericity	df	120
	P	0.000***

Note: ***, **, * represent 1%, 5%, and 10% significance levels, respectively.

The results from Table 2 show that the value of KMO is 0.595, meanwhile, the results from Bartlett's spherical test show that the significance p-value is 0.000***, which presents significance at the level, rejecting the original hypothesis that there is a correlation between the variables and that the principal component analysis is valid. In Fig 3. and Fig. 4 gravel plot, at principal component 7, the characteristic root is lower than 1.0, and the contribution of variable explanation reaches 63.234, so the principal component is made to be 7 for principal component analysis.

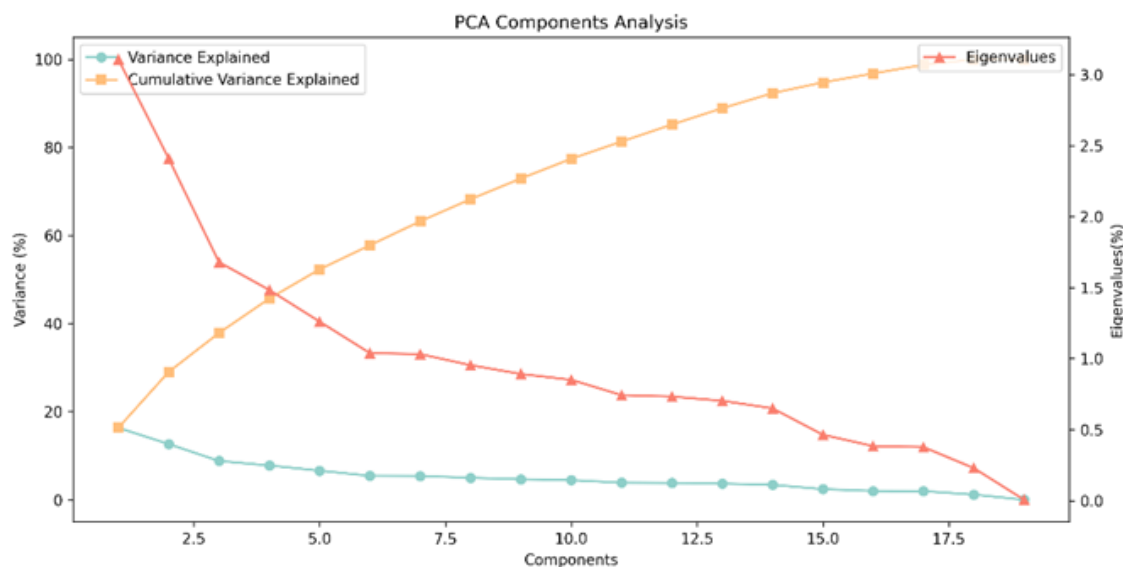


Figure 3. PCA computers analysis

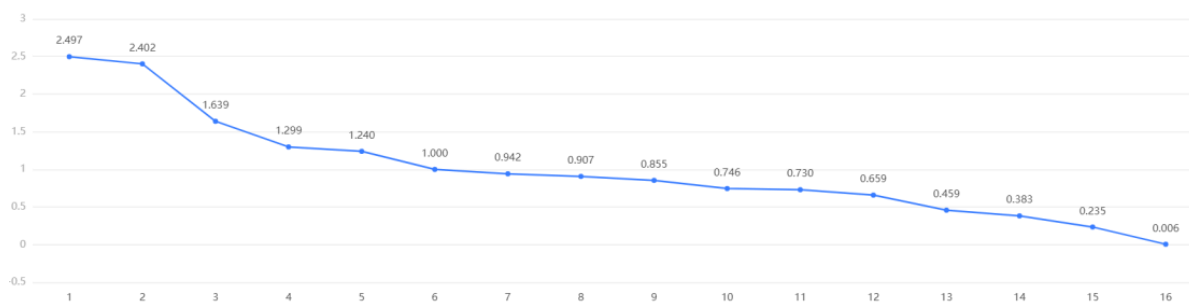


Figure 4. Diagram of crushed stone

3.2. LGBM model

LightGBM is a gradient boosting decision tree algorithm (GBDT) with gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) [4]. It uses the following techniques to improve efficiency and reduce memory usage.

Histogram (histogram) based decision tree algorithm: Traditional decision trees use a pre-sorting algorithm (e.g. in XGBoost) to find the optimal split point, which is very accurate but also very slow [5]. LightGBM discretizes continuous eigenvalues into bins (boxes) so that the number of segmentation points is reduced, greatly reducing computational complexity.

Gradient-based One-Side Sampling (GOSS): In order to maintain the accuracy of the data distribution, GOSS retains data samples with large gradients and down samples data samples with small gradients (i.e., data samples that have been predicted accurately), which reduces the amount of computation without significantly affecting the data distribution.

Exclusive Feature Bundling (EFB): In real-world data, many features are sparse (i.e., most of the values are 0.) EFB bundles these features and reduces the number of features, as well as reduces memory usage and speeds up computation.

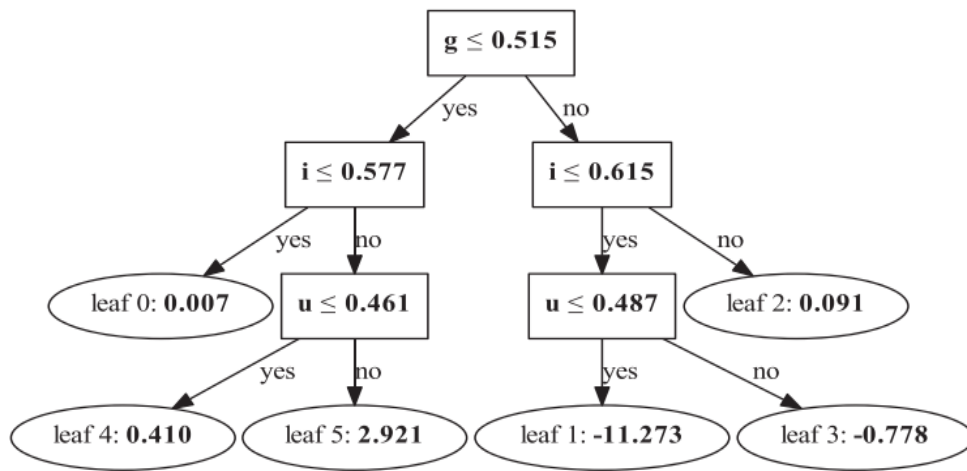


Figure 5. The regression decision tree of GBDT

As shown in Fig. 5, the regression decision tree of GBDT fitting T_{eff} as shown in Fig.5, the regression decision tree of GBDT fitting, with photometric data. The box represents feature selection according to the mean square error, and the ellipse represents the prediction result. The regression decision tree fits the residuals (the difference between the real value and the predicted value) of the (-1) th regression decision tree. With an increase in m , the fitting result of the m th regression decision tree tends to be 0. The final fitting result of GBDT is the sum of the fitting results of samples in each regression. The final fitting result of GBDT is the sum of the fitting results of samples in each regression decision tree.

4. Experiments

4.1. Evaluation

There are many indexes in machine learning to evaluate the classification effect of the classifier, here we mainly use Accuracy, Recall, Precision, F1 and AUC for evaluation.

Recall can be calculated using the following formula.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (1)$$

Precision can be calculated using the following formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

The F1 score is the reconciled average of Precision and Recall and can be calculated using the following formula:

$$F1 \text{ score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Accuracy can be calculated using the following formula:

$$Accuracy = \frac{True \text{ Positives} + False \text{ Negatives}}{True \text{ Positives} + True \text{ Negatives} + False \text{ Positives} + False \text{ Negatives}} \quad (4)$$

Whereas AUC is used to measure the performance of binary classifiers, it represents the area under the ROC curve.

4.2. Comparison with other Models

We choose LGBM, XGB, SVC, MLP these classical models for comparison, using 5-fold cross validation method for verification, the results are shown in Table 3.

Table 3. Comparison

	Accuracy	Recall	Precision	F1	AUC
LGBM	0.88	0.86	0.91	0.88	0.96
XGB	0.88	0.86	0.90	0.88	0.95
SVC	0.86	0.81	0.92	0.86	0.94
MLP	0.88	0.86	0.90	0.88	0.95
LR	0.83	0.79	0.87	0.83	0.92

5-fold cross-validation

LGBM XGB SVC MLP LR

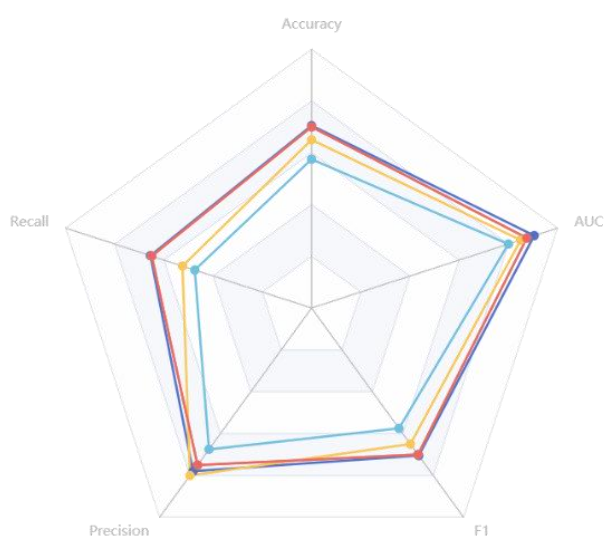


Figure 6. 5-fold cross-validation Radar Chart

As can be seen from Table 3 and the radar chart Fig.6, LGBM works best and improves relative to binary logistic regression.

In more detail, from Fig.7 ROC plot [6], it can be seen that the AUC of the LGBM model in the test set is 0.9588 and the AUC of the LR model is 0.9237. This indicates that the LGBM model has a better overall performance than the LR model on this particular task, and the TPR of both the LGBM and the LR rises with increasing FPR, but the curve of the LGBM is always located in the LR above, and the ROC curve of the LGBM model is closer to the upper left corner than that of the LR model, which means that it is able to reach high TPR and low FPR faster, indicating that the LGBM outperforms the LR at most threshold settings.

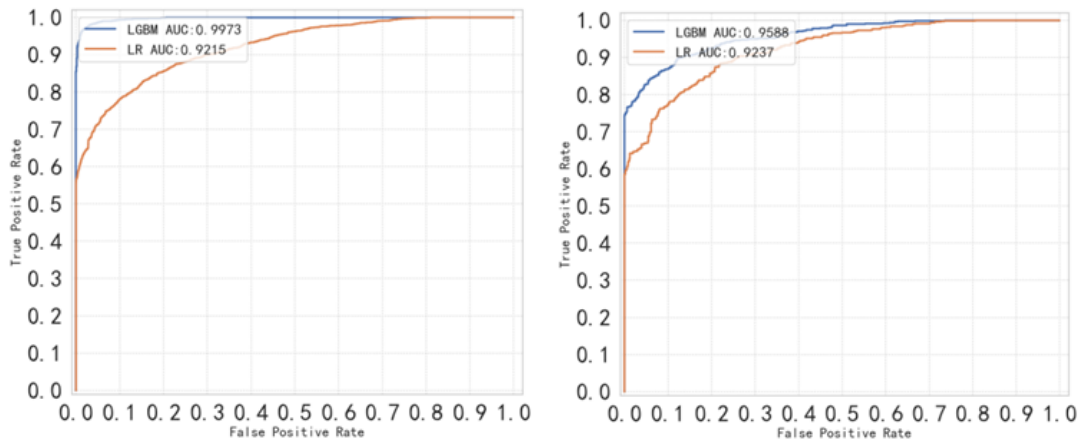


Figure 7. Training set ROC plot Test set ROC plot

4.3. Momentum Prediction

We re-do the modeling, using the best performing LGBM model, and select the classic 2023 Wimbledon men's singles final match in which 20-year-old Spanish rising star Carlos Alcaraz defeated 36-year-old Novak Djokovic for real-time performance visualization. As Fig. 8, the real-time performance scores of Carlos Alcaraz for a five-set match are represented by five different colored curves.

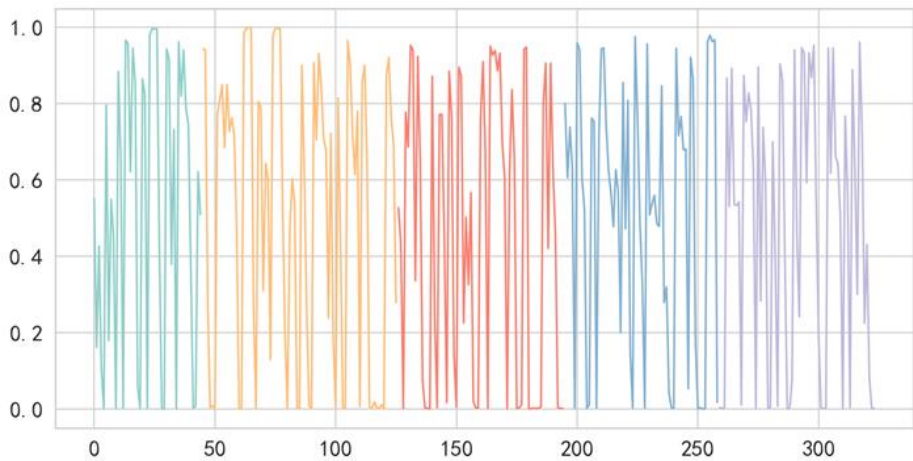


Figure 8. the real-time performance scores of Carlos Alcaraz for a five-set match

The actual scoring situation is compared with the Score, such as Fig. 9, which helps us to visualize the trend in the game.

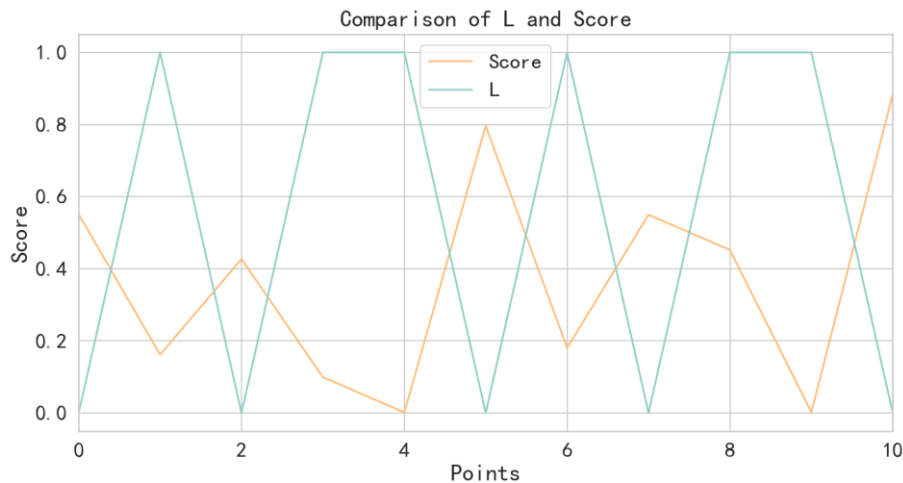


Figure 9. Comparison of Land score

Finally, the performance scores of Carlos Alcaraz for the given match points are given and it can be seen through Table 4 that the performance scores of Carlos Alcaraz in the key match point moments are 0.507649, 0.510365, 0.430196, 0.541306, and 0.479347. These scores can be used as an indicator to assess his performance in the key moments.

Table 4. The performance scores of Carlos Alcaraz in the key match point moments

S_1	0.507649
S_2	0.510365
S_3	0.430196
S_4	0.541306
S_5	0.479347

In summary, based on our rebuilt LGBM model, we have observed the momentum changes in the game through real-time performance visualization and comparative analysis, and given Carlos Alcaraz's performance scores at match point moments. These results provide valuable information to coaches and players, helping them to make more accurate tactical decisions.

5. Conclusion

This study provides an in-depth study of momentum in sports based on the 2023 Wimbledon final. We interpolated the missing values of the data through BP neural network technique, selected the data using PCA method, and quantified and visualized the momentum using LGBM model. The study shows that our model is highly accurate and performs well, providing valuable information and strategy suggestions to coaches and players, while also summarizing the role of momentum in the game.

The strengths of this study lie in the data-driven decision support, momentum recognition capabilities, and serve strategy optimization. However, we also recognize the data dependency of the model and the challenge of dynamic adjustment. To further enhance model performance, we recommend improving data quality and completeness and exploring real-time adjustment strategies to accommodate momentum changes during matches. In summary, this study will provide a useful reference and impetus for the development of sports, and provide new ideas and expansion space for future research.

References

- [1] Blackman E G, Tarduno J A. Mass, energy, and momentum capture from stellar winds by magnetized and unmagnetized planets: implications for atmospheric erosion and habitability [J]. Monthly Notices of the Royal Astronomical Society, 2018, 481 (4): 5146 - 5155.
- [2] Nickolas S, Shobha K. An unsupervised neural network approach for imputation of missing values in univariate time series data [J]. Concurrency and Computation: practice and experience, 2021, 33 (9).
- [3] Zhang W, Wei Y. Regression based principal component analysis for sparse functional data with applications to screening growth paths [J]. 2015.
- [4] Aziz, Musheer R, Baluch, et al. LGBM: a machine learning approach for Ethereum fraud detection [J]., 2022, 14 (7): 1 - 11.
- [5] ZHAO Zheng, YU Xiaojie, XIONG Yuzheng, et al. PM2.5 prediction model based on regression analysis and decision tree algorithm [J]. Yangtze River Information and Communication, 2022, 35 (11): 9 - 11.
- [6] Y. J V, W. E S, Hajime U, et al. Erratum to "ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models" [J] Clin Epidemiol. 126C (2020): 207-16] [J]. Epidemiology, 2020, (republished).