Journal of Intelligent Learning Systems and Applications, 2024, 16, 242-261
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Stock Type Prediction Based on Multiple Machine Learning Methods

Zhonger Zhu, Wansheng Wang

College of Mathematics and Physics, Shanghai Normal University, Shanghai, China
Email: zhuzhonger@outlook.com, wswang@shnu.edu.cn

How to cite this paper: Zhu, Z.E. and Wang, W.S. (2024) Stock Type Prediction Based on Multiple Machine Learning Methods. *Journal of Intelligent Learning Systems and Applications*, **16**, 242-261.
https://doi.org/10.4236/jilsa.2024.163013

Received: June 6, 2024
Accepted: August 25, 2024
Published: August 28, 2024

Abstract

Stocks in the Chinese stock market can be divided into ST stocks and normal stocks, so to prevent investors from buying potential ST stocks, this paper first performs SMOTEENN oversampling data preprocessing for the ST stock category, and selects 139 financial indicators and technical factor as predictive features. Then, it combines the Boruta algorithm and Copula entropy method for feature selection, effectively improving the machine learning model's performance in ST stock classification, with the AUC values of the two models reaching 98% on the test set. In the model selection and optimization, this paper uses six major models, including logistic regression, XGBoost, AdaBoost, LightGBM, Catboost, and MLP, for modeling and optimizes them using the Optuna framework. Ultimately, XGBoost model is selected as the best model because its AUC value exceeds 95% and its running time is less. Finally, the XGBoost model is explained using the SHAP theory and the interaction between features is discovered, further improving the model's accuracy and AUC value by about 0.6%, verifying the effectiveness of the model.

Keywords

Stock Classification, Boruta Algorithm, Copula, Machine Learning, Interaction

1. Introduction

As a popular research in the financial market, stock classification mainly focuses on stock types (such as ST shares and non-ST shares), the classification of listed companies and the prediction of stock price rise and fall. Liu and Liao used support vector machines, decision trees, and logistic regression to explore the relationship between financial indicators and stock investment value [1]. Wu *et al.* optimized feature selection and clustering effect by entropy weight method

and improved FCM clustering [2]. Borovkova and Tsiamal applied the LSTM algorithm to predict intraday stock price movements [3]. Anbalagan and Maheswar proposed the method of decision making in Indian stock market based on fuzzy measure method [4]. Jones and Hensher used a mixed Logistic model to construct financial crisis warnings [5].

In terms of feature selection and model interpretation, researchers have adopted a variety of methods. K.K. Perabodha *et al.* explore the influence of features on model output using Boruta algorithm and Shapley Additive Interpretation (SHAP) [6]. Li Xiaoning *et al.* selected features based on multi-layer genetic algorithm (GA) and eliminated redundant features [7]. Amini N and Mahdavi M *et al.* selected the optimal combination of features through multiple feature reduction methods and predicted COVID-19 patient mortality [8]. Khalid Y Aram *et al.* proposed a knapsack maximum interval feature selection method based on SVM [9]. Krivorotko and Marias Osnovskaia *et al.* used the Covasim model and Optuna optimizer to assess epidemic spread, While SrinivasPolipireddy *et al.* applied XGBoost model and Optuna framework for hyperparameter tuning in cardiovascular disease diagnosis [10].

To sum up, many scholars have studied stock types and achieved rich research results, but most of them only use machine learning models with a single model for prediction. In addition, the decision-making process of machine learning models is not explained, and there is a lack of systematic methodology to guide feature selection and optimization process in feature selection. Therefore, this paper optimizes on this basis, adopts a variety of machine learning models combined with feature selection methods to select the model, and applies SHAP theory to explain the model.

2. Model Description

In machine learning, commonly used classification models include logistic regression, XGBoost, LightGBM, ADABoost, Catboost, and multilayer perceptrons (MLP). In addition, feature selection theories such as Boruta algorithm and Copula entropy are included.

2.1. Logistic Regression (Abbreviated: LR)

Logistic regression is a simple way to learn classification. Although this is a method called regression, it is a common binary classification pattern. Generally used objective functions include: square loss function, absolute loss function, cross entropy and so on. On this basis, a method of parameter estimation based on the maximum likelihood estimation method is proposed. Classification is performed by the probability of classifying each type of data. Because the dependent variable in logistic regression is 0 - 1, it is different from the case that the dependent variable in general regression is continuous. Therefore, logical transformation must be carried out based on linear regression. This operation needs to be changed by adding sigmoid activation function, so that the depend-

ent variable becomes continuous variable, so as to meet the modeling needs. The Sigmoid function is expressed as:

$$g(z)=\frac{1}{1+e^{-z}} \tag{1}$$

The assumed functional form of logistic regression is:

$$P(y=1|x;\theta)=h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T x}} \tag{2}$$

2.2. XGBoost (Abbreviated: XGB)

The core idea of XGBoost model can be summarized as follows: iterative operation. It mainly uses CART regression tree to realize the transformation of classifier from weak to strong, and improves the decision tree model according to the gradient boosting algorithm, so as to improve the classification accuracy. The objective function of the algorithm is as follows:

$$Obj^{(t)}=\sum_{i=1}^{n} l\left(y_i \hat{y}^{(t-1)}+f_i(x_i)\right)+\Omega(f_t)+cons \tag{3}$$

where, l is the loss function defined in general, $\Omega(f_t)$ is the regular term, and finally the constant term is added. The regularization term defines the complexity of the model and helps prevent overfitting.

2.3. LightGBM (Abbreviated: LGB)

LightGBM uses a technique called "histograph-based Learning" that processes data efficiently, reduces memory consumption and speeds up training. This is done by splitting the data into histograms, thus reducing the computational complexity. During the growth of the decision tree, LightGBM uses a technique called "GOSS" (Gradient-based One-Side Sampling) to more effectively select the leaf nodes that need to be split, thus improving the accuracy of the model. It supports parallel training and multi-classification tasks and can handle large-scale data sets. LightGBM first divides the data set into multiple straight squares (bins) and builds a histogram for each square to accelerate feature selection. In each iteration, LightGBM selects a straight grid to split, calculates the gain after the split, and then selects the split point with the maximum gain. After selecting the best splitting point, LightGBM grows a decision tree until the maximum depth is reached or the stopping condition is met. In order to prevent overfitting, LightGBM introduces regularization terms, such as the minimum data number of leaf nodes and the minimum gain of leaf nodes. The final prediction is made by combining the prediction results of multiple decision trees. In classification problems, the voting method is usually used to select the final category.

2.4. ADABoost (Abbreviated: ADA)

The basic idea of ADABoost algorithm is to combine multiple weak classifiers to form a powerful classification system. These weak classifiers usually choose a single level of decision tree. ADABoost uses an iterative approach in which only

one weak classifier is trained in each iteration before the next round of application. In N iterations, there are N weak classifiers, N − 1 are trained and their parameters do not change. In this case, the relationship between weak classifiers is as follows: the probability that N weak classifiers miss the previous n − 1 weak classifiers is large, and the final classification result depends on the overall effect of these classifiers. The ADABoost algorithm uses two weights, namely the data weight and the weak classifier weight. On this basis, the weak classifier determines the classification error of each decision point by weighting, and uses these minimum errors to determine the weight of each decision point. The larger the weight of the classifier is, the larger the proportion of the weak classifier in the final decision is.

2.5. CatBoost (Abbreviated: Cat)

Catboost is a class of Boosting algorithms, whose accuracy is better than XGBoost and LightGBM. Catboost is a learning tool based on fewer parameters and support for classification type and accuracy. Its main role is to effectively and reasonably deal with the characteristics of classification type, solve the problems of gradient offset and prediction offset, reduce the occurrence of overfitting, and further enhance the accuracy and generalization of the algorithm.

2.6. MLP (Abbreviated: MLP)

MLP is a feedforward artificial neural network used for supervised learning. It propagates information through multiple layers, each containing multiple neurons. These neurons communicate with each other through connections and receive input from the previous layer. Each neuron computes an output based on an input and activation function and passes the result to the next layer.MLP consists of an input layer, a hidden layer, and an output layer. The input layer receives the data, the hidden layer processes and transforms the data, and the output layer produces the prediction or classification results. At training time, the MLP uses a backpropagation algorithm and gradient descent to update the weights and biases of the network to reduce the difference between the prediction and the actual output. This optimization process is repeated until preset conditions are met.

2.7. Boruta Algorithm

The Boruta algorithm aims to weed out seemingly important but actually useless features in the tree model. It is based on two core concepts: the shaded feature and the binomial distribution. Shadow features simulate irrelevant features by randomly shuffling the original feature values, and are added to the original data set for training. If the original feature is less important than its shadow feature, it is considered invalid and eliminated. The binomial distribution iteratively selects features, setting thresholds to classify features into three categories: useless (red reject region), uncertain (purple uncertain region), and useful (green accept region). Through these

steps, Boruta algorithm can identify and eliminate invalid features.

2.8. Copula Entropy

Copula entropy is a new concept of entropy defined by MA [11] and SUN in this paper published in 2008. It can be used to measure the full-order correlation between random variables. First, let and are the marginal distribution functions of random variables and respectively, and are the joint distribution functions of the two.

Let $u = F_X(x)$ and $v = F_Y(y)$, after that, it is easy to get from the equation:

$$f_{X,Y}(x,y) = C\left(F_X(x), F_Y(y)\right) = C(u,v) \tag{4}$$

is called the joint density function. Therefore, the interaction entropy of variables, is the negative value of Copula entropy:

$$H_C(U,V) = -\iint C(u,v)\log C(u,v)\,dudv = -M(X,Y) \tag{5}$$

3. Data Preprocessing

3.1. Introduction to Data

The research data in this paper comes from tushare library. In this paper, stocks traded on the Shanghai Stock Exchange and Shenzhen Stock Exchange in China's stock market since January 1, 2016 are selected through the tushare library command to screen out the stocks that have become intermediate ST stocks since January 1, 2016.

139 characteristics were determined by two directions of financial index and technical factor. If the stock is ST stock, the data selection period is from January 1, 2016 to the time when the stock starts ST stock. If the stock is a non-ST stock, the data selection period is from January 1, 2016 to December 1, 2022. The result is 2706 stock data, each with a data set of 140 variables. There are 139 feature variables and 1 column of class labels in this data set. Some variable names and their corresponding meanings are described in **Table 1**. Besides, the data scale is shown in **Table 2**. After SMOTEENN, the sample size reaches 4317.

3.2. Data Cleaning

Since the order of magnitude and dimension between each feature are not consistent, it is not reasonable to directly apply the original data for modeling operation. Therefore, in order to ensure the consistency of data features in the order of magnitude and dimension, make the data comparable, and make the subsequent classification and prediction analysis reasonable, 139 columns of features are normalized in this work. The data were normalized directly using the Python preprocessing package sklearn and then combined into normalized data with mean 0 and variance 1.

3.3. Feature Selection

First of all, the Boruta algorithm based on XGBoost is used to test the feature

correlation. Since the relationship between 139 features selected and whether the stock is ST stock has not been tested, the Boruta algorithm is applied in this paper for verification. Through model establishment, the variables that finally fall in the green area are as follows Table 3.

After the selection of variables, this paper considers the correlation between variables. Due to the high degree of similarity of information contained in variables, the accuracy of model classification will be affected.

Table 1. Partial variable names.

Attribute	Name	Label
Features	eps	Basic earnings per share
	current_ratio	Current ratio
	quick_ratio	quick_ratio
	cash_ratio	cash_ratio
	invturn_days	Days of inventory turnover
	arturn_days	Days of turnover of accounts receivable
	inv_turn	Inventory turnover
	ar_turn	Accounts receivable turnover rate
Target	ST	If it is ST stock, it is 1, otherwise it is 0

Table 2. Data description.

Dimensionality	Samples	Resampling	Train Set	Test Set	Validation
139 times*1	2706	4317	70%	30%	5 fold cross

Table 3. 11 characteristic variations.

Variables	Label
q_roe	Return on equity (single quarter)
profit_to_gr	Net profit/gross operating income
ebit	EBIT
undist_profit_ps	Undistributed profit per share
bps	Net asset value per share
fixed_assets	Total fixed assets
netprofit_yoy	Growth rate of net profit attributable to shareholders of parent company (%)
macd_dif	The difference between the short-term and long-term average price of a share price
rsi_12	12-day relative strength indicator
pct_change	Price limit
equity_yoy	Net asset growth rate

Therefore, the correlation coefficient index is used in this paper to measure the information coincidence degree or redundancy degree among independent variables. For variables with large absolute correlation coefficients, some features can be retained, while others can be deleted to reduce the feature dimension and serve as a reference for variable screening. In this paper, the correlation between variables is described from linear and nonlinear perspectives, as shown in the following Figure 1.

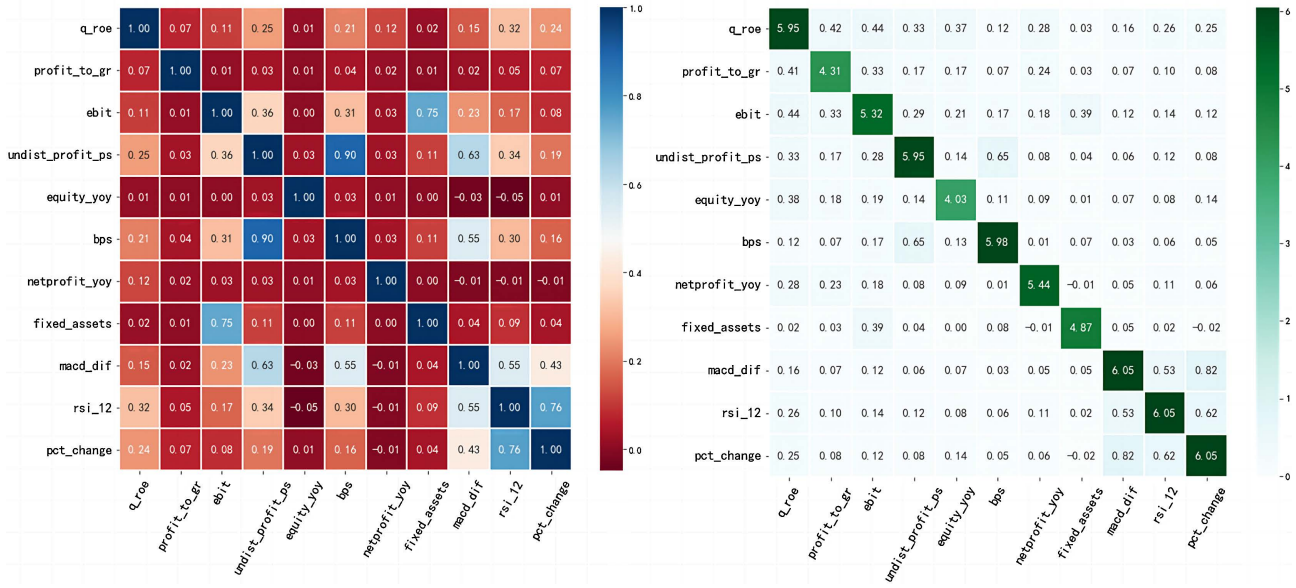

Figure 1. Thermodynamic group diagram of variable correlation coefficients.

It can be seen from the upper left figure that the technical factor has a high linear correlation among the Pearson correlation coefficient of 11 independent variables, and the linear correlation between the total fixed assets and EBIT is also high. However, it can be seen from the upper right figure that the Copula entropy of the two is 0.39, and the nonlinear correlation is low. Besides, it can be seen that the Copula entropy has a high correlation with itself. But there is less mutual information between the remaining variables. However, the correlation between the three factors of stock technical performance is higher than 0.5, and the mutual information between undistributed profit per share and net asset per share is also larger, reaching 0.65.

Therefore, this paper makes selection based on the economic meaning of the feature itself, and considers the principle of minimum mutual information between this variable and other variables. Therefore, the seven variables identified in this paper are as follows: Quarterly return on equity (q_roe), netprofit/total operating income (profit_to_gr), earnings before interest and tax (ebit), net assets per share (bps), total fixed assets (fixed_assets), netprofit attributable to shareholders of the parent company (netprofit_yoy), rise and fall (pct_change) and the year-on-year growth rate of net assets (equity_yoy). The English names of the variables will be used for subsequent analysis.

3.4. Feature Modeling

According to the above, the feature selection method chosen in this paper is Boruta algorithm combined with Copula entropy. The former is used to select the relationship between feature variables and target variables, and the latter is used to remove the non-linear information redundancy between feature variables. Therefore, this paper compares the six machine learning models mentioned above, compares the Boruta algorithm with the boruta algorithm combined with the Coplua entropy method, takes the AUC value as the measurement index, and obtains the following Table 4.

As can be seen from Table 4, among the six algorithms, Bo-Copula feature selection method on XGBoost, LightGBM and Catboost models is superior to Boruta method alone, but three models are superior to Boruta algorithm. The AUC value of two models in the Boruta-Copula feature selection reaches about 98%, so this paper believes that the Boruta-Copula feature selection method has certain advantages.

Table 4. Comparison of feature selection models.

	LR	XGB	LGB	ADA	Cat	MLP
Bo-Copula	0.90511	0.98147	0.98213	0.94771	0.97512	0.94979
Boruta	0.91747	0.96037	0.97319	0.95333	0.97378	0.95377

4. Empirical Analysis

After feature selection is completed, this paper conducts initial modeling.The data set is segmented to verify the prediction accuracy of each classification model, and the data set is segmented at a ratio of 7:3. Six models, including logistic regression, XGBoost, LightGBM, ADABoost, Catboost and MLP, were selected to train the classification prediction model.

4.1. Modeling without Tuning

First, the built-in model classifier of Python software is used for training.No parameters were set in this training, and the default parameters of the classifier were selected for training. The training results were shown in Table 5.

At the same time, the AUC, six model classification confusion matrices and other model evaluation indicators were plotted into curves to obtain Figure 2 and Figure 3. Recall adopts macro method, F1 Score is weighted.

As can be seen from Figure 2 and Table 5, the accuracy of the initial models in the test set is higher. In addition, in order to identify ST stocks as much as possible, this paper pays more attention to the recall rate. As can be seen from Figure 3, the six thermal maps are logistic regression model, XGBoost, LightGBM, ADAboost, CatBoost, and MLP. Then, the recall rate of the model before tuning is higher. In order to prevent overfitting, the model will be cross-validated and the model hyperparameters will be tuned to find the optimal hyperparameters so that the model can achieve better performance.

Table 5. Evaluation index values of the model before tuning.

	Accuracy	recall	F1 Score
LR	0.906636	0.907321	0.906630
XGB	0.977623	0.977547	0.977622
LGB	0.983796	0.983886	0.983798
ADA	0.949074	0.949444	0.949083
Cat	0.980710	0.980676	0.980709
MLP	0.951389	0.951497	0.951394

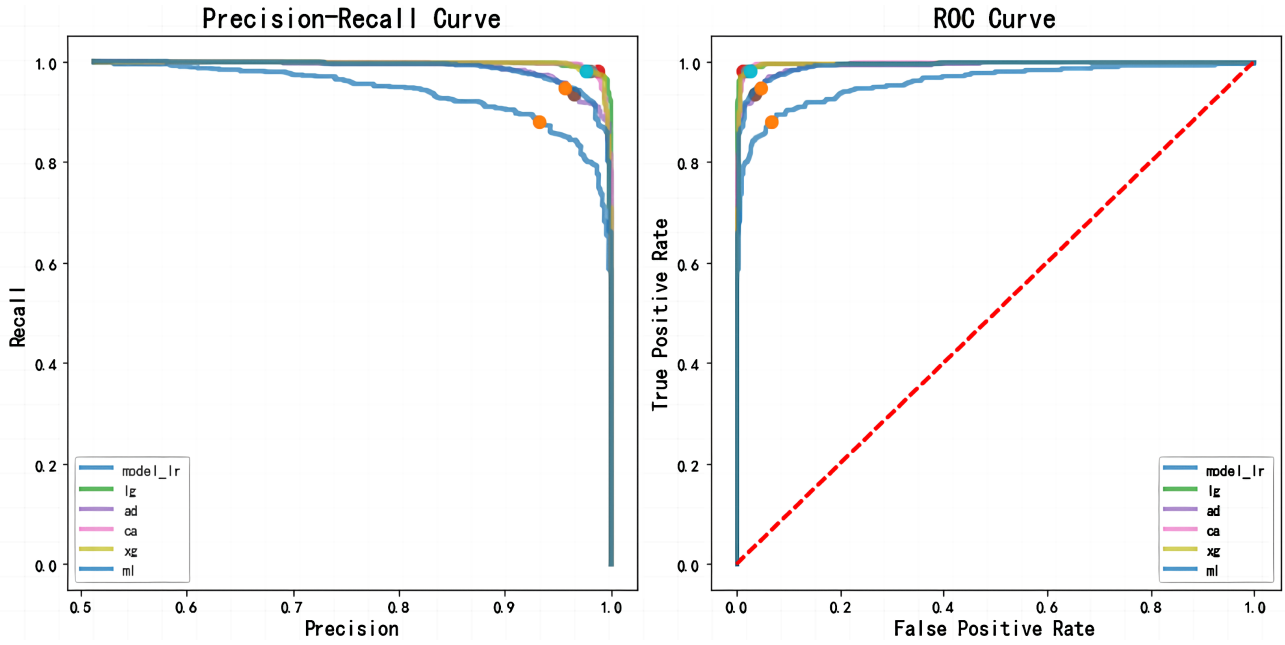

Figure 2. Comparison of ROC curves of prediction models before tuning.

4.2. Tuning

After the initial modeling is completed, Optuna framework is applied to check parameter tuning in this paper. Optuna is an automatic hyperparameter optimization framework, and it is very convenient to apply Optuna framework in machine learning. By setting parameters that need to be tuned, such as leaf nodes, maximum depth and other hyperparameters, and establishing an optimization framework, 20 times of path tuning were set through hierarchical cross-validation, and the maximum AUC was selected as the tuning standard in this paper. Finally, the optimal parameters were obtained and saved to make preparation for the subsequent training. Only some hyperparameters of some models are shown in this section. The training set was cross-verified by 5-fold to avoid overfitting, and the hyperparameters of each model were set for optimization training. Finally, the hyperparameter Settings of the six parameters after tuning are obtained, as shown in Tables 6-11. These hyperparameters provide a basis for the following model selection.

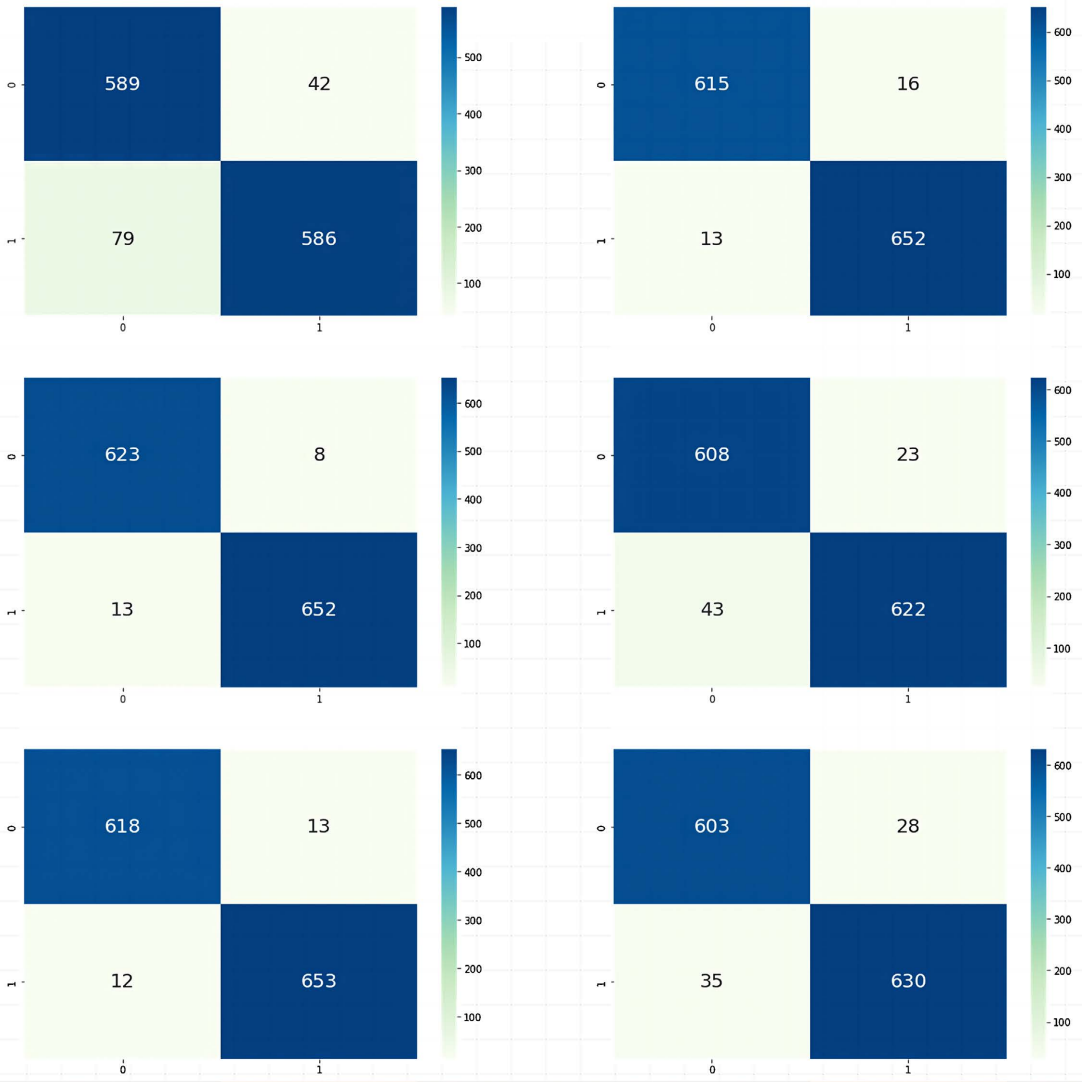

Figure 3. Thermodynamic composition of confusion matrices before tuning.

Table 6. Table of logistic regression hyperparameters.

Hyperparameter	Value
C	20
Penalty	L2

Table 7. Table of XGBoost hyperparameters.

Hyperparameter	Value
max_depth	3
learning_rate	0.1
n_estimators	63
min_child_weight	2
gamma	0.061992

Continued

alpha	0.001271
lambda	0.830266
colsample_bytree	0.577063
subsample	0.314197

Table 8. Table of LighrGBM hyperparameters.

Hyperparameter	Value
n_estimators	72
learning_rate	0.268540
num_leaves	43
max_depth	6
min_data_in_leaf	200
lambda_l1	5
lambda_l2	5
min_gain_to_split	8.559152
bagging_fraction	0.7
bagging_freq	1
feature_fraction	0.9

Table 9. Table of ADA hyperparameters.

Hyperparameter	Value
learning_rate	0.2178889
n_estimators	59

Table 10. Table of Cat hyperparameters.

Hyperparameter	Value
max_depth	3
learning_rate	0.08
n_estimators	92
max_bin	216
min_data_in_leaf	192
l2_leaf_reg	0.000129
subsample	0.518710

As shown in Table 10, the hyperparameter optimization of CatBoost model includes max_depth and learning_rate, etc. The optimization range of each hyperparameter is set through Optuna framework, and the optimization is carried out with the maximum AUC value, and the optimal value of each hyperparameter is finally obtained. At the same time, Optuna tuning framework can visualize

the process of searching hyperparameters, such as CatBoost model to get a Figure 4, in searching hyperparameters.

Table 11. Table of MLP hyperparameters.

Hyperparameter	Value
alpha	0.000596
learning_rate_init	0.0098860
beta_1	0.679091
beta_2	0.916951

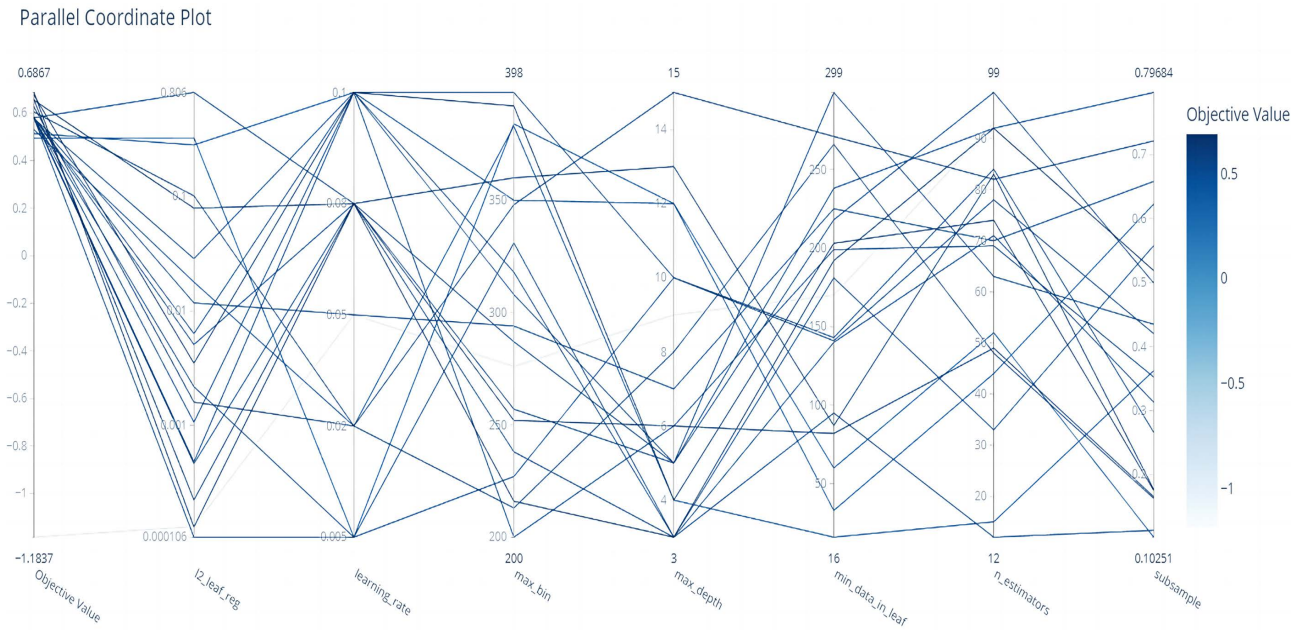

Figure 4. Latitude and longitude diagram of parameter coordinates of CatBoost model.

As shown in Figure 4, the hyperparameter tuning process of CatBoost model starts from the subsample of the training set. After the number of learners (n_estimators), the minimum number of samples of a leaf node (min_data_in_leaf), the maximum depth of the tree (max_depth), etc., the optimal value is reached. 20 different tuning paths can be intuitively obtained through the longitude and latitude coordinate diagram.

4.3. Modeling after Tuning

After model tuning, the final hyperparameters of the six models are obtained, so the final model selection is carried out. This paper carried out cross-validation to reduce the probability of overfitting, and obtained the values of each evaluation index as follows Table 12. The evaluation indicators do not specify methods.

Compared with the results before parameter tuning, from the perspective of accuracy, the accuracy of models except MLP decreased, indicating that the models before tuning had overfitting to a certain extent, and the accuracy of

XGBoost, ADABoost, Catboost and MLP models after tuning exceeded 95%. In terms of recall indicators, all six models decreased except MLP, with XGB, Catboost and MLP performing better. Taking F1 Score as a measure, it can be found that XGBoost, ADABoost, Catboost and MLP are four models with F1 Score values of more than 95% and perform well. Based on the tune-up run time and the test run time, the XGBoost model has less time. Based on the above five indicators, this paper chooses XGBoost model as the model of stock classification prediction.

Similarly, the confusion matrix group diagram is shown in Figure 5.

Table 12. Values of model evaluation indexes after optimization.

	Accuracy	Recall	F1 Score	Parameter Setting time	Run time
LR	0.907407	0.881381	0.907264	0.26581 s	0.048839 s
XGB	0.952160	0.957958	0.953662	4.99447 s	0.430995 s
LGB	0.937500	0.947447	0.939687	0.93704 s	0.087580 s
ADA	0.951389	0.944444	0.952309	6.215535 s	1.429417 s
Cat	0.952932	0.957958	0.954375	91.396693 s	0.898349 s
MLP	0.962191	0.977477	0.963731	134.874886 s	6.600946 s

It can be seen that in the optimized model, it is actually ST stocks, but XGBoost, Catboost and MLP perform better in the data predicted to be non-ST stocks. It can be seen from the confusion matrix of XGBoost model that there are only 28 samples of ST stocks predicted to become non-ST stocks by XGBoost model. Similarly, it can be seen from the confusion matrix of CatBoost model that there are only 28 samples of ST stocks predicted to become non-ST stocks by CatBoost model. In addition, according to the prediction of MLP model, the number of samples of non-ST stocks predicted to become ST stocks was only 34. Therefore, combined with the confusion matrix, it can be seen that the prediction efficiency of XGBoost model is also relatively good.

5. Model Visualization

Since the machine learning model is not very interpretable, similar to the black box model, this paper uses the help of SHAP library to explain how the XGB model performs ST stock classification. Since SHAP is a model-based post-interpretation method, its core idea is to calculate the marginal contribution of features to the model output, and explain the "black box model" from the overall and local levels. SHAP establishes an additive interpretation model in which all properties are treated as research targets for the model. The existing attribution methods can not express the interaction well, but need to divide the interaction among attributes. To get the interaction effect, we use the SHAP interaction value. The interaction value of SHAP can both explain the interaction of individual predictions and guarantee consistency. In SHAP theory, feature

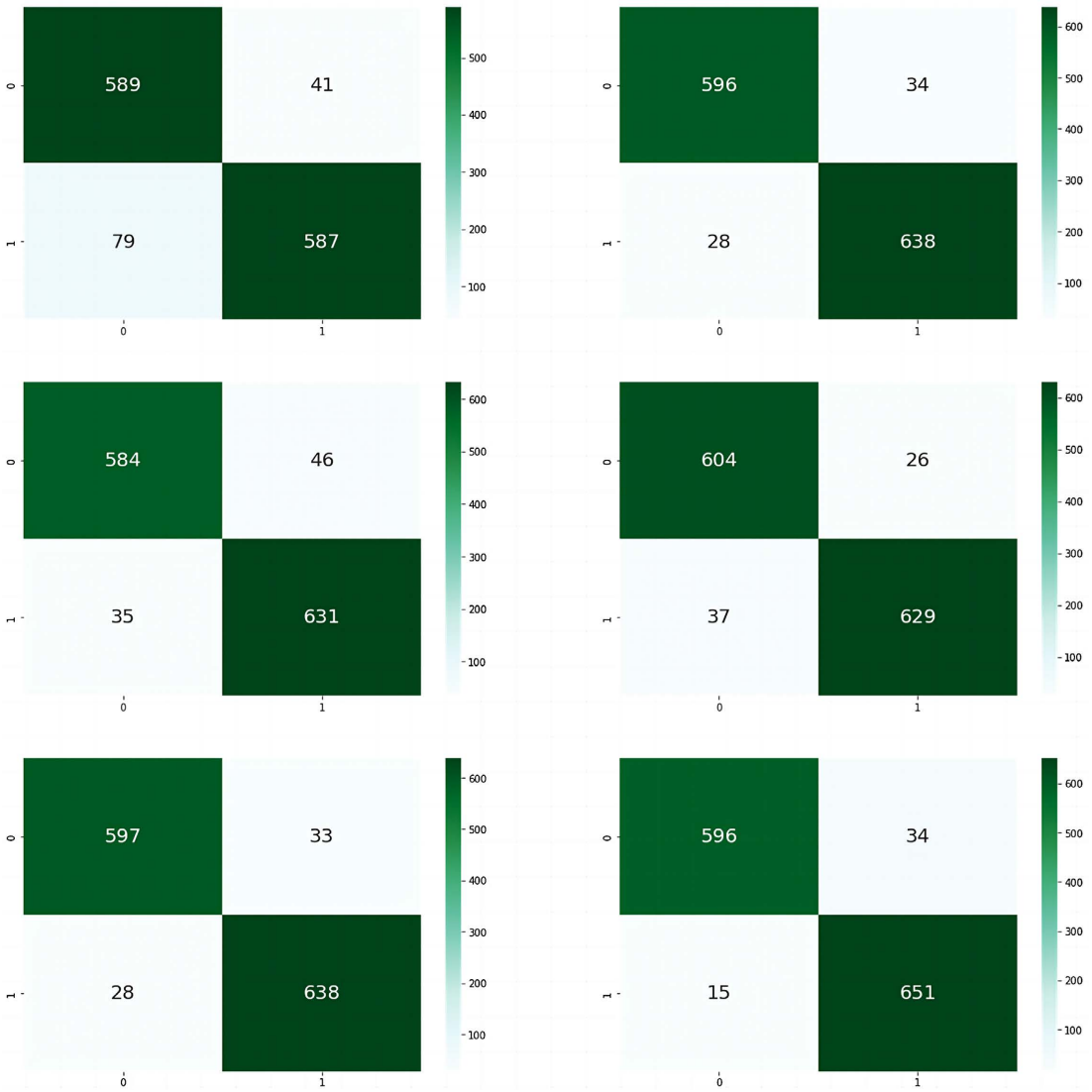

Figure 5. Thermodynamic composition of confusion matrices after tuning.

attribution usually affects the input property, assigning a property to a property, and extracting the influence from the main effect to get more information. On this basis, a distribution matrix based on feature and feature interaction is established, which reflects the prediction effect of these two features on XGBoost model.

5.1. Waterfall Plot

The waterfall statement can be used to obtain a waterfall plot for a sample value of the model.

For the sake of illustration, the sample selected in **Figure 6** is the 0^{th} sample. From **Figure 6**, the vertical axis represents the seven features of the model input, and E[f(X)] represents the expectation of all samples f(x), *i.e.* the average of the predicted values of Model. Predict (X). f(x) represents the 0^{th} sample, and f(x) has a value of 3.95, which represents the predicted value of the 0^{th} sample.

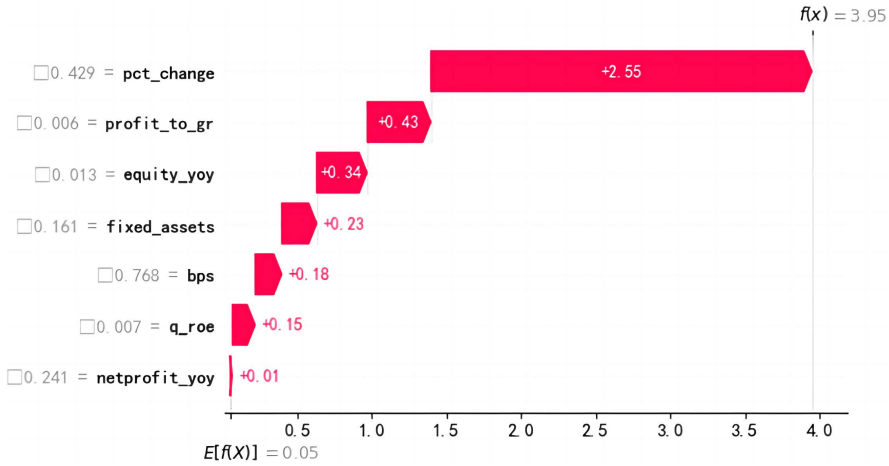

Figure 6. Waterfall diagram of single sample characteristics.

Therefore, it can be seen from the direction of the waterfall diagram in the figure that the seven features all play a positive gain effect on the 0th sample, and the gain contribution of the rise and fall is the largest, which is +2.55.

5.2. Feature Dependence Graph

This paper uses the scatter statement to plot feature dependencies.

It can be seen from Figure 7 below that the X-axis is the value range of q_roe under the overall sample, and the Y-axis corresponds to the shap value under the value of q_roe. It can also be seen from the figure that most samples are concentrated near 0 and samples less than 0 correspond to higher shap values. Because this feature has interaction effect with other features, when the feature value of q_roe is the same, there will be different shap values. Therefore, it is found in

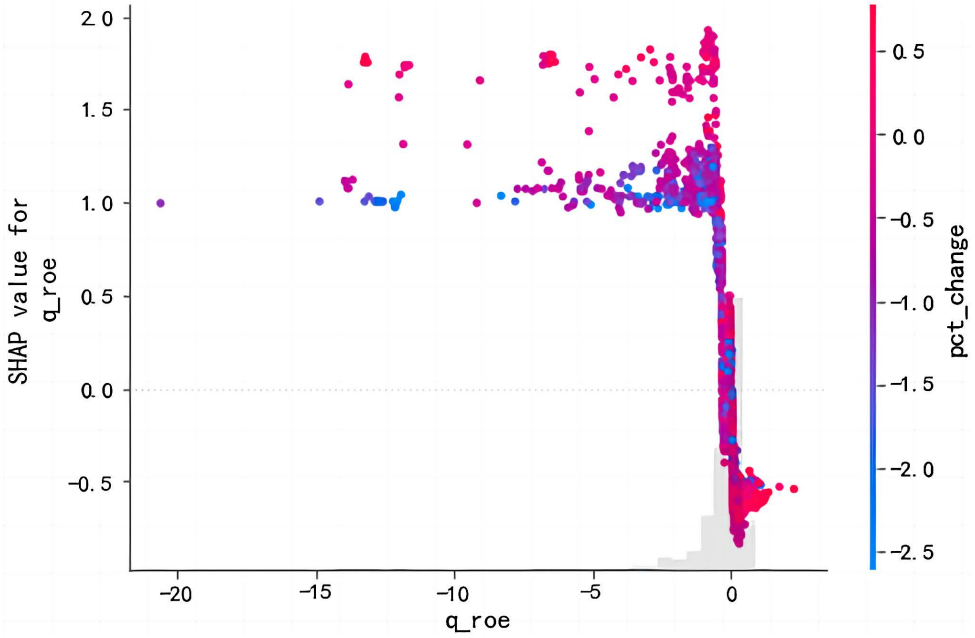

Figure 7. Feature dependency diagram.

this paper that the shap value of the part with low fluctuation is generally less than 0 for the part with high q_roe, which has a negative impact.

5.3. Characteristic Density Scatter

This paper uses the summary statement in the SHAP package to draw the scatter plot of feature density. As can be seen from Figure 8, x axis represents the shap value, y axis represents the feature ranking, and the seven features are sorted according to the average absolute value of shap. Therefore, this paper finds that variables such as pct_change and q_roe are of high importance to the model. Among them, the shap value of the point with high rise and fall is less than 0, showing a negative influence. And the distribution of the characteristic samples of the rise and fall is relatively dispersed. And the ranking is the highest, indicating that the rise and fall characteristics are highly relevant to the model. However, it can be seen from the net profit yoy distribution that most of the distributed points are concentrated near the shap value of 0, so it has no influence on most stocks and only affects a few stocks in this paper.

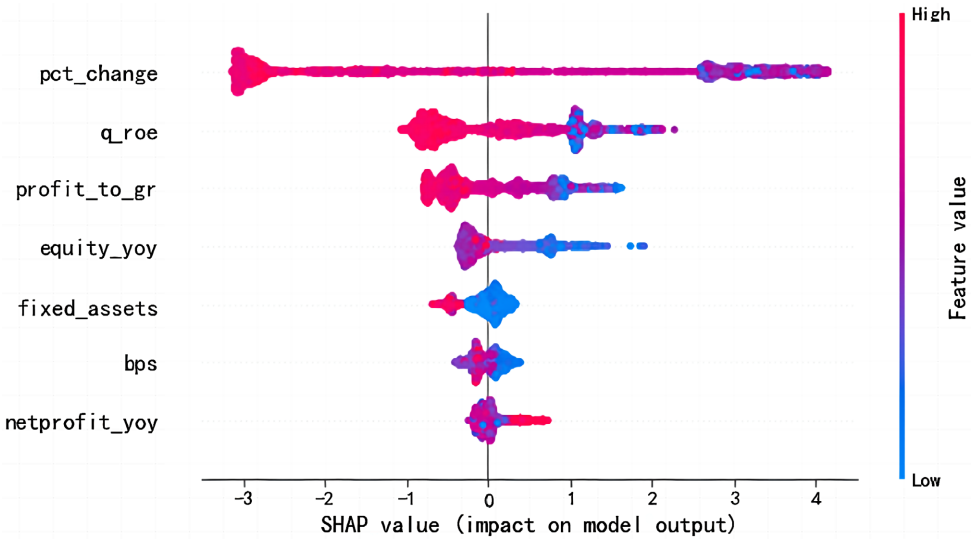

Figure 8. Scatter plot of characteristic.

5.4. Feature Interaction Analysis

This paper uses the interaction values statement for interaction analysis.As can be seen from Figure 9, there is an interaction between pct_change and q_roe, and between pct_change and profit_to_gr among the seven variables in this paper. In addition, when the relationship between pct_change and profit_to_gr is studied in this paper, it can be seen from Figure 10 that when pct_change value is concentrated at 0, its shap value is distributed scattered, and it is approximately distributed on a line with profit_to_gr value. In addition, the part with pct_change greater than 0 has a small impact on profit_to_gr, while the part with pct_change less than 0 has a large impact on profit_to_gr. Therefore, it can be seen that there is an interaction between the two features.

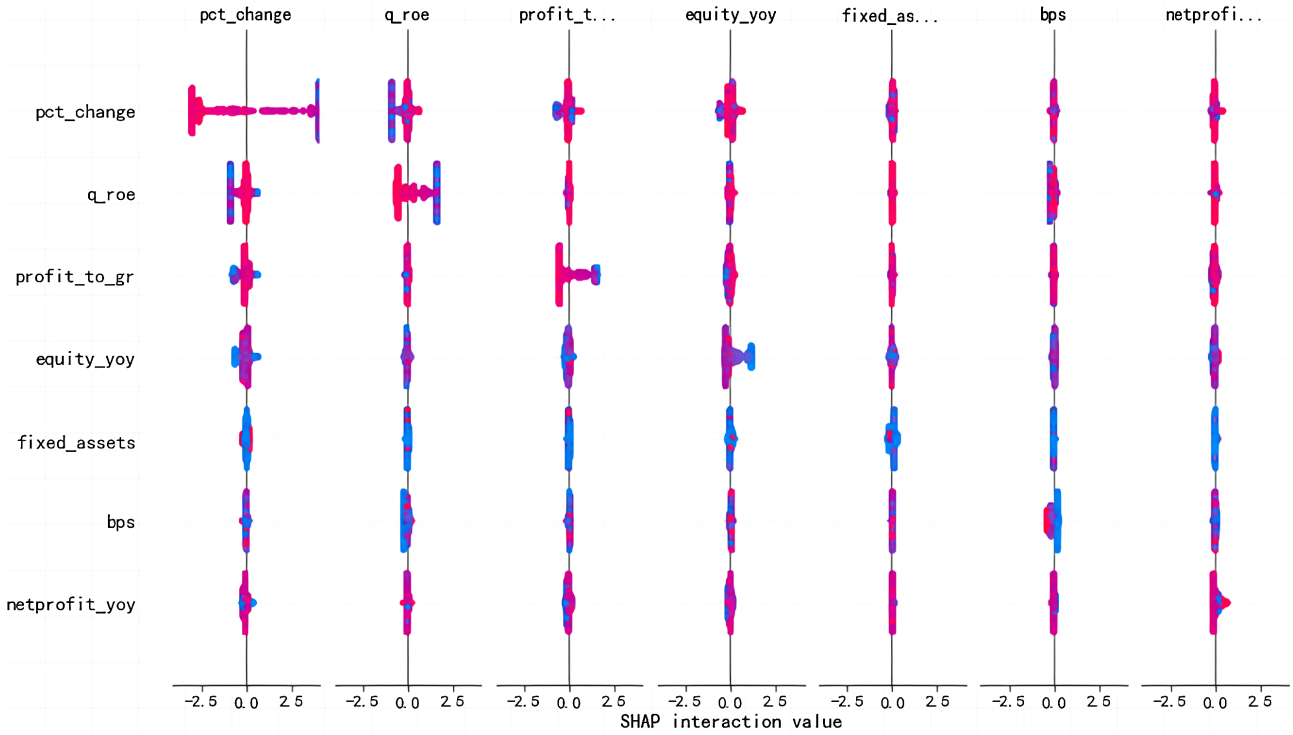

Figure 9. Feature interaction diagram.

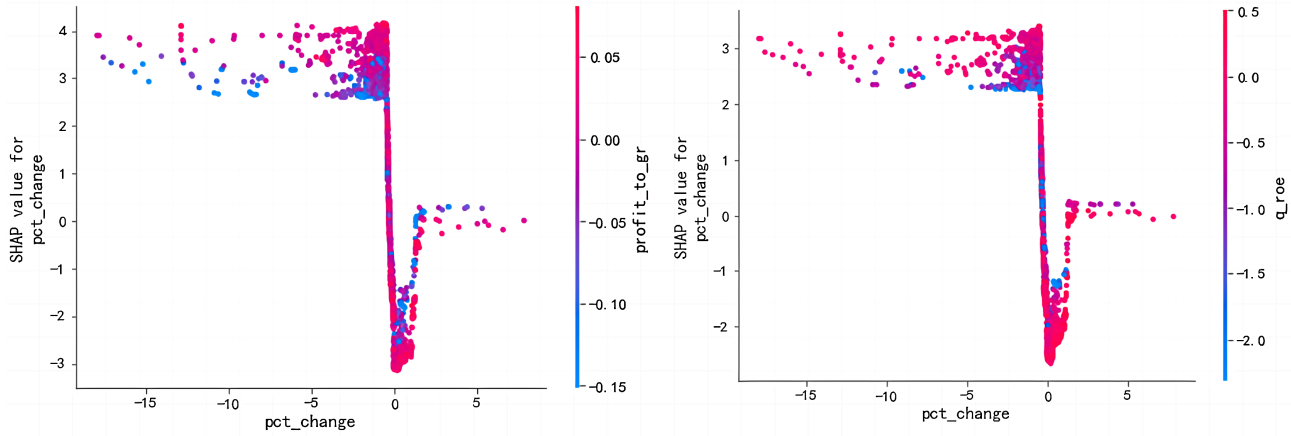

Figure 10. Interaction diagram of feature variables

Table 13. Comparison of interaction models.

	Feature num.	Accuracy	Recall	F1 Score	AUC
XGB	7	0.952160	0.957958	0.953662	0.951995
XGB	9	0.957562	0.959581	0.958863	0.957497

Similarly, this paper finds that pct_change and q_roe also interact. Therefore, in this paper, these two groups of variables with corresponding interactive characteristics are trained in the XGBoost model, and the results in Table 13 are obtained.

At the same time, this paper draws the confusion matrix diagram of model classification after the interaction effect is added.

It can be seen from Table 13 that the accuracy rate of XGBoost model increased by more than 0.005, the recall rate increased by about 0.002, the F1 value increased by about 0.005, and the AUC increased by about 0.006 through the features added by interaction. In addition, it can be seen from Figure 11 that the left is the confusion matrix generated after the interaction effect is added, and the right is the matrix generated without the interaction effect. It can be seen that after the interaction effect is added, the proportion of correct prediction increases, while the proportion of actual ST stocks but predicted non-ST stocks decreases. Therefore, the XGBoost model selected in this paper is an excellent classification model, and the features after adding interaction effects are also relatively comprehensive interpretation model.

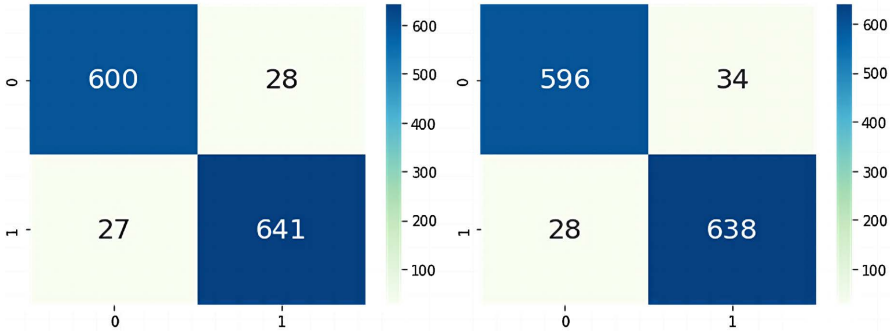

Figure 11. Comparison of interaction models.

6. Conclusions

ST stocks are a big problem for investors, and for most investors, how to avoid the investment of stocks into ST stocks is one of the goals of investment stocks. For a company, ST stock reflects the company's financial predicament. Therefore, the ST stock classifier set up in this paper based on the research of many scholars provides certain reference value for investors and companies.

This paper studies the classification of ST shares from the following three aspects:

1) Feature selection. In the actual stock market, investors will look at the financial data of the company and the technical factors of the stock in the past as the basis for investment in the stock. Therefore, this paper selects classical financial indicators studied by previous scholars, such as current asset-liability ratio, asset-liability ratio and other indicators, and adds many indicators reflecting the quality of the company's operation, combined with the performance of the stock market and technical factors, so as to reflect the performance of the stock in a more comprehensive way.

2) Feature redundancy. In this paper, 139 feature indicators are selected through feature selection as feature variables to study the classification of ST stocks, but it will cause feature redundancy, resulting in excessive resource con-

sumption and long time running during modeling. Different from the previous random forest to select features, this paper cannot select features through random forest due to the setting of more features, and because the algorithm of random forest to select features cannot fully explain whether the features are really related to the target variables. Therefore, this paper combines Boruta algorithm with Copula entropy method and uses Boruta-Copula method. On the one hand, it reasonably explains the problem that feature variables are accepted by target variables, on the other hand, it reduces the redundancy of linear and nonlinear information between feature variables, and finally selects 7 explanatory variables. Moreover, ST stock classification is verified by six machine learning models, and the performance of Boruta-Copula method is better than that of Boruta method to some extent.

3) Comparison and interpretation of models. In this paper, six classical machine learning models are studied, and the comparison of six models is obtained through the tuning of Optuna framework. In the actual stock market, the classical logistic regression model is easy to operate and explain, but the model can not fit the distribution under nonlinear conditions reasonably, so the prediction accuracy is poor. To solve this problem, this paper adopts five mainstream machine learning models in recent years, namely XGBoost, ADABoost, LightGBM, Catboost and MLP, to model and optimize the samples after resamping, and finally selects the samples by accuracy rate, recall rate, running time and other indicators. Choose the XGBoost model as reasonable and realistic as possible to conduct the modeling research of ST stock. At the same time, in order to explain the application of the model, this paper uses SHAP theory to visualize the model and avoid "black box algorithm" to a certain extent. Based on the interaction diagram of SHAP theory, this paper added interaction variables to the original model and further enhanced the model. Finally, the accuracy and recall rate of the model with interaction variables were higher than that of the original model, about 96%. It shows that the Boruta-Copula-XGBoost model combined with the interaction is more powerful to explain the classification prediction problem of ST shares.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Liu, X.J. and Liao, A.H. (2021). Application of SVM, Decision Tree and Logistic Regression Algorithm in Stock Classification and Prediction. *Proceedings of the* 2021 *International Conference on Financial Management and Economic Transition* (*FMET* 2021), 27-29 August 2021, 64-68. https://doi.org/10.2991/aebmr.k.210917.011

[2] Wu, Z.Y., Chen, G.D. and Yao, J.Y. (2019) The Stock Classification Based on Entropy Weight Method and Improved Fuzzy C-Means Algorithm. *Proceedings of the*

4*th International Conference on Big Data and Computing* (*ICBDC*' 19), New York, 10 May 2019, 130-134.

[3] Borovkova, S. and Tsiamas, I. (2019) An Ensemble of LSTM Neural Networks for High-Frequency Stock Market Classification. *Journal of Forecasting*, **38**, 600-619. https://doi.org/10.1002/for.2585

[4] Anbalagan, T. and Maheswari, S.U. (2015) Classification and Prediction of Stock Market Index Based on Fuzzy Metagraph. *Procedia Computer Science*, **47**, 214-221. https://doi.org/10.1016/j.procs.2015.03.200

[5] Jones, S. and Hensher, D.A. (2004) Predicting Firm Financial Distress: A Mixed Logit Model. *The Accounting Review*, **79**, 1011-1038. https://doi.org/10.2308/accr.2004.79.4.1011

[6] Kannangara, K.K.P.M., Zhou, W., Ding, Z. and Hong, Z. (2022) Investigation of Feature Contribution to Shield Tunneling-Induced Settlement Using Shapley Additive Explanations Method. *Journal of Rock Mechanics and Geotechnical Engineering*, **14**, 1052-1063. https://doi.org/10.1016/j.jrmge.2022.01.002

[7] Li, X., Yu, Q., Tang, C., Lu, Z. and Yang, Y. (2022) Application of Feature Selection Based on Multilayer GA in Stock Prediction. *Symmetry*, **14**, Article 1415. https://doi.org/10.3390/sym14071415

[8] Amini, N., Mahdavi, M., Choubdar, H., Abedini, A., Shalbaf, A. and Lashgari, R. (2022) Automated Prediction of COVID-19 Mortality Outcome Using Clinical and Laboratory Data Based on Hierarchical Feature Selection and Random Forest Classifier. *Computer Methods in Biomechanics and Biomedical Engineering*, **26**, 160-173. https://doi.org/10.1080/10255842.2022.2050906

[9] Aram, K.Y., Lam, S.S. and Khasawneh, M.T. (2022) Linear Cost-Sensitive Max-Margin Embedded Feature Selection for Svm. *Expert Systems with Applications*, **197**, Article 116683. https://doi.org/10.1016/j.eswa.2022.116683

[10] Krivorotko, O., Sosnovskaia, M., Vashchenko, I., Kerr, C. and Lesnic, D. (2022) Agent-Based Modeling of COVID-19 Outbreaks for New York State and UK: Parameter Identification Algorithm. *Infectious Disease Modelling*, **7**, 30-44. https://doi.org/10.1016/j.idm.2021.11.004

[11] Jian, M. (2019) Variable Selection with Copula Entropy. arXiv: 1910.12389. https://doi.org/10.48550/arXiv.1910.12389