Hindawi
Mobile Information Systems
Volume 2022, Article ID 6542652, 12 pages
https://doi.org/10.1155/2022/6542652

Research Article

Novel Key Indicators Selection Method of Financial Fraud Prediction Model Based on Machine Learning Hybrid Mode

Hongsheng Xu,[1,2] **Ganglong Fan**,[1,2] **and Yanping Song**[1,2]

[1]*College of Electronic Commerce, Luoyang Normal University, Luoyang, 471934 Henan, China*
[2]*Henan Key Laboratory for Big Data Processing & Analytics of Electronic Commerce, Luoyang Normal University, Luoyang, 471934 Henan, China*

Correspondence should be addressed to Hongsheng Xu; xhsls@lynu.edu.cn

Received 13 January 2022; Revised 18 February 2022; Accepted 11 March 2022; Published 28 March 2022

Academic Editor: Chia-Huei Wu

In recent years, the financial fraud event of listed companies has occurred continuously. The financial fraud has brought huge losses to the capital market and investors, hindering the investment allocation mechanism of the capital market. The current financial fraud prediction model can judge the company that may conduct financial fraud in advance. So, this can reduce economic losses. The key factor to construct the financial fraud prediction model is how to select the evaluation indicators. This paper analyzes the existing indicators selection method and finds the problem of low prediction accuracy. A key indicators selection method of prediction model based on machine learning hybrid mode is proposed. First, the contribution degree of the selected algorithm and model is ranked according to the features. The support vector machine with good classification effect and heterogeneity with other models is used as the intermediate evaluation model. A variety of selected indicators from machine learning are tested for AUC on the intermediate model. Well-performing machine learning models are selected and combined into multiple hybrid modes. These hybrid models are tested for AUC again. Experiments demonstrate that the hybrid mode of Lasso method and random forest performed best in the AUC test. The repetition indicators of the hybrid model are then selected as important indicators of the prediction model. Finally, the correlation of the indicators is tested, and the indicators beyond the threshold are removed. The selected key indicators effectively improve the accuracy of financial fraud prediction.

1. Introduction

With the high development of social economy, people need more accurate data and information when making economic decisions. The financial indicators provided by the financial report of listed companies are the basis to measure the development status of the enterprise. If more listed companies issued a false financial report. On the one hand, it will seriously mislead investors, affect investors' investment decisions, and reduce their investment interest. On the other hand, it will greatly reduce the spontaneous resource allocation efficiency of the capital market, harm the development of the whole capital market, and even endanger the benign development of social economy and national economy.

Therefore, the financial data of listed companies are analyzed, and the financial fraud prediction model with high accuracy is constructed. It is very necessary to effectively predict the fraud behavior of listed companies. The financial fraud prediction of listed companies can help to truly and accurately reflect the operating conditions of the enterprise, improve the efficiency of the capital market, and promote the healthy and stable development of the capital market.

Current research on financial fraud prediction models has been mostly reported. The author compares the application effect of the traditional BP network, decision tree, and other models in the financial statement fraud identification model. By comparing the prediction accuracy of the model, it is concluded that the Bayesian network has a higher

prediction accuracy [1]. This paper is a combination of machine learning algorithms to first select fake feature value through artificial neural networks and support vector machines, and then construct financial fraud identification models using four types of decision trees [2]. Nurul Herawati uses M-Score models in the field of financial and financial analytics, combined with data mining technology that can more effectively identify financial fraud behavior [3]. The paper uses the text mining method to analyze the content in the financial report, and finds the fraud information in language structure of the text content [4]. The authors used decision tree and classified regression tree of machine learning models to predict the financial fraud of listed companies in the United States [5]. The experimental results show that the methods such as random forest, support vector machine, and neural network are well used in the financial fraud identification model of listed companies, among which the random forest performs best in the test set [6].

The authors firstly used the Mann–Whitney test and correlation analysis with principal component analysis to identify the financial indicators for building the model [7]. The author conducted a financial fraud study on listed companies in Malaysia, involving a total of 65 fraudulent samples and 65 non-fraudulent samples. The purpose of the research is to find suitable financial indicators for predicting financial fraud [8]. This paper builds different financial fraud identification models based on data mining technology, and conducts financial fraud identification tests combined with samples. The experimental results show that the recognition efficiency of indicators combination model of random forest and Relief algorithm is the highest [9]. The author uses a variety of statistical methods to extract features, adopts the method of three neural network fusion models, and uses the AUC value as the evaluation indicators to predict the listed companies with financial fraud [10].

In short, the key factor to build a financial fraud prediction model is how to select the evaluation indicators. Then, the existing methods have the problem of low prediction accuracy in the indicators selection. This paper proposes a novel key indicators selection method of financial fraud prediction model based on machine learning hybrid mode. First, the contribution degree of the selected algorithm and model is ranked according to the features. We select these models as Pearson's coefficient, Lasso method, multiple linear regression model, random forest model, XGBoost, and decision tree. The support vector machine with few parameters and good classification effect is used as the intermediate evaluation model. A variety of selected indicators from machine learning are tested for AUC on the intermediate model. Well-performing machine learning models are selected according to the experimental results and combined into multiple hybrid modes. These hybrid models are tested for AUC again. The optimal hybrid mode is finally determined, and the repetition indicators in this hybrid mode are used as an important indicator of the prediction model. After correlation test of indicators, the indicators beyond the threshold are removed, and the key indicators of financial fraud prediction model are finally obtained. The selected key indicators are more accurate when making fraud predictions.

2. Data Preprocessing

The dataset collected in this article comes from 11,310 financial data from 2,660 listed manufacturing companies in the first five years. To prevent the data from involving sensitive content, data masking has been conducted. There are 91 fraud samples in these financial data, and the proportion of samples for whether financial fraud is 1 : 124, similar to the actual situation, the relevant indicators are 361. Finally, use the FLAG column to indicate whether the fraud (0: normal, 1: fake). Here, the collected data is first to read by using Excel to obtain multiple data tables. These tables are merged to remove duplicate attributes, and then deletion of moderately irrelevant features is performed to prevent excessive features from causing overfitting, as well as to prevent noise problems. The specific operation is as follows:

(1) It is considered that the data for each year are relatively independent, regardless of the time series problem, so the three features of "ACT_PUBTIME," "END_DATE_REP," and "PUBLISH_DATE" can be deleted

(2) Financial fraud prediction mainly considers financial data, and removes the six features of non-financial indicators "REPORT_TYPE," "FISCAL_PERIOD," "ACCOUTING_STANDARDS," "CURRENCY_CD," "FISCAL_PERIOD," and "MERGED_FLAG." Additional features with a variance of 0 are removed using variance filtering. Finally, we get the financial data sample table, as is shown in Table 1

2.1. Processing of Missing Values. Missing data values is one of the problems frequently encountered in data analysis. For the missing values in Table 1, models and algorithms may not be used for the data without processing. At the same time, improper methods and means of processing missing values may lose a lot of information, which may get wrong conclusions in data analysis. Missing data is divided into three categories: MCAR (Missing Completely at Random), MAR (Missing at Random), and MNAR (Missing not at Random). How to fill in the missing values is a key issue in the research.

2.1.1. Two Methods of Filling in the Missing Values

(1) Delete the Missing Value. There are mainly simple deletion methods and weight methods. The simple deletion method is to directly delete samples with missing values, which is the most direct way to delete data. This method applies to a large sample size but a small missing proportion (such as 5%); the weighting method means that when the type of the missing values is MNAR, bias can be reduced by weighting the complete data. After marking the incomplete data, the complete data cases are given different weights, and the weight of the cases can be obtained by logistic or probit regression. The weighting method is not ideal for multiple missing properties.

(2) Possible Value Interpolates the Missing Values. The main idea is to interpolate missing values with the most likely

Table 1: Sample table of financial data.

Hin	NOTES_PAYABLE	INT_PAYABLE	INTAN_ASSETS	T_LIAB	OTH_RECEIV	CIP	…	FLAG
20260864.19	43137262.89		35853006.19	199014246.6	3584940.6	63918166.78	…	0
25300	1641824324		198151920.9	3745813682	85626756.05	47840630.38	…	0
150309070.2	33933660.12		161767971.2	991771386.2	37519236.14	498490316.2	…	0
	135439801.1	444138.21	133556485.4	1132316949	67249473.03	589941.46	…	0
		37133.33	7648295.11	103051049.6	775584.93	5852628.26	…	0
			81818076.32	122502787.3	11574909.7	23776020.83	…	0
	15867887.4		18069593.1	101698231.2	2214365.14		…	0
	48110000	52833.34	23280391.4	179528298.5	3127042.06	408878.15	…	0
	191162947.4		81162818.28	762841837.1	20761013.49	30169688.61	…	0
82165039.57	197499986.4	12115563.06	506346335.6	2142225797	26651977.7	112975382.2	…	0
	57160740.76		244048187	1911119201	18937852.67	183144656.5	…	0
	4000000		8601000	101149100	414500	1079500	…	0
455832602.9	357750000		433867311.4	7485796120	135391987.6	108686448	…	0
	63346265.54		30744297.61	369280322	4100364.75	79193425.27	…	0
25534371.17	330467802.8		192516535.6	2255491128	183031871.6	134753825.4	…	0
		41705890	61048005.18	6014023828	171552625.7	567848801.8	…	0
125784488	55000000	736216.35	228561097	1495908699	42011631.46	1062885591	…	0
	24267329.85	7777.77	34716946.33	132601183.6	873642.4	933681.85	…	0
4147201.25	171770	425778.78	8582593.46	553913707.1	43086003.84	117853510.9	…	0
34324301.94		274456.67	101165025.8	729939458.8	10358632.69	69771256.84	…	0

values than less information loss resulting from full deletion of incomplete samples. There are several common methods:

(1) Mean interpolation. The properties of the data are divided into numerical and non-numerical methods. If the missing value is numerical type, the missing value is interpolated with the average of the values for that feature; if the missing value is non-numerical type, the missing value is supplemented with the mode (the value with the highest frequency)

(2) Mean interpolation of the same kind. The use of mean interpolation has a disadvantage: all missing values on attribute $x^{(t)}$ containing missing values are filled with the mean value of the attribute, which may lead to a decrease in accuracy when the classification algorithm is performed subsequently. The idea of similar mean interpolation is to first classify the samples and then interpolate the missing values with the mean of the samples of that class

Known dataset $D = (x_1, y_1), (x_2, y_2) \cdots (x_N, y_N)$, where $x_i = (x_i^{(1)}, x_i^{(2)}, \cdots x_i^{(d)})^T$, $i = 1, 2, \cdots, N$. Assuming attribute $x^{(t)}$ contains missing values, divides the dataset into $D_l = (x_1, y_1)$, $(x_2, y_2) \cdots (x_l, y_l)$ and $D_u = (x_1, y_1), (x_2, y_2) \cdots (x_u, y_u)$, where $x^{(t)}$ contains valid values on D_l and missing values on D_u.

First, use the hierarchical clustering algorithm to cluster D_l. Let the result of the clustering be k clusters, calculate the mean value $\mu_1, \mu_2 \cdots, \mu_k$ of these k clusters on $x^{(t)}$.

For $x_i \in D_l$, $\widehat{x}_i^{(t)} = x_i^{(t)}$.

For $x_i \in D_u$, first cluster it, assuming it belongs to cluster $C_k (1 \le k \le K)$, then $\widehat{x}_i^{(t)} = \mu_k$.

(3) ML (Max Likelihood). Under conditions with random missing, assuming that the model is correct for the complete sample, a maximum likelihood estimation of the unknown parameters can be performed by the marginal distribution of the observed data. The usual calculation method of parameter estimation for maximum likelihood is EM (Expectation-Maximization). The proposed method is suitable for large samples. But this approach can fall into local extrema, convergence is not fast, and computationally complex

2.1.2. Processing of the Missing Values in This Paper. Due to the uneven miss rate of the rows and columns of the dataset, the number of fake samples and normal samples is greatly different. What's more, the causes of the Loss rate of Row are different from the Loss rate of Columns, so we used different methods for rows and columns.

(1) For column data missing are shown in Figure 1. This paper believes that the absence of serious financial feature indicates that this feature plays little role in judging financial fraud in the financial analysis, so removes the indicators with a missing rate of greater than 70%

(2) For the row loss rate as shown in Figure 2. The mean value of the sample attribute of whether it is fake or not may be quite different. And the classification algorithm is used in this paper. In order to prevent

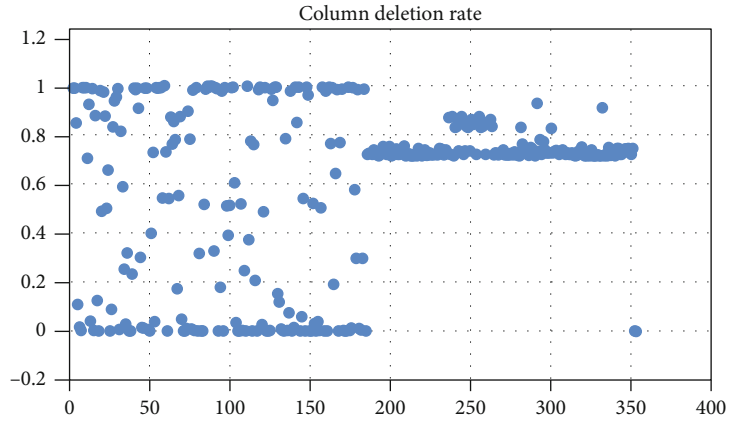

Figure 1: Scatter plot of the column missing rate.

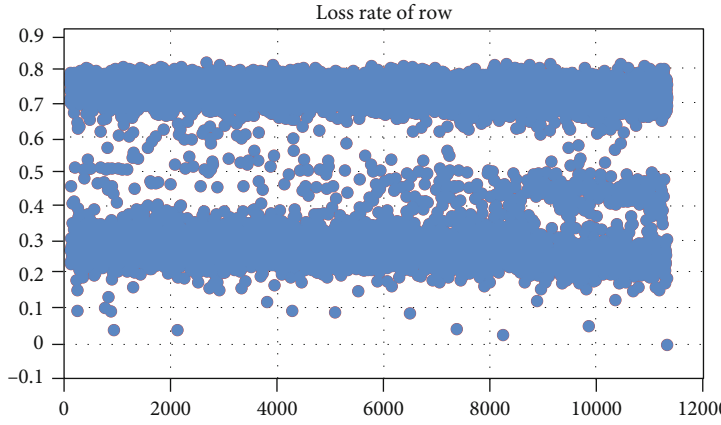

Figure 2: Scatter plot of the row missing rate.

the blurring of the discriminant boundary in the filling of missing values, we use a similar mean to interpolate the missing values, and when the loss rate is too large, using the similar mean interpolation causes the excessive error. The practice of this article is the samples are first divided into two groups of normal data and fake data group according to whether the sample is fake, the mean filling is used for samples with normal data missing rates within 70%; samples with a missing rate of more than 70% are deleted; for the absence of fake data, because of the few data being fake, the deletion will lead to the poor generalization ability of the construction model, so we keep all the fake samples, and the missing value is filled with the mean

2.2. Data Standardization. The samples selected in this paper are mainly numerical indicators, including different categories of numerical indicators, and there are large differences between different factors in the range of values and units of measure so that we are unable to compare, weight, and other subsequent operations on the different indicators, so we need to standardize the indicators data.

Common data standardization methods are Min-Max, Z-score, decimal scaling, quantitative feature binarization, etc. Since the data selected in this paper may have extremes, we use the Z-score normalization method. Z-score normalization is a normalization method to transform the data into standard normal distribution. The specific calculation formula is:

$$\bar{x} = \frac{x - \bar{x}}{\text{std}(x)}, \tag{1}$$

where the$\bar{x}$ represents the mean value and std(x) represents the standard deviation.

Considering that the following situation may occur in the real environment: ① data will be continuously input into the model, and the mean and variance cannot be obtained; ② the training set is simulating data in a real environment and cannot directly use, it is own mean and variance; ③ in real environments, single data cannot be normalized. To solve these three problems, we first obtain the parameters (mean, variance) in the training set; the entire dataset is then standardized using the Z-score normalization method. The specific flow is shown in Figure 3.

In Figure 3, fit represents the mean and variance obtained according to the training dataset, returning a Scalar object; Transform means that according to the obtained mean and variance, the Z-score method is used to

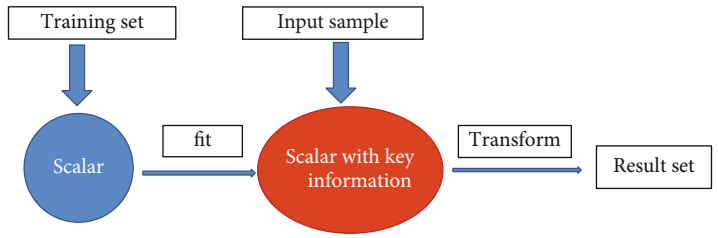

Figure 3: Standardized flow figure of sample data.

Table 2: Financial data schedule.

	Quantity
Fake samples	8598
Normal samples	91
Factors	106

standardize the training set and test set data at the same time. The final result is obtained.

Finally, the dataset treated with missing values is normalized using the above Z-score normalization method in the experiment, and the data results are shown in Table 2.

3. Feature Selection Methods Based on Multiple Machine Learning Models

The most appropriate feature selection method for the financial fraud prediction model is the feature selection based on multiple machine learning algorithms and models, such as Pearson, LR, RF, and DT. These machine learning models themselves have a mechanism of scoring features, which are easily applied to feature selection tasks. We first introduce the feature selection based on the machine learning model. Finally, these algorithms and models are used to obtain the relevant features.

3.1. Feature Selection Method Based on Machine Learning Model

3.1.1. Feature Selection Based on Pearson's Correlation. Pearson's correlation coefficient measures the magnitude of a degree of linear correlation, and the greater the absolute value of the correlation coefficient is, the stronger the degree of linear correlation is. The range of the correlation coefficients is found between [-1,1]. Assuming two variables, X and Y, then there are:

(i) The X and Y variables are not correlated when the correlation coefficient is 0

(ii) When the values of X and Y values increase or decrease at the same time, the two variables are positively correlated, and the correlation coefficients are between 0 and 1

(iii) When the value of X increases and the Y value decreases, or the X value decreases while the Y value increases; the two variables have a negative correlation, with correlation coefficients between-1 and 0

Its formula is:

$$\rho(X \cdot Y) = \frac{\sum(X\text{-}\bar{X})(Y\text{-}\bar{Y})}{\sqrt{\sum(X\text{-}\bar{X})^2\sum(Y\text{-}\bar{Y})^2}}. \tag{2}$$

In formula (2), X and Y are two variables, respectively. $\bar{X}$ and $\bar{Y}$ denote the mean values of X and Y, respectively.

The correlation coefficient is defined only when the standard deviation of both variables is not zero. Scope of application of Pearson's correlation coefficient:

(1) There is a linear relationship between the two variables, both with continuous data

(2) Overall normal distribution of two variables, or near-normal unimodal distribution

(3) The two variables are in pairs and each pair is independent of each other

3.1.2. Feature Selection Based on Lasso. Lasso (least absolute shrinkage and selection operator) is a regression method suitable for multicollinearity problems and can implement feature selection while parameter estimation. The Lasso method is a compression estimation method of reducing the set of variables. It constructs a penalty function that can compress the coefficients of variables and change some regression coefficients to 0, thus achieving the purpose of variable selection [11].

Lasso regression is performed by imposing a penalty term on the coefficients of the model, so that some coefficients tend to 0 based on the least-squares estimation, to achieve the purpose of variable selection, and also avoid overfitting, ensure the interpretability and simplicity of the model.

Consider the following linear model:

$$Y = X\beta + \varepsilon. \tag{3}$$

Y is the vector of $n \times 1$, X is the matrix of $n \times p$, and ε is the vector of $n \times 1$.$\beta = (\beta_1, \beta_2, \cdots, \beta_p)$ is the regressive coefficient variable of $p \times 1$. Its Lasso is estimated at:

$$\widehat{\beta} = \arg\min_{\beta}\left\{\|Y - X\beta\|_2^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \tag{4}$$

Among them, $\lambda\sum_{j=1}^{p}|\beta_j|$ is the penalty term, and $\lambda > 0$ is the reconciliation parameter, used to control the penalty strength of the model and in turn to control the number of explained variables. When λ is small, the weight ratio of the first part of the upper formula will increase, to minimize the sum of the overall residual squares, so that more explained variables will be added to the model; when λ is large, the weight ratio of the second part of the formula above increases, the regression coefficients of many explanatory variables are compressed and tend to be 0.

3.1.3. Multiple Linear Regression Feature Selection. The multiple linear regression model is a model that explains the dependent variable by using two or more explanatory variables. Let Y is the dependent variable, $X_1X_2\cdots, X_k$ are k different variables described as explanatory variables. Where X_1 is constantly equal to 1, the multiple linear regression model is performed as in Equation (5):

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i (i = 1, 2, \cdots, n), \quad (5)$$

where $\mu_i(i = 1, 2, \cdots, n)$ is a random perturbation term; the parameter $\beta_1, \beta_2, \cdots, \beta_k$ is called a regression coefficient. Hypothesis,

$Y = (Y_1 Y_2 \vdots Y_n), X = (X_{11} \quad X_{21} \quad \cdots X_{k1} X_{12} X_{22} \cdots X_{k2} \vdots \vdots \vdots \vdots X_{1n} X_{2n} \cdots X_{kn}), \beta = (\beta_1 \ \beta_2 \vdots \beta_k), \mu = (\mu_1 \ \mu_2 \vdots \mu_n)$. Then, formula (5) is expressed in the matrix form as formula (6).

$$Y = X\beta + \mu \quad (6)$$

The relationship between one or more independent variables is modeled using the least-squares function. Regression analysis in mathematical statistics is a statistical analysis method used to determine the quantitative interdependent relationship between two or more variables. According to the principle of the least-squares method, the estimate of $\beta_i(i = 0, 1, 2, \cdots, k)$, the value of b_i $(i = 0, 1, 2, \cdots, k)$, should be let by

$$Q = \sum_{a=1}^{n} \left(y_a - \hat{y}_a\right)^2 = \sum_{a=1}^{n} [y_a - (b_0 + b_1 x_{1a} + b_2 x_{2a} + \cdots + b_k x_{ka})]^2 \longrightarrow \min. \quad (7)$$

According to the definition of the least-squares method, in the linear regression model, the estimate of the regression coefficient that minimized the residual sum of squares is called the least-squares-estimation. Equivalent to the smallest $\beta^\wedge$ that makes $\mu^{\wedge\prime}\mu^\wedge = (Y - X\beta^\wedge)'(Y - X\beta^\wedge)$ up, among it, $\mu'^\wedge$ is the transpose of $\mu^\wedge$. To minimize $\beta^\wedge$, $\mu^{\wedge\prime}\mu^\wedge = (Y - X\beta^\wedge)'(Y - X\beta^\wedge)$ can be seen as a function about $\beta^\wedge$, then the first-order partial derivative of $\beta^\wedge$ must be 0, namely, $\partial\mu^{\wedge\prime}\mu^\wedge/\partial\beta^\wedge = -2X'Y + 2X'X\beta^\wedge = 0$. Thus, we get equation $X'X\beta^\wedge = X'Y$, and so $\beta^\wedge = (X'X)^{-1}X'Y$.

3.1.4. Feature Selection Based on Random Forest Model. In machine learning, random forest (RF) is a classifier containing multiple decision trees, and the category of its output is determined by the mode of categories output by the individual trees, and the underlying classifier that constitutes the random forest is called the decision tree. The random forest has the advantages of high accuracy, robustness, and easy to use, making it one of the most popular machine learning algorithms today. There are two ways to calculate feature importance in random forest: one method is based on OOB error, called MDA (Mean Decrease Accuracy); another method is based on Gini impurity, called MDG (Mean Decrease Gini). Both methods are the more the numerical decreases and the more important the representation features [12].

The MDA is specifically described below:

(1) A random forest model is trained to test the OOB error for each tree in the model using the out-of-bag sample data

(2) The value of the variable v in the out-of-bag sample data is randomly shuffled to retest the OOB error of each tree

(3) The mean of the difference in OOB error for the two tests is the measure of the importance of a single tree to the variable v

The calculation formula is:

$$\mathrm{MDA}(v) = \frac{1}{\mathrm{ntree}} \sum_{t} \left(\mathrm{errOOB}_t - \mathrm{errOOB}_t'\right). \quad (8)$$

The MDG is specifically described below: Gini-based variable importance is measured by the degree of reduced Gini purity due by use of variable v. At classification node t, the Gini coefficient impurity is:

$$G(t) = 1 - \sum_{K=1}^{Q} p^2(k/), \quad (9)$$

where Q represents the total number of categories for the target variable, $p(k/t)$represents the conditional probability that the target variable is class k in the node t. The Gini nonpurity drop value for each tree was calculated from the formula, and then the results were averaged across all the trees.

3.1.5. XGBoost Feature Selection. The extreme gradient lifting algorithm (XGBoost) is a tree-based Boosting algorithm [13]. Compared with the traditional gradient improvement decision tree algorithm, the XGBoost algorithm innovatively uses the second derivative information of the loss function, so that the XGBoost algorithm can converge faster, ensure high solution efficiency, and also increase the scalability. Because a function meets the second-order derivable condition, the function can be used as a custom cost function when appropriate. Another advantage of the XGBoost algorithm is that it borrows the column sampling method in the

random forest algorithm and further reduces the computation and overfitting.

XGBoost calculates which feature to select as a segmentation point based on the gain of the structure fraction, and the importance of a feature is the sum of its occurrences in all trees. That is, the more times an attribute is used to build a decision tree in a model, the relative importance it is. The XGBoost algorithm can be expressed as:

$$\widehat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in F, \tag{10}$$

where K indicates the number of trees, $F = \{f(x) = \omega_{q(x)}\}(q : R^m \longrightarrow T, \omega \in R^T)$represents the function space of the model, and $f_K(x_i)$ represents the classification results of the i sample in the K tree. As seen from the expression of the XGBoost algorithm, the model is a collection of iterative residual trees, one tree is added to each iteration, and each tree eventually forms the linear combination of the K tree by learning the residuals of the former $(N-1)$ tree.

3.1.6. Feature Selection Based on Decision Tree. The structure is similar to that of the tree and consists of oriented edges and nodes. There are three main types of nodes, include the root node, intermediate node, and leaf node. At the top of the decision tree is the root node, which contains the most informative properties; at the bottom of the decision tree are the leaf nodes, representing the results of the classification; between the root and leaf nodes are intermediate nodes used for the testing of feature properties [14].

When using the decision tree, the original sample is divided into the training set and the test set, first training a decision tree with the strongest generalization ability, and then predicting using the test set to calculate the generalization error. When training a decision tree, feature selection is about deciding whether to use an indicator as a division basis (as an intermediate node) to help with the classification. The general decision tree algorithm relies on three criteria, namely, information gain, information gain ratio, and Gini index.

Information entropy is an indicator used to measure the uncertainty of sample sets, and the more uncertain the set is, the greater the information entropy. Assuming that the proportion of class i samples in sample set D is $p_i(i = 1, 2, 3, \cdots, |y|)$, then the information entropy of sample set D is defined as:

$$\mathrm{Ent}(D) = -\sum_{i=1}^{|y|} p_i \log_2 p_i. \tag{11}$$

For dataset D, assuming that feature A is selected as the decision tree judgment node, then the information entropy after the action of feature A is defined as:

$$\mathrm{Ent}_A(D) = -\sum_{j=1}^{k} \left|\frac{D_j}{D}\right| \times \mathrm{Ent}\left(D_j\right). \tag{12}$$

It was found that decision trees would tend to choose those features with more attribute values when adopting information gain as the criterion for feature selection. To reduce the possible adverse effects of this decision tree preference, the C4.5 decision tree algorithm improves on the original decision tree algorithm. It uses the information gain ratio as a criterion for selecting the optimal partition properties. Information gain ratio is defined as:

$$\mathrm{Gain\_ration}(D, A) = \frac{\mathrm{Gain}(D, A)}{IV(A)}. \tag{13}$$

Where in equation (13), $IV(A)$ is known as a fixed value of feature A:

$$IV(A) = -\sum_{j=1}^{k} \left|\frac{D_j}{D}\right| \times \log_2 \left|\frac{D_j}{D}\right|. \tag{14}$$

Gini index is also a criterion for a decision tree to select the optimal partition property, representing the probability that a randomly selected sample is misclassification in the set of samples. For sample set D, assuming K categories, the probability of samples belonging to the Kth category is set to A. Then, the Gini index of this probability distribution is:

$$\mathrm{Gini}(p) = \sum_{k=1}^{k} p_k(1 - p_k) = 1 - \sum_{k=1}^{k} {p_k}^2. \tag{15}$$

For the dataset D, its Gini index is:

$$\mathrm{Gini}(D) = 1 - \sum_{k=1}^{k} \left(\frac{|C_k|}{|D|}\right)^2, \tag{16}$$

where $|C_k|$ represents the number of samples belonging to category k in the sample set D.

We have a test at each internal node representation on one attribute by using the tree structure of the decision tree based on the selected features, and test the output on the branches of the tree. Child nodes are generated recursively from top to bottom according to the selected feature evaluation criteria until the decision tree stops growing when the dataset is not separable. The necessary pruning is performed to narrow down the tree structure and alleviate overfitting.

3.2. Experiments of Important Feature Selection. The training set and the test set are divided by using the train_test_split method of model_selection in machine learning sklearn; the ratio of the training set to the test set is determined as 7 : 3. To prevent the sample division from affecting the prediction results, random seeds are set when dividing the training set and the test sets. "FLAG" serves as a label column, and other indicators serve as feature columns. Using

the data from the training set, for the Pearson correlation coefficient method, all correlation indicators and label columns are imported to find the top 20 features for the correlations with the label column; For the Lasso method, a total of nonzero16 metrics with a weight coefficient are found; the top 20 features are ranked on the model using all correlation metrics to find the top features, respectively. Features are ranked using feature importance on the LR, RF, XGBoost, and DT models, respectively, finding the top 20 features. The main purpose is to obtain the top 20 attributes using sklearn's ranking of feature_importances_ attribute values. The inherent mechanism of DT is to discriminate the effect of features on the purity increase of nodes based on the actual discriminating criteria such as GINI, information entropy, and information entropy gain. The parameters used on LR, RF, XGBoost, DT models for indicators selection are shown in Table 3. These two algorithms of Pearson's and Lasso do not use parameters. Table 3 shows the names and values of these parameters.

Finally, through experiments, the important indicators related to financial fraud are selected by each algorithm and model, as shown in Table 4.

Table 3: Parameters list of machine learning model.

No	Model name	Parameters	
		Name	Value
1	LR	penalty	L1
		C	1.0
		solver	liblinear
		multi_class	over
2	RF	n_estimators	10
		max_features	20
		splitter	best
3	XGBoost	booster	gbtree
		max_depth	6
		learning_rate	0.3
4	DT	n_estimators	10
		splitter	best
		max_depth	3

4. Key Indicators Selection Method of Financial Fraud Prediction Model Based on Machine Learning Hybrid Mode

4.1. Intermediate Evaluation Indicator. In this paper, the performance measures of the classification model adopt the confusion matrix and True Positive Rate, False Positive Rate, and AUC. In the binary classification problem, the confusion matrix is a second-order matrix. Among them, this paper uses normal samples as positive (0) and fake samples as negative (1), using *TPFPTNFN* to indicate four cases of whether the classifier predicted or correct on the dataset.

TP: True Positive, predict positive class to be positive;

FP: False Positive, predict negative class to be positive;

TN: True Negative, predict negative class to be negative;

FN: False Negative, predict positive class to be negative;

The representation in the confusion matrix is shown in Table 5.

TPR (True Positive Rate) represents the proportion of true classes in the samples predicted to be positive. The calculation formula is:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{17}$$

FPR(False Positive Rate) represents the proportion of true classes in samples predicted to be negative. The calculation formula is:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{18}$$

The AUC is the area under the ROC curve, giving the average performance value of the classifier. AUC values can be used to evaluate the effect of binary classification problem. The closer the AUC is to 1.0, the higher the authenticity of the detection method. The calculation formula is:

$$\mathrm{AUC} = \frac{1 + \mathrm{TPR\text{-}FPR}}{2}. \tag{19}$$

4.2. Key Indicators Selection Method Based on Machine Learning Hybrid Mode. Features are ranked using feature importance on the Pearson, Lasso, LR, RF, XGBoost, and DT models, respectively, to find the top 20 features. In order to prevent the homogeneous model from restricting the selection of indicators, the selected intermediate model should be heterogeneous with the above six models. The support vector machine (SVM) with fewer parameters and good effects is used as the intermediate model to conduct AUC tests on the indicators selected by various models. Specifically, for the features selected by each model, the corresponding data in the feature columns of the training set are, respectively, selected as the feature columns of the new training set, and imported into the support vector machine model for retraining. The data corresponding to the test set are selected as a new test set feature column for prediction.

We first test the values of the AUC on the support vector machines for the features selected by the Pearson, Lasso, RF, XGBoost, DT, LR. The experimental results are shown in Figure 4:

The models and algorithms used in this paper are divided into three categories, include tree model (DT, RF, XGBoost), algorithm (Lasso, Pearson), and basic model (Logistic). The models with the first AUC ranking are selected from these three categories, and the final models selected are RF, Lasso, and Logistic. Then, these three models with good performance are mixed. In this paper, the network search mechanism is used to exhaustively list all the combined models, so that the optimal hybrid model can be compared easily. Finally, all the hybrid models are formed as RF+ Lasso, Lasso+ Logistic, RF+ Logistic, and

Table 4: Selection table of important indicators based on machine learning methods.

Num	Model					
	Pearson	Lasso	LR	RF	XGBoost	DT
1	AR	INT_PAYABLE	T_LIAB_EQUITY	INT_PAYABLE	T_ASSETS	AVAIL_FOR_SALE_FA
2	PREPAYMENT	T_LIAB	T_PROFIT	T_LIAB	T_LIAB	T_LIAB
3	OTH_RECEIV	C_OUTF_FR_INVEST_A	T_COGS	C_OUTF_FR_INVEST_A	N_CE_END_BAL	INT_PAYABLE
4	AVAIL_FOR_SALE_FA	INTAN_ASSETS	N_CF_FR_INVEST_A	INTAN_ASSETS	AVAIL_FOR_SALE_FA	N_CE_END_BAL
5	CIP	LT_EQUITY_INVEST	N_CF_OPERATE_A	LT_EQUITY_INVEST	LT_EQUITY_INVEST	T_REVENUE
6	DEFER_TAX_ASSETS	DILUTED_EPS	N_CF_FR_FINAN_A	DILUTED_EPS	REVENUE	C_OUTF_FR_FINAN_A
7	OTH_NCA	NOPERATE_EXP	N_INCOME	NOPERATE_EXP	C_INF_FR_OPERATE_A	C_FR_SALE_G_S
8	NOTES_PAYABLE	ADVANCE_RECEIPTS	T_EQUITY_ATTR_P	ADVANCE_RECEIPTS	C_FR_OTH_INVEST_A	INVEST_INCOME
9	INT_PAYABLE	C_PAID_TO_FOR_EMPL	C_OUTF_OPERATE_A	C_PAID_TO_FOR_EMPL	N_INCOME_ATTR_P	OTH_RECEIV
10	PAID_IN_CAPITAL	C_PAID_FOR_DEBTS	REVENUE	AVAIL_FOR_SALE_FA	FOREX_EFFECTS	CASH_C_EQUIV
11	C_PAID_OTH_FINAN_A	LT_AMOR_EXP	C_PAID_G_S	C_PAID_OTH_FINAN_A	C_PAID_TO_FOR_EMPL	C_OUTF_FR_INVEST_A
12	N_CF_FR_INVEST_A	C_FR_OTH_INVEST_A	T_CL	C_FR_OTH_INVEST_A	C_PAID_FOR_OTH_OP_A	ADVANCE_RECEIPTS
13	C_FR_MINO_S_SUBS	MINORITY_INT	NOPERATE_EXP	FOREX_EFFECTS	MINORITY_GAIN	INTAN_ASSETS
14	N_CHANGE_IN_CASH	T_NCL	C_OUTF_FR_INVEST_A	C_INF_FR_OPERATE_A	INTAN_ASSETS	NOPERATE_INCOME
15	GAIN_INVEST	FIXED_ASSETS	MINORITY_INT	GAIN_INVEST	INT_PAYABLE	T_EQUITY_ATTR_P
16	FOREX_EFFECTS	C_PAID_INVEST	FOREX_EFFECTS	T_ASSETS	C_PAID_OTH_FINAN_A	BIZ_TAX_SURCHG
17	C_FR_OTH_OPERATE_A		T_NCL	N_CE_END_BAL	T_NCA	C_PAID_FOR_TAXES
18	N_CF_FR_FINAN_A		T_COMPR_INCOME	CFSGS_R	MINORITY_INT	C_PAID_TO_FOR_EMPL
19	ASSETS_IMPAIR_LOSS		AVAIL_FOR_SALE_FA	C_OUTF_FR_INVEST_A	DILUTED_EPS	COGS
20	NOPERATE_EXP		FIXED_ASSETS	T_NCA	AVAIL_FOR_SALE_FA	LT_EQUITY_INVEST

Table 5: Confusion matrix.

True	Predict	
	0	1
0	TP	FN
1	FP	TN

RF+ Lasso+ Logistic. The repeated features in these hybrid models are selected, and finally these hybrid models are tested again on the support vector machine for AUC values. The experimental results are shown in Figure 5.

The experimental results in Figure 5 show that the features jointly selected by Lasso and the RF performed best on the SVM. Therefore, the corresponding selected indicators are shown in Table 6.

4.3. Correlation Analysis of the Key Indicators. Due to the internal logic of the financial statement data itself, there is also some degree of autocorrelation between the key features selected in this paper, which will affect the accuracy of the model estimate. Therefore, before substituting into the model, we first have to analyze multicollinearity and remove the features with high partial autocorrelation. We use

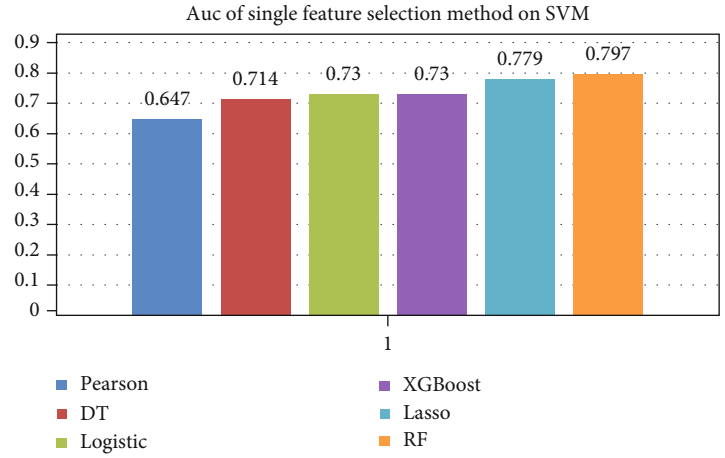

Figure 4: AUC comparison result of individual feature selection methods on the SVM.

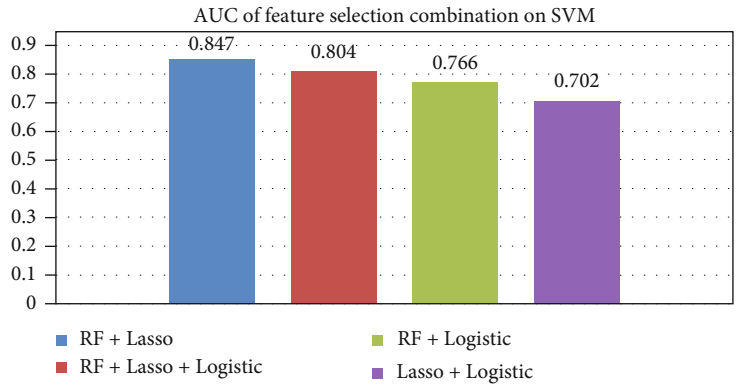

Figure 5: AUC comparison result of repeated features selected by hybrid models on the SVM.

Table 6: Important indicators selected on the optimal hybrid model.

Num	Feature	Num	Feature
1	INT_PAYABLE	5	DILUTED_EPS
2	C_OUTF_FR_INVEST_A	6	NOPERATE_EXP
3	INTAN_ASSETS	7	ADVANCE_RECEIPTS
4	LT_EQUITY_INVEST	8	C_PAID_TO_FOR_EMPL

autocorrelation to conduct experiments on the explanatory power of the features in the table described above and to analyze the multicollinearity problem of these features. After autocorrelation analysis, we remove features with autocorrelation greater than the threshold.

Correlation coefficients are the amount of the degree of linear correlation between the studied variables, generally indicated by the letter r. Due to the different subjects, the correlation coefficient is defined in many ways, including the Pearson correlation coefficient. The correlation is considered strong if the absolute value of the correlation coefficient r for A and B is above 0.7; correlation coefficient r between 0.3 and 0.7, correlation is weak; correlation coefficient r below 0.3, no correlation. In this paper, the Pearson coefficient method is used to calculate the linear correlation coefficient between the two features. The corresponding calculation formula is:

$$\rho(X \cdot Y) = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}. \quad (20)$$

The correlation experiment is performed on the 8 indicators selected in Table 6, and the correlation between the

Figure 6: Correlation figure of the key indicators.

Table 7: The final key indicators.

Num	Feature
1	INT_PAYABLE
2	C_OUTF_FR_INVEST_A
3	INTAN_ASSETS
4	LT_EQUITY_INVEST
5	DILUTED_EPS
6	NOPERATE_EXP
7	ADVANCE_RECEIPTS

indicators is calculated by formula (20). The experimental results are shown in Figure 6.

As is shown in Figure 6, the autocorrelation value for both features "INTAN_ASSETS" and "C_PAID_TO_FOR_EMPL" is 0.76, greater than the set threshold of 0.7. Therefore, these two features belong to the strong correlation feature, and considering the correlation of the features, the "C_PAID_TO_FOR_EMPL" feature is deleted. Final key indicators for the prediction model are shown in Table 7.

5. Conclusion

This paper mainly analyzes the selection method of key indicators in the financial fraud prediction model, and proposes the key indicators selection method based on the machine learning hybrid model because of the existing feature selection method with low prediction accuracy. First for the dataset, preprocessing includes missing value processing and standardization of the data. The feature selection methods in multiple machine learning models are then described and the selected top 20 features for each model. Support vector machine is used as the intermediate model for AUC testing, the top models are combined, and the repeated features in the hybrid model are selected as the pre-selection features. Tested the selected pre-selected features, and the model combined with the highest AUC values is selected. Finally, through the correlation experiment of the indicators, the final key indicators are obtained. The novel key indicators selection method based on machine learning hybrid model

is proposed here in this paper and effectively improves the prediction accuracy. This novel key indicators method provides an important basis for the construction of the financial fraud prediction model. However, the number and types of models for feature selection in the experiment are large, and the optimization of all parameters will cause a large time complexity, thus affecting the experimental effect. Therefore, the default parameters are used in this paper. Future work will focus on parameter optimization of the models and building a financial fraud prediction model by using the selected key indicators.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work was supported by the National Natural Science Funds of China (61272015) and 2022 Henan Province Science and Technology Research Project: "Construction and Application of Intelligent Ontology in Internet of Things Based on Semantic Concept Model" (222102210316).

References

[1] E. Kirkos and C. Spathis, "Data mining techniques for the detection of fraudulent financial statements," *Expert Systems with Applications*, vol. 32, no. 4, pp. 995–1003, 2007.

[2] C.-l. Jan, "An effective financial statements fraud detection model for the sustainable development of financial markets: evidence from Taiwan," *MDPI*, vol. 10, no. 2, p. 513, 2018.

[3] N. H. Tarjo and N. Herawati, "Application of Beneish M-score models and data mining to detect financial fraud," *Procedia - Social and Behavioral Sciences*, vol. 211, no. 211, pp. 924–930, 2015.

[4] L. Purda and D. Skillicorn, "Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection," *Contemporary Accounting Research*, vol. 32, pp. 1193–1223, 2015.

[5] S. Chen, "Detection of fraudulent financial statements using the hybrid data mining approach," *Sprinter Plus*, vol. 5, no. 1, pp. 2–16, 2016.

[6] H. U. A. N. G. Zhigang, L. I. U. Jiajin, and L. I. N. Chaoying, "A comparative study on the cutting-edge methods in identifying financial statement fraud of listed companies based on machine learning," *Journal of Systems Science and Mathematical Sciences*, vol. 40, no. 10, pp. 185–203, 2020.

[7] Z. Haili, *The recognition research of Chinese listed companies' financial fraud*, Southwestern University of Finance and Ecnomics, 2016.

[8] H. Dalnial, A. Kamaluddin, Z. M. Sanusi, and K. S. Khairuddin, "Accountability in financial reporting:detecting fraudulent firms," *Procedia-Social and Behavioral Sciences*, vol. 145, pp. 61–69, 2014.

[9] F. Bingchun, "Construction of financial fraud recognition model based on data mining technology," *Financial and Accounting Communications*, vol. 5, pp. 93–97, 2019.

[10] C. Jingbo and C. A. I. Zhijie, "Prediction of financial fraud of listed companies based on deep learning model," *Mathematical Modeling Its Applications*, vol. 3, no. 10, pp. 54–59, 2021.

[11] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society*, vol. 67, no. 1, pp. 91–108, 2005.

[12] Q. Wu, Y. Ye, H. Zhang, M. K. Ng, and S. S. Ho, "ForesTexter: an efficient random forest algorithm for imbalanced text categorization," *Knowledge-Based Systems*, vol. 67, pp. 105–116, 2014.

[13] B. Pan, "Application of XGBoost algorithm in hourly PM2.5 concentration prediction," *IOP Conference Series: Earth and Environmental Science*, vol. 113, article 012127, 2018.

[14] Y. Xia, C. Liu, Y. Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," *Expert Systems with Applications*, vol. 78, pp. 225–241, 2017.