Inzinerine Ekonomika-Engineering Economics, **2024, 35(4), 439–452**

Financial Distress Prediction Using an Artificial Neural Network Integrating a Two-Stage Feature Selection Method

Zijiao Zhang[1], Shiyou Qu[2], Chong Wu[3,*]

School of Management, Harbin Institute of Technology
92 West Dazhi str., Nan Gang District, Harbin City, China
E-mail: [1]zhangzijiao@hit.edu.cn; [2]qushiyouhit1@163.com; [3]wuchong@hit.edu.cn; *Corresponding author*

https://doi.org/10.5755/j01.ee.35.4.30924

Financial distress prediction (FDP) is critical for companies, banks, and investors, and artificial neural networks (ANN) have been proven to be an efficient method for FDP. However, the "curse of dimensionality" in FDP not only increases the computational complexity, but also reduces the prediction accuracy. To solve this problem, this paper takes an ANN model as the basic classifier and presents a new two-stage feature selection method integrated with multiple filters and a wrapper method. The financial data of Chinese listed companies are applied for comparative analysis to verify the effectiveness of the constructed method. The results demonstrate that the proposed method achieves a smaller feature subset and better predictive effect than other methods, thus solving the "curse of dimensionality" more effectively and improving the accuracy. In addition, SHapley Additive exPlanations (SHAP) and Partial Dependence Plots (PDP) are employed to investigate the relative importance of selected features. Their results increase the credibility of the proposed model, giving users more confidence in using this "black box" model.

Keywords: *Financial Distress Prediction; Feature Selection; Artificial Neural Network; Genetic Algorithm; SHapley Additive exPlanations.*

Introduction

When a company is in financial distress, it not only affects its sustainable operations, but may also lead to an increase in bad debts for banks and financial losses for investors. Therefore, it is essential to accurately predict a company's financial distress, especially in increasingly competitive and uncertain markets. Accurate financial distress prediction (FDP) models can help managers recognize deteriorating financial conditions early and alert them to take timely action. In addition, banks can effectively avoid bad loans, and investors can avoid financial losses by using accurate models for FDP. Conversely, if FDP models are not precise enough, investors may miss good investment opportunities, and banks may lose customers, resulting in lower returns. To sum up, it is critical to establish accurate FDP models for these stakeholders.

Financial distress prediction assesses whether a company will suffer financial distress by constructing a prediction model. This model unveils the functional relationship between the financial data at time *t-m* and the financial status at *t*. Exploring effective FDP models has always been an essential research topic that attracts scholars and practitioners (Karas & Reznakova, 2017). More and more machine learning models are being widely applied in FDP. While machine learning models improve accuracy to some extent, other problems in the forecasting process can affect the accuracy of these models and their use by stakeholders.

The "curse of dimensionality", which refers to a multitude of redundant indicators reflecting the financial situation, is an urgent problem to be solved in FDP (Du *et al.*, 2020). Too much dimensionality not only increases the computational burden, but also leads to overfitting of predictive models and reduces accuracy. Therefore, mining critical information from massive data without compromising the final prediction results is a vital step in FDP. Feature selection (FS), which aims to search for key information, is the main way to solve this problem. However, the existing FS methods have their shortcomings.

In addition, it is also difficult for stakeholders to trust machine learning models as most of them are difficult to interpret (Ariza-Garzon *et al.*, 2020). Due to their complexity, machine learning models cannot reveal their internal mechanisms, and their predictions are often difficult to explain and validate. These complex machine learning models have the "black box" problem (Shin, 2021). However, the interpretability of FDP models is critical for stakeholders. These stakeholders want to know not only the predicted outcomes, but also whether the results are reliable. Improving the interpretability of machine learning models is also an essential content in FDP. Therefore, with regard to the two problems of the "curse of dimensionality" and "black box" in FDP, our study makes the following contributions. First, this paper proposes an ANN model integrating a two-stage feature selection method to address the "curse of dimensionality" problem. In this method, an ANN model is taken as the basic classifier for FDP and the new two-stage method is applied to FS. The new two-stage method focuses on the fusion of multiple filters and a wrapper to overcome the shortcomings of existing FS methods. The proposed method can quickly and efficiently obtain a feature subset with a good prediction effect. In addition, SHapley Additive exPlanations (SHAP) and Partial Dependence Plots (PDP) are employed in this paper to investigate the relative importance of the selected features, which can improve the interpretability and credibility of the "black box" model.

The remaining paper is organized as follows. Section 2 is a literature review. Section 3 describes the details of our proposed method. Section 4 gives the case study and discusses the results. Finally, Section 5 outlines our conclusion.

Literature Review

This section aims to systematically organize and summarize the previous literature to draw out why to build the ANN model integrating the two-stage FS method for FDP. Therefore, a literature review is conducted from two aspects of predictor construction and feature selection in FDP. Among them, the two-stage FS method is the main innovation point, so the literature review of feature selection is detailed.

According to previous research, FDP predictors have transitioned from statistical methods (Beaver *et al.*, 2005) to machine learning models (Tsai *et al.*, 2021). The hypothesis of sample distribution limits the applications of traditional statistical methods. However, machine learning models do not need to obey this assumption, and their accuracy is higher, so they are getting more and more attention. Now, the popular machine learning methods in FDP are artificial neural networks (Halim *et al.*, 2021), decision trees (Ben Jabeur *et al.*, 2020), support vector machines (Kim *et al.*, 2018), random forest models (Malakauskas & Lakstutiene, 2021), adaptive boosting models (Sun *et al.*, 2020), extreme gradient boosting models (Liu *et al.*, 2019) and categorical boosting models (Ben Jabeur *et al.*, 2021) and so on. ANN has been proven to be an effective FDP method in previous studies, and ANN is designed as the predictor in this paper.

Previous literature concludes that financial indicators are key for FDP (Bellovary *et al.*, 2007; Dimitras, *et al.,*1996; Mensah, 1984). With the development of society and technology, financial indicators have become more sophisticated and complex, and the number of financial indicators that can be collected and stored has increased. Thus, there are more and more financial indicators for FDP, leading to the “curse of dimensionality”. FS is an essential and widely used technology to solve the problem of the “curse of dimensionality”. FS is to select the most effective feature subset from original feature set according to specific criteria, which is a complex process. Obtaining an optimal feature subset is usually intractable, and many FS problems have been proven to be NP-hard (Jeong *et al.*, 2015).

Many scholars have developed various FS methods, which can be divided into three categories: filters, wrappers and embedded models. In filters, all features are evaluated by the general characteristics of data (Deng *et al.*, 2019). A feature subset is formed by selecting features with higher scores and deleting those with lower scores. Evaluation criteria include mutual information (Bennasar *et al.*, 2015), information gain (Omuya *et al.*, 2021), Pearson’s correlation (Liu *et al.*, 2020), t-test (Zhu *et al.*, 2020) and Relief (Ul Haq *et al.*, 2020). As filters are independent of prediction algorithms, the feature subset obtained by filters may not perform well (Solorio-Fernandez *et al.*, 2020). Nevertheless, filters are widely applied in high-dimensional data problems due to their high computational efficiency (Sayed *et al.*, 2019).

Unlike filters, wrappers need to specify predictors in advance and use prediction accuracy as a criterion for evaluating feature subsets (Nouri-Moghaddam *et al.*, 2021). To improve the prediction performance, wrappers should find a more suitable feature subset from the candidate subsets. From the perspective of computational time, searching all subsets is neither practical nor efficient, especially when the number of features is extensive. Heuristic search is a standard method that adds or removes one or more features at a time from the candidate feature subset, including forward search, backward search and two-way search (Hancer *et al.*, 2015). Compared to the exhaustive method, these search ways reduce computational cost, but may fall into a local optimum. Metaheuristics, such as Tabu search (Yan *et al.*, 2018), genetic algorithm (Mehanovic *et al.*, 2021), simulated annealing (Mafarja & Mirjalili, 2017) and ant colony optimization (Ghosh *et al.*, 2020), have been employed as search strategies. In general, the combination of search strategies and classification models can yield feature subsets with high classification effectiveness. However, they are much less computationally efficient than filters due to the large search space.

Embedded methods take feature selection as part of the training process, selecting a feature subset that contributes most to the prediction when building a predictor (Jiang *et al.*, 2020). They need to choose features and determine model parameters simultaneously, making it difficult to find a globally optimal solution. Regularization is the most common embedded method that minimizes the fitting error while reducing the coefficient (Tayal *et al.*, 2014). The underlying hypothesis is the linear relationship between predicted variables and variables, which may not exist in high dimensional datasets. Overall, the computational cost of embedded methods is lower than that of wrappers, but higher than that of filters (Guo *et al.*, 2019). In addition, the feature subset selected by embedded methods may not achieve good prediction results.

A systematic summary of existing FS methods shows that each type has advantages and disadvantages. Filters have a fast computational speed, but the accuracy of the feature subset obtained by them may be relatively low. Wrappers are less computationally efficient than filters, but can produce feature subsets with high classification performance. Therefore, filters and wrappers have a certain complementarity. This paper aims at fusing these two methods, and proposes an ANN integrating a two-stage FS method to quickly and efficiently obtain feature subsets with good predictive effect.

Methodology

Among these two contributions, the first belongs to the methodological innovation, which aims to propose an ANN model that integrates a two-stage feature selection method to address the “curse of dimensionality” problem. Therefore, we mainly present the proposed method in the methodology part. In this method, ANN is designed as a classifier, and the two-stage FS method is applied for feature selection. The two-stage FS method is innovative and is described in more detail. The second is mainly application innovation. Most scholars are devoted to exploring more accurate FDP models, but seldom pay attention to the interpretability of the models. SHAP and PDP are relatively mature methods widely applied to improve the interpretability of complex models. Therefore, they are not described in this section and are briefly introduced in the case application. To sum up, we mainly introduce the new method proposed in this paper in the methodology part.

The new two-stage method focuses on the fusion of different types of FS methods, integrating multiple filter and wrapper methods to quickly and efficiently obtain a feature subset with an outstanding prediction effect. For a large number of indicators in FDP, it first uses filters to remove irrelevant and redundant indicators. This greatly reduces the computational cost of the wrapper algorithm. In the filter stage, multi-criteria including information gain (IG), information gain ratio (IGR) and Spearman's rank correlation coefficient (SRCC) are applied to avoid the accidental situation caused by a single criterion. In the wrapper part, genetic algorithm (GA) is employed to search for feature subsets, and the accuracy of the predictor constructed by ANN is designed as the evaluation function. This can ensure the feature subset by this method has high prediction ability.

Multi-Criteria Filter Method

Filters are attracting more and more attention due to the emergence of high-dimensional data. This method applies filters in the first stage to preliminarily select features. The reasons are as follows. First, the filter criteria are often simple and easy to design and understand. Second, predictors are not considered, so their computation is small. Finally, filters can effectively delete some irrelevant indicators, improving the efficiency and accuracy of the subsequent wrapper model. Unlike other filter methods, multi-criteria, including IG, IGR and SRCC, are designed as the evaluation criteria to screen all features to avoid accidental situations caused by a single criterion.

(1)Information measurement

IG and IGR are measures of how much information a feature provides. Before mastering their definition, we need to clarify the concepts of entropy and conditional entropy. Entropy measures the uncertainty of a random variable. Let $Y = \{y_1, y_2, \cdots, y_i, \cdots, y_n\}$be a random variable, $p(y_i)$ is the probability when $Y = y_i$. $H(Y)$ is the information entropy of Y, and the calculation formula is given in Equation 1.

$$\mathrm{H(Y)} = -\sum_i p(y_i)\, log_2\, (p(y_i)) \quad (1)$$

Assume that $X = \{x_1, x_2, \cdots, x_i, \cdots, x_n\}$ is also a random variable. The conditional entropy $H(Y|X)$ reflects the uncertainty of the random variable Y when the value of the random variable X is known. The conditional entropy $H(Y|X)$ is defined as the mathematical expectation of conditional probability distribution entropy of Y under given condition X. $H(Y|X)$is calculated by Equation 2.

$$H(\mathrm{Y}|X) = \sum_{i=1}^{n}(x_i) H(Y|X = x_i) \quad (2)$$

In FS, IG is often used as an evaluation criterion to judge how much information a feature brings to a classification system (Jadhav *et al.*, 2018). The more information a feature adds, the more important it is. When a feature is deleted from the feature set, the total amount of information of the remaining feature set will change. The difference in the information amount between before and after the deletion is the amount of information this feature brings to the classification system (Reineking, 2016). $IG(Y, X)$ is calculated by Equation 3.

$$IG(Y, X) = H(Y) - H(Y|X) \quad (3)$$

However, there is a drawback in that IG is partial to features with more values. It is not appropriate to regard IG as the only criterion for information measurement. This paper applies IGR as an evaluation standard of feature selection. $IGR(Y, X)$ is calculated by Equation 4.

$$IGR(Y, X) = {IG(Y, X)}\Big/{SPI(X)} \quad (4)$$

where $SPI(X)$ is the splitting information, which can be calculated by Equation 5.

$$SPI(X) = -\sum_{i=1}^{n} \frac{x_i}{X} log_2 \frac{x_i}{X} \quad (5)$$

where $\frac{x_i}{X}$ is the probability that x_i occurs in X.

IGR is the result of information gain divided by splitting information (Dai and Xu, 2013). When the number of values contained in a feature is smaller, the splitting information is smaller. Therefore, when IGR is the criterion in filters, the features with fewer values are more easily selected. So, IG and IGR are taken as the information measurement criteria to screen the initial features in this study.

(2)Correlation coefficient

Correlation coefficient is a statistical index that reflects the close relationship between two features. Methods of measuring correlation coefficients include Pearson, Spearman and Kendall. SRCC is generally designed as a criterion when the data do not follow conventional distributions or data distribution is unknown. SRCC is required due to the uncertainty in the distribution of the indicator data collected in FDP. SRCC can be regarded as Pearson's correlation coefficient between two ranked variables. For n samples, all raw data are needed to convert to ranked data.

Let $R = \{R_1, R_2, \cdots, R_i, \cdots, R_n\}$ and $S = \{S_1, S_2, \cdots, S_i, \cdots, S_n\}$ denote the rank variables of X and Y, respectively, and SRCC is calculated by Equation 6.

$$\rho = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2 \sum_i (S_i - \bar{S})^2}} \quad (6)$$

where $\bar{R}$ and $\bar{S}$ are the means of the ranked variables X and Y, respectively. Through simple transformation, SRCC can be obtained by Equation 7. SRCC indicates the correlation direction of X (independent variable) and Y (dependent variable).

$$\rho = 1 - \frac{6\sum(R_i - S_i)^2}{n(n^2 - 1)} \quad (7)$$

IG and IGR are information measures, and SRCC is a dependency measure. The union of the feature subsets obtained by different criteria is taken as the result of the multi-criteria filter method. The multi-criteria complement each other, which can effectively prevent the randomness of a single criterion. In addition, the filters decrease number of features, which significantly reduces the computational complexity and improves the accuracy of the wrapper algorithm in the second stage.

The Wrapper Model Based on GA and ANN

Wrappers require a predefined predictor and apply its accuracy as the evaluation criterion for feature subsets. Thus, wrappers ensure that the selected feature subsets have good prediction performance. There are two tasks in a wrapper model. First, we need to set up a predictor in advance to calculate the accuracy of feature subsets. There are many methods available for FDP. However, it would take a lot of time to calculate the accuracy of each feature subset, and there are many feature subsets. Therefore, the predefined predictor should have a fast prediction speed. With a simple structure and fast speed, ANN can meet the enormous computational demands in a wrapper.

The second is to choose a search strategy. The exhaustive method can list all feature subsets. However, it is time-consuming and impractical to test the classification effect on all feature subsets (Lopez-Santillan *et al.*, 2020). So it is not appropriate to use an exhaustive method. Meta-heuristic search strategies are effective. It randomly generates several feature subsets, and then updates them with heuristics to gradually get the optimal feature subset. GA, a standard meta-heuristic method, is adopted in this paper. The main reason is that GA has many advantages in combinatorial optimization problems. First, binary is a common coding way in GA, which is suitable for combinatorial optimization and easy to understand. Besides, there is no need to recommend the optimization equation, the optimization efficiency is high, and it is easy to jump out of the local optimal solution in GA. And FS is a typical combinatorial optimization problem.

Therefore, the integration model based on GA and ANN is the main way in the second stage to obtain a feature subset with higher classification accuracy. In this method, GA is the search strategy, and the classification accuracy calculated by ANN is designed as the individual fitness function. The main contents are listed below.

(1) Coding

Coding is the transformation of solutions of practical problems to chromosomes of GA. The problem of FS can be viewed as a binary classification of features: retention or deletion. If the value of a certain position in a chromosome is 0, the corresponding feature is deleted; on the contrary, if the value is 1, the corresponding feature is retained. The length of the chromosomes is determined by the number of features, which is easy to understand and ensures the encoding efficiency.

(2) Fitness function

In evolutionary theory, individual fitness represents the ability to adapt to the environment and reproduce. The fitness function should be consistent with the goal of an actual problem, so the design of fitness functions is crucial to achieving the optimal solution. The accuracy of feature subsets is directly intended as the fitness function, which helps to find feature subsets with higher classification performance. ANN, with a simple structure and fast speed, is constructed as the classifier to calculate the accuracy of each individual in the evolution process.

(3) Selection

Selection is the process of choosing superior individuals from a population and eliminating inferior ones according to their fitness values. Then the selected individuals are taken as parents to produce children through crossover and mutation. In this paper, the roulette wheel method is adopted to select individuals. If the group size is n and the fitness value of the i-th individual is F_i, the probability of the i-th individual being selected is P_i, as shown in Equation 8.

$$P_i = \frac{F_i}{\sum_i^n F_i} \tag{8}$$

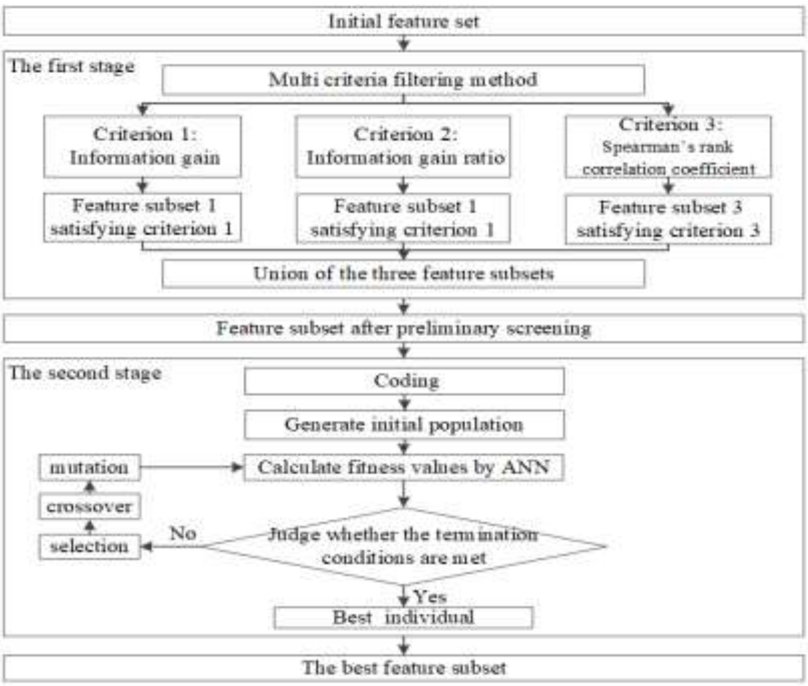


Figure 1. The Detailed Flow Chart of the Proposed Two-Stage FS Method

(4) Crossover

Crossover refers to the replacement and recombination of some genes of two selected parental individuals to produce a new individual. There are several crossover ways, including single-point, two-point, multi-point and average crossover (Kerschke *et al.,* 2019). Single-point crossover is employed in the proposed model. A crossover point is randomly chosen, and the genes behind the point on one chromosome are exchanged with those at the same position on the other chromosome. For example, Parent 1: 1 1 0 1 1 1 ˇ1 0 1

Parent 2: 1 0 1 1 0 1 ˇ0 1 1

Child 1: 1 1 0 1 1 1 0 1 1

Child 2: 1 0 1 1 0 1 1 0 1

(5) Mutation

Mutation preserves genetic diversity and prevents premature convergence. Mutation manipulation is the alteration of specific genes in a small number of individuals, which is the same as gene mutation in biology. Classically, the basic process of mutation manipulation is as follows. First, a mutation probability is set to determine whether each individual in the population has a mutation. If there is, a gene is randomly selected from the chromosome and the gene is changed from its original state. Such as,

Individual: 1 0 0ˇ1 0 0 1 0 1

New individual: 1 0 0 0 0 0 1 0 1

In summary, the flowchart of the two-stage FS model based on multi-criteria filter method and a wrapper is shown in Figure 1.

Case Analysis

This section tests the effectiveness and superiority of the method proposed in this paper. First, the proposed method is applied to feature selection and financial distress prediction to demonstrate its validity. Second, the two-stage FS method is compared with other classical FS methods to verify its superiority. Third, the effectiveness of the method proposed in this paper is validated in the manufacturing industry. Finally, to illustrate the importance of selected features in the constructed model, SHAP and PDP are introduced to enhance the interpretability and credibility of the "black box" model.

The 596 listed companies in China are studied, with their 2018 financial data and 2020 financial status as samples. Companies whose shares are subject to special treatment (ST) are considered to be in financial distress. A company is labeled as "ST" when one of the following two situations occurs. (1) Its net profits are negative for two consecutive years. (2) Its net assets in the latest fiscal year are less than registered capital. Obviously, listed companies labeled as ST are in financial distress. Here, 81 listed companies have been ST, and 515 listed companies haven't. Previous studies have shown that financial indicators are good at predicting financial distress. The financial data are collected through the Stock Financial Indicators (FININD) statement of the RESSET database. We collated all the indicators in the financial statements, deleted those with many missing values, and finally summarized 50 financial indicators, as shown in Table 1. The 50 indicators have the following characteristics. First, they can reflect the financial situation of listed companies well. Second, they are as comprehensive as possible, reflecting profitability, solvency, operating capacity, cash flow, and earnings per share. Finally, the data of the indicators can be collected.

Table 1

The Initial Feature Set

Indicators	Description	Indicators	Description
Incmope	Operating income	BasEPS	Basic earnings per share
Netinvincm	Net investment income	DilutEPS	Diluted earnings per share
Finexp	Financial expenses	EPS	Earnings per share (diluted)
OpePrf	Operating profit	ROE	Return on equity (diluted)
Totalprf	Total profit	ROECut	Return on equity (dilution cut)
IncTax	Income tax	ROEW	Return on equity (weighted)
Minprf	Minority profit	TotNcurlia	Total non-current liabilities
Netprf	Net profit	Nopeexp	Non-operating expenses
MoneFd	Monetary Funds	NPwiomin	Net profit (without minority profit and loss)
Accrecv	Accounts receivable	NRecProLos	Non-recurring profit and loss
Invtr	Inventory	NetprfCut	Net profit after cutting non-recurring profit and loss
Totcurass	Total current assets	NCFbyope	Net cash flow by operating
Intanass	Intangible assets	NCFfropePS	Net cash flow from operating per share
Totass	Total assets	NCFfrinv	Net cash flow from investment
Stloan	Short term loan	NCFfrfin	Net cash flow from financing
Taxpay	Tax payable	IncrinCCE	Net increase in cash and cash equivalents
Intpay	Interest payable	CCEatend	Cash and cash equivalents at the end of the period
Totcurlia	Total current liabilities	LTequinv	Long-term equity investment
Nopeincm	Non-operating income	Empsalpay	Employee salary payable
Totlia	Total liabilities	Shrcap	Paid in capital (or share capital)
Capsur	Capital surplus	SHEwioMin	Shareholders' equity (without minority interests)
Surres	Surplus reserve	MinSHE	Minority shareholders' equity
Retear	Retained earning	TotSHE	Total shareholders' equity
NAPS	Net assets per share	TotliaSHE	Total liabilities and shareholders' equity
Totshr	Total share capital	TotNcurass	Total non-current assets

The Two-Stage Feature Selection Method for FDP

(1) The results in the first stage

A multi-criteria filter method is established from the perspectives of IG, IGR and SRCC to avoid the accidental situation caused by a single criterion. The values of 50 features of 596 listed companies in China are collected. It has been verified that any feature does not conform to standard distribution functions, such as normal distribution, Poisson distribution, and exponential distribution. So, the piecewise method is applied to a discrete distribution, and the piecewise frequency is calculated as the probability of an interval value distribution. Regarding whether a company is in ST as variable Y, its information entropy can be calculated according to Equation 1.

$$H(y) = -\left(\frac{81}{596} \times log_2 \frac{81}{596} + \frac{515}{596} \times log_2 \frac{515}{596}\right) = 0.5734$$

IG and IGR are information measurements. Then, the IG values of 50 features are calculated according to Equation 3, as shown in Table 2. According to Equations 4 and 5, the IGR values of 50 features are computed by Python programming, as shown in Table 3.

SRCC is a standard for measuring the correlation between features. According to Equations 6 and 7, the SRCC values of 50 features to Y are calculated by Python programming, as shown in Table 4.

Table 2

The IG of 50 Financial Features

Features	IG	Features	IG	Features	IG	Features	IG	Features	IG
ROECut	0.2812	ROEW	0.2388	TotSHE	0.0486	Totass	0.0347	Intpay	0.0174
NetprfCut	0.2459	Retear	0.1735	NCFfrinv	0.0446	TotliaSHE	0.0347	NCFfrfin	0.0166
NPwiomin	0.2450	NAPS	0.1175	SHEwioMin	0.0440	Taxpay	0.0339	Accrecv	0.0158
BasEPS	0.2423	CCEatend	0.1028	MoneFd	0.0436	TotNcurass	0.0310	Shrcap	0.0149
DilutEPS	0.2423	NCFfropePS	0.0892	MinSHE	0.0423	NRecProLos	0.0288	Invtr	0.0143
EPS	0.2423	IncTax	0.0813	Incmope	0.0407	Totlia	0.0266	Totshr	0.0139
ROE	0.2414	Minprf	0.0694	Netinvincm	0.0386	Totcurlia	0.0248	Nopeincm	0.0128
OpePrf	0.2408	NCFbyope	0.0653	Empsalpay	0.0378	LTequinv	0.0233	Capsur	0.0106
Totalprf	0.2396	IncrinCCE	0.0582	TotNcurlia	0.0358	Intanass	0.0193	Nopeexp	0.0095
Netprf	0.2396	Surres	0.0511	Totcurass	0.0347	Finexp	0.0180	STloan	0.0094

Table 3

The IGR of 50 Financial Features

Features	IGR	Features	IGR	Features	IGR	Features	IGR	Features	IGR
ROECut	0.4793	NetprfCut	0.3442	IncrinCCE	0.0583	Totcurlia	0.0495	Empsalpay	0.0378
Totalprf	0.4084	NAPS	0.2613	NCFfrinv	0.0576	Netinvincm	0.0486	STloan	0.0364
NPwiomn	0.4029	Retear	0.2464	Taxpay	0.0561	TotNcurlia	0.0470	Invtr	0.0304
OpePrf	0.4016	Nopeexp	0.1636	Incmope	0.0544	TotNcurass	0.0464	LTequinv	0.0270
BasEPS	0.3957	IncTax	0.1587	TotSHE	0.0540	SHEwioMin	0.0440	Finexp	0.0259
DilutEPS	0.3957	CCEatend	0.1514	Surres	0.0531	NCFfrfin	0.0430	Capsur	0.0217
EPS	0.3957	NCFfropePS	0.1024	Totcurass	0.0496	MinSHE	0.0428	Shrcap	0.0193
Netprf	0.3886	MoneFd	0.0959	Totass	0.0496	Totlia	0.0426	Intanass	0.0193
ROE	0.3838	Minprf	0.0723	TotliaSHE	0.0496	Accrecv	0.0423	Intpay	0.0180
ROEW	0.3773	NCFbyope	0.0688	NRecProLos	0.0495	Totshr	0.0407	Nopeincm	0.0154

Table 4

The SRCC of 50 Financial Features (Absolute Value of the SRCC)

Features	SRCC	Features	SRCC	Features	SRCC	Features	SRCC	Features	SRCC
ROECut	0.5004	ROEW	0.4460	MoneFd	0.2672	Taxpay	0.1746	Totcurlia	0.0969
BasEPS	0.4925	Retear	0.4294	Surres	0.2611	TotNcurass	0.1658	Nopeexp	0.0878
DilutEPS	0.4915	NAPS	0.3864	NCFfrinv	0.2558	Totcurass	0.1598	Intpay	0.0833
EPS	0.4899	CCEatend	0.3709	IncrinCCE	0.2467	Intanass	0.1539	Finexp	0.0748
OpePrf	0.4872	NCFfropePS	0.3275	Incmope	0.2411	LTequinv	0.1373	NCFfrfin	0.0635
Totalprf	0.4862	IncTax	0.3270	MinSHE	0.2282	Invtr	0.1370	Accrecv	0.0537
Netprf	0.4857	NCFbyope	0.3192	Empsalpay	0.2208	Totlia	0.1141	Shrcap	0.0468
NPwiomin	0.4834	Minprf	0.3017	Netinvincm	0.2026	Nopeincm	0.1101	Totshr	0.0338
NetprfCut	0.4812	TotSHE	0.2867	Totass	0.1841	TotNcurlia	0.1077	Capsur	0.0268
ROE	0.4464	SHEwioMin	0.2815	TotliaSHE	0.1841	NRecProLos	0.1035	STloan	0.0043

Information measurements reflect the information amount each feature provides for judging financial status, and the correlation coefficient shows the degree of dependence between each feature and financial situation. The multi-criteria complement each other, which can effectively prevent the randomness of a single criterion. The results of the multi-criteria filter model are obtained by combining the information measurement results with the correlation coefficient measurement results. There are 35 features, so the feature subset C is {BasEPS, DilutEPS, EPS, ROE, ROECut, ROEW, Incmope, Netinvincm, OpePrf, Nopeexp, Totalprf, IncTax, NPwiomin, Minprf, Netprf, NRecProLos, NetprfCut, NCFbyope, NCFfropePS, NCFfrinv, IncrinCCE, CCEatend, MoneFd, Totcurass, Totass, Empsalpay, Taxpay, TotNcurlia, Surres, Retear, SHEwioMin, MinSHE, TotSHE, TotliaSHE, NAPS}.

(2) The results in the second stage

The first stage decreases the number of features from 50 to 35, significantly narrowing the search space in the subsequent wrapper. In the second stage, an integration method based on ANN and GA is designed as a wrapper to search for the optimal feature subset with high accuracy. A binary code is applied, and the number of features determines the length of chromosomes. The relevant parameters are designed as shown in Table 5.

Table 5

Related Parameter Designed in GA

Parameter	Setting value
Population size	10
Hybridization probability	0.85
Mutation probability	0.1
Chromosome length	35
Iterative Number	100

ANN is employed to calculate the prediction accuracy, which is designed as the fitness evaluation function in GA. In ANN models, the sigmoid function is set as the activation function. Set the number of hidden layers to 1, the learning rate to 0.1, and the momentum factor to 0.01. 300 samples are randomly selected from 596 as training samples to train the model, and the remaining 296 are test samples. The classification accuracy of the test samples is the fitness value of each individual in GA. After 100 iterations, the optimal feature subset A is obtained.

The optimal individual is [1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1]. So the optimal feature subset A is {BasEPS, DilutEPS, EPS, ROECut, Incmope, Netinvincm, Nopeexp, Totalprf, IncTax, NPwiomin, NRecProLos, NetprfCut, NCFbyope, NCFfropePS, IncrinCCE, Totass, Empsalpay, Taxpay, TotNcurlia, Surres, Retear, MinSHE, NAPS }.

The calculation process is shown in Figure 2. After 3.70 hours of 100 iterations, the optimal prediction accuracy is 96.2838 %. After filter and wrapper methods, the initial 50 features are screened to form a subset with 23 features. This section shows that the two-stage FS model proposed in this paper is available.

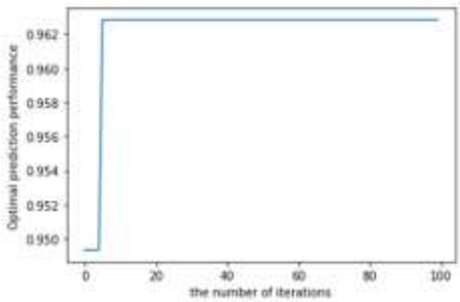


Figure 2. The Results of the Second Stage

Comparative Analysis of Multiple FS Methods

The main methodological innovation of this paper is the two-stage FS model. Therefore, a comparative analysis is performed to evaluate the predictive effect of the feature subsets obtained by the proposed method, wrapper and multi-criteria selection method. The three methods are employed for feature selection by taking the financial data of 596 Chinese listed companies as a sample set. Then, the corresponding feature subsets A, B and C are applied to FDP. By comparing their prediction performance, we can judge whether the method proposed in this paper is more effective. In the first stage, the multi-criteria filter method calculates the feature subset C. The feature subset A has been derived by the two-stage method. Therefore, we should build a wrapper model for the corresponding feature subset B.

(1) Feature selection based on wrapper

The initial feature set is selected directly by a wrapper without any prior filters. The basic idea of the wrapper model based on GA and ANN is consistent with that of the wrapper introduced in the previous section, and the results are shown in Figure 3.

In Figure 3, the optimal accuracy is 94. 9324%. After 100 iterations, the optimal individual is[1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1]. The optimal feature subset B is {BasEPS, DilutEPS, EPS, ROE, Netinvincm, Finexp, OpePrf, Nopeexp, IncTax, NPwiomin, NRecProLos, NCFbyope, NCFfrfin, MoneFd, Invtr, Intanass, Totass, STloan, Totcurlia, Capsur, SHEwioMin, MinSHE, TotSHE, TotliaSHE, NAPS, Totshr}.

The optimal accuracy of feature subset B is 94. 9324 %, lower than the classification accuracy of feature subset A. Furthermore, feature subset B contains 26 features, which is three more than feature subset A. Finally, the wrapper model based on GA and ANN takes 6.38 hours, almost twice as long as the two-stage method. Thus, compared with the feature subset B obtained by the wrapper, the feature subset A obtained by the two-stage method contains fewer features, has a higher classification accuracy, and takes less time.

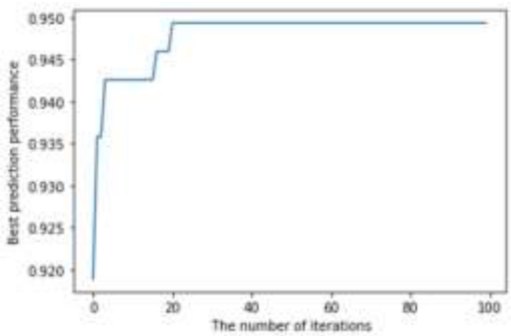


Figure 3. The Result of the Wrapper

(2)The comparison of prediction results

In order to verify the superiority of the two-stage FS approach, the feature subsets A, B, and C are employed for discriminant analysis. And their classification results are compared in this section. The sample set is the financial data of 596 listed companies in China. Among them, 81 listed companies are in ST, and 515 listed companies are normal. If all samples are for discriminant analysis, it makes little sense to compare the accuracy. If all companies are identified as normal, the accuracy is as high as 86.4 %, and the differences between different feature subsets are small. Since the number of samples needs to be greater than five times the number of features, at least 175 data samples are required. 119 enterprises are randomly selected from 515 normal enterprises, and 81 ST samples are added. A total of 200 companies are taken as the experimental data.

Discriminant analysis models are constructed using the feature subsets A, B, and C, respectively. Table 6 and Table

7 are the results of Fisher discriminant analysis. Table 6 shows the eigenvalues of the Fisher discriminant models. Table 7 gives the test results of the Fisher discriminant function. It can be seen that there are significant differences in the means of each group. Therefore, it is effective to build these Fisher discriminant models. Both Table 6 and Table 7 show that the three discriminant models are effective.

Table 6

Eigenvalue

Feature subset	*Eigenvalue*	*% of Variance*	*Cumulative %*	*Canonical correlation*
A	1.051a	100	100	0.716
B	1.040a	100	100	0.714
C	1.290a	100	100	0.750

a. First 1 canonical discriminant functions were used in the analysis.

Table 7

Wilks' Lambda

Feature subset	*Wilks' Lambda*	*Chi-square*	*df*	*Sig.*
A	0.488	134.698	21	0.000
B	0.490	132.968	23	0.000
C	0.437	152.002	29	0.000

Table 8

The Classification Result of Feature Subset A[b,c]

		Is it ST	*0*	*1*	*Total*
original	*count*	0	116	3	119
		1	14	67	81
	%	0	97.5	2.5	100.0
		1	17.3	82.7	100.0
Cross-validated	*count*	0	113	6	119
		1	17	64	81
	%	0	95.0	5.0	100.0
		1	21.0	79.0	100.0

a. Cross-validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b.91.5 % of original grouped cases correctly classified.
c.88.5 % of cross-validated grouped cases correctly classified.

Tables 8, 9 and 10 are the prediction results of these Fisher discriminant models constructed by the feature subsets A, B, and C, respectively, and the results are compared in Figure 4. In the original grouped cases, the accuracy of the model constructed by feature subset A is 91.5 %, then that by feature subset C is 89.5 %, and that by feature subset B is 86 %. In the cross-validated grouped cases, feature subset A has the highest accuracy of 88.5 %, followed by feature subset C with 85.0 % and feature subset B with the worst accuracy of 84 %. The prediction accuracy of the model constructed by feature subset A is higher than that of feature subsets B and C.

Table 9

The Classification Result of Feature Subset B[b,c]

		Is it ST	*0*	*1*	*Total*
original	*count*	0	110	9	119
		1	19	62	81
	%	0	92.4	7.6	100.0
		1	23.5	76.5	100.0
Cross-validated	*count*	0	107	12	119
		1	20	61	81
	%	0	89.9	10.1	100.0
		1	24.7	75.3	100.0

a. Cross-validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b.86.0 % of original grouped cases correctly classified.
c.84.0 % of cross-validated grouped cases correctly classified.

Table 10

The Classification Result of Feature Subset C[b,c]

		Is it ST	*0*	*1*	*Total*
original	*count*	0	117	2	119
		1	19	62	81
	%	0	98.3	1.7	100.0
		1	23.5	76.5	100.0
Cross-validated	*count*	0	111	8	119
		1	22	59	81
	%	0	93.3	6.7	100.0
		1	27.2	72.8	100.0

a. Cross-validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b.89.5 % of original grouped cases correctly classified.
c.85.0 % of cross-validated grouped cases correctly classified.

However, comparing accuracy alone for unbalanced class samples is not appropriate. The F1 score is adopted to measure the classification effect for a complete comparison. The F1 score is the harmonic mean of recall and precision. As noted in Fig.5, in the original grouped cases, F1 score of feature subset A is 0.89, that of feature subset C is 0.86, and that of feature subset B is 0.82. In the cross-validated grouped cases, F1 score of feature subset A is 0.85, then that of feature subset C is 0.80, and that of feature subset B is 0.79. On the whole, the two-stage FS method not only contains fewer features, but also has higher predictive performance. These further illustrate that the method proposed in this paper can effectively deal with the "curse of dimensionality", thereby improving the accuracy of FDP.

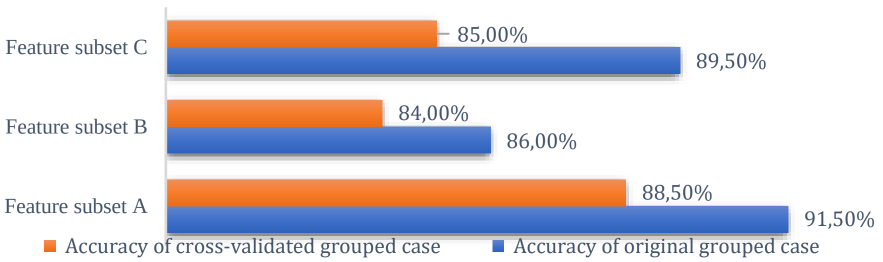


Figure 4. Comparison of Classification Accuracy of three Feature Subsets

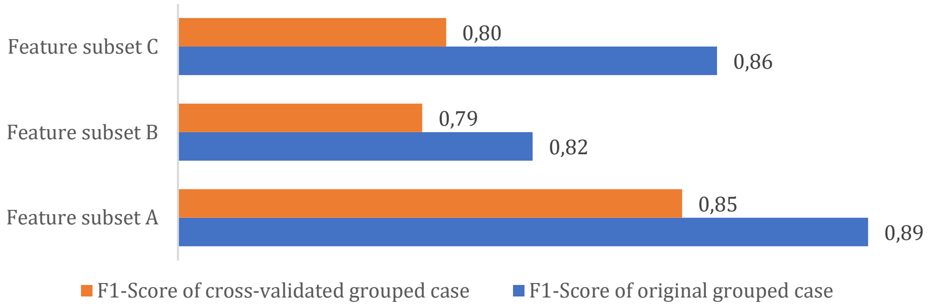


Figure 5. Comparison of F1 Score of three Feature Subsets

(3) The comparison of results in manufacturing

To examine the performance of the proposed method on the micro market, this paper chooses the manufacturing industry as an example. The main reason is that the manufacturing sector accounts for a relatively large share of the total market, and the number of samples in this sector is relatively large. Similarly, the discriminant analysis models are constructed using the feature subsets A, B, and C, respectively. The number of manufacturing enterprises in all samples is 311, of which 53 firms are in ST and 258 are normal. To ensure the validity of the models, 148 of the 258 normal samples are randomly selected and combined with the 53 distressed samples to form a test set of 200 samples.

Table 11

Eigenvalue

Feature subset	*Eigenvalue*	*% of Variance*	*Cumulative %*	*Canonical correlation*
A	0.860[a]	100	100	0.680
B	0.944[a]	100	100	0.697
C	1.069[a]	100	100	0.719

a. First 1 canonical discriminant functions were used in the analysis.

Table 12

Wilks' Lambda

Feature subset	*Wilks' Lambda*	*Chi-square*	*df*	*Sig.*
A	0.538	116.021	22	0.000
B	0.514	123.642	24	0.000
C	0.483	133.022	30	0.000

Based on the results of Fisher discriminant analysis of the manufacturing industry in Tables 11 and 12, it is clear that all three discriminant models are valid. Fig. 6 shows the accuracy results of the discriminant analysis models using the three feature subsets on manufacturing, both for original grouped cases and cross-validated grouped cases. For original grouped cases, the accuracy of the model constructed by feature subset A is 92.5 %, that by feature subset B is 91.5 %, and that by feature subset C is 90.5 %. For cross-validated grouped cases, the accuracy of feature subset A is 90 %, that of feature subset B is 87 %, and that of feature subset C is 86.5 %. It can be inferred that for a specific manufacturing industry, the feature subset A obtained by the method constructed in this paper can enhance the accuracy of FDP. Comparing the discriminant models on the test set with all the samples in the last subsection, the accuracy of all three discriminant models on the manufacturing samples has increased. Because there are only 53 distressed samples in the manufacturing sector, exacerbating the class imbalance, which inevitably leads to an increase in accuracy.

To avoid class imbalance leading to a biased refutation of accuracy metrics, and to ensure the reliability of comparisons on different sample sets, this subsection similarly calculates the F1 scores of the discriminant models constructed from the three feature subsets. As shown in Fig. 7, the F1 score of the discriminant model created by feature subset A is the highest in both the original grouped cases and the cross-validated grouped cases. However, the F1 scores in the specific market are lower than those in the overall market. This is because all these feature selection methods are data dependent, and the feature subsets A, B, and C are computed from all samples. In conclusion, the results on the test set of the manufacturing industry also prove the effectiveness of the method proposed in this paper

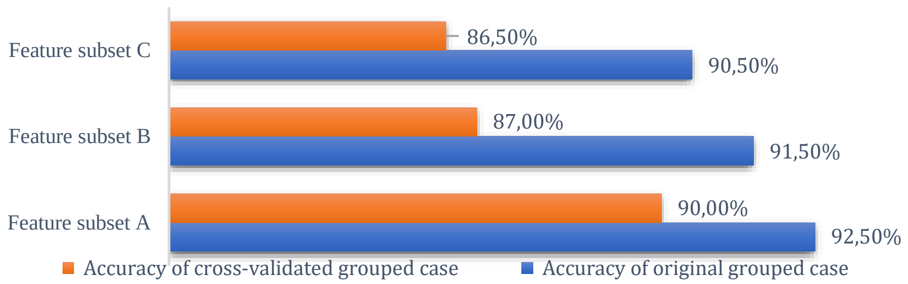


Figure 6. Comparison of Classification Accuracy of three Feature Subsets in Manufacturing

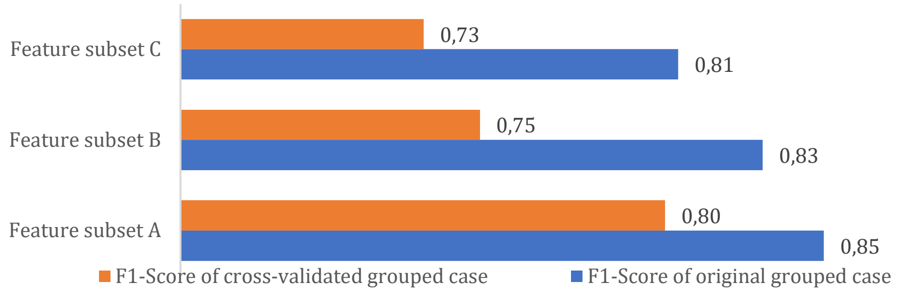


Figure 7. Comparison of F1 Score of three Feature Subsets in Manufacturing

Feature Importance

According to Friedman (2001), measuring feature importance can test the effect of each feature on FDP models. This paper introduces SHAP to quantify feature importance. Since SHAP was put forward by Lundberg & Lee in 2017, it has been frequently used to explain complex models. In SHAP, the Shapely values are defined as the average of the marginal contributions that can measure the influence of a feature. Shaply value can be calculated as Equation 9.

$$\phi_j(x) = \sum_{s\subseteq N\setminus\{j\}} \frac{|s|!(m-|s|-1)!}{m!}\left[f_x\left(s \cup \{x_j\}\right) - f_x(s)\right] \quad (9)$$

where $\phi_j(x)$ is the Shapley value of x_j, x_j represents a feature, N is the set of all input features, s is a feature subset, m is the number of features, and f_x corresponds to the prediction of the feature subset.

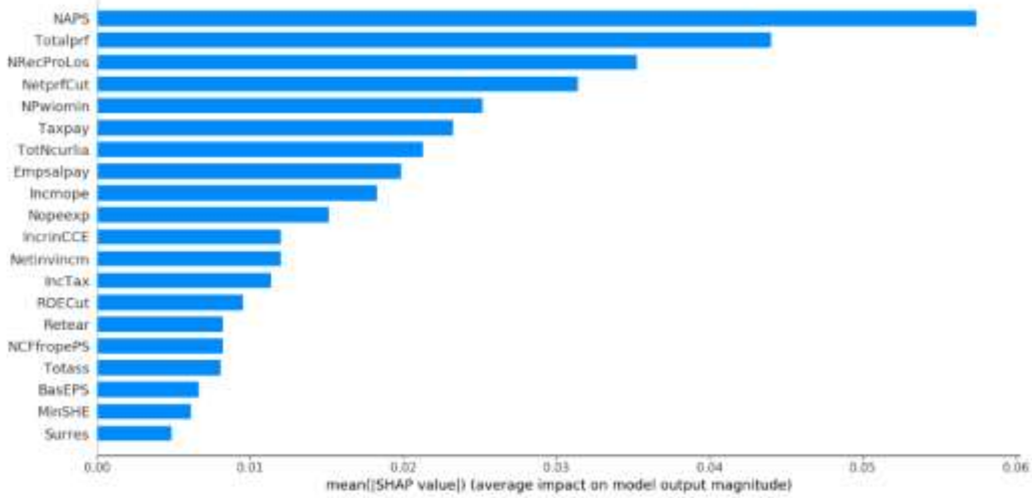
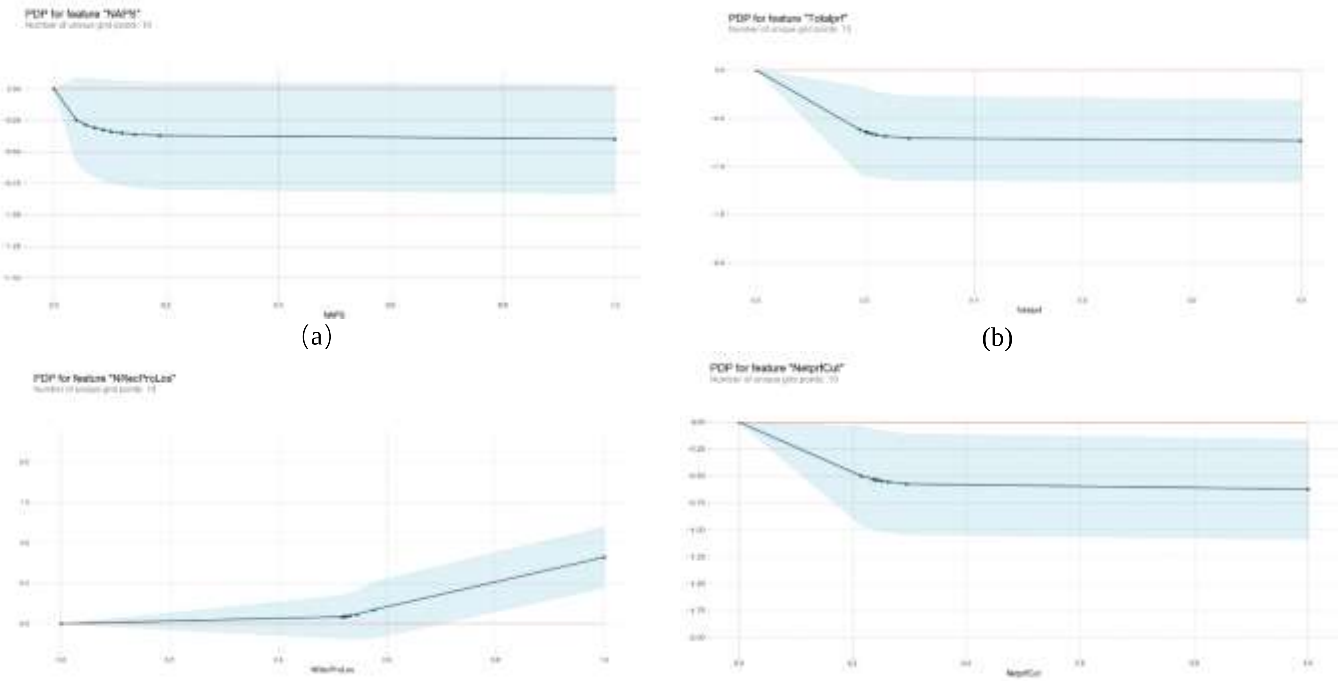


Figure 8. The Ranking of Feature Importance

Figure 8 shows the importance ranking of the selected features in the FDP model constructed in this paper. It can be concluded that these features including NAPS (net assets per share), Totalprf (total profit), NRecProLos (non-recurring profit and loss), NetprfCut (net profit after cutting non-recurring profit and loss), NPwiomin (net profit without minority profit and loss), Taxpay (tax payable), TotNcurlia (total non-current liabilities) are more important than other features, which is very similar to many previous research results. The analysis results of the feature importance of the FDP model constructed in this paper suggest that managers and financial institutions in China should pay more attention to the important features such as NAPS, Totalprf, NRecProLos and NetprfCut.

Furthermore, the PDPs of the most influential features (normalized value) are illustrated in Figure 9. There is a nonlinear relationship between the features and the probability of financial distress. The NAPS indicates the present value of the assets owned by each share. Generally speaking, the higher the NAPS, the better. Fig. 9(a) suggests that the probability of financial distress decreases to some extent as the NAPS increases. The Totalprf is a major indicator of the profitability of a company. The stronger the profitability of an enterprise, the smaller the probability of a financial crisis. Figure 9(b) displays an increase in Totalprf lowers the likelihood of bankruptcy. NRecProLos refers to the income and expenses that are not directly related to the main business, but impact the true and fair reflection of a company's normal profitability. Figure 9(c) exhibits the impact of NRecProLos on the prediction results. However, an increase in NRecProloss raises the probability of financial distress to some extent. Like Totalprf, NPwiomin is also a crucial indicator of profitability. It can be concluded from Figure 9(d) that an increase in NetprfCut reduces bankruptcy probability. In short, the results shown in the PDPs are in line with the prior knowledge of financial managers, which increases the credibility of the "black box" model.



(a) (b) (c) (d)

Figure 9. The PDPs of Some Important Features

Conclusion and Discussion

The current turbulent economic environment is exacerbating the occurrence of financial distress among companies. Corporate managers, financial institutions and investors urgently need accurate and credible FDP models. While machine learning models improve accuracy to some extent, other problems in the forecasting process can affect the accuracy of these models and their use by stakeholders. First, the “curse of dimensionality” not only increases the computational burden, but also leads to overfitting of prediction models and reduces the accuracy of predictions. In addition, the interpretability of FDP models is critical for users. Most complex machine models have the “black box” problem, which makes users very cautious when using these models. Therefore, the main objective of this paper is to address the “curse of dimensionality” and “black box” issues in FDP to improve the accuracy and credibility of the predictive model. First, this paper proposes an ANN model integrating a two-stage FS method to efficiently and quickly achieve a feature subset with a good prediction effect. Furthermore, SHAP and PDP are employed to improve the interpretability and credibility of the FDP model constructed in this paper.

In this method, ANN is designed as a classifier, and a two-stage method is applied for feature selection. The two-stage FS method aims to overcome the shortcomings of the existing FS methods, efficiently solve the “curse of dimensionality” and improve the prediction accuracy. In the first stage, multi-criteria such as IG, IGR and SRCC are adopted to screen features to avoid accidental faults caused by a single criterion. Some irrelevant indicators are deleted, and the number of features is reduced from 50 to 35, which lessens the subsequent computations and the possibility of model overfitting. In the second stage, an integration method based on ANN and GA is designed as a wrapper to ensure the accuracy of the selected feature subset. ANN models are constructed to calculate the accuracy of FDP as the evaluation standard of the chromosomes in GA. Ultimately, the initial 50 features are screened to form a subset with 23 features, and the accuracy of the prediction model is 96.2838 %. The two-stage feature selection method proposed in this paper can achieve a feature subset with good prediction effects in a relatively short time, effectively balancing the predictive effect and time cost.

In addition, we compare the two-stage method constructed in this paper with other classical FS methods and judge which method obtains the feature subset with the best prediction effect. Feature subsets A, B and C are achieved by the two-stage FS, wrapper and multi-criteria filter methods, respectively. The three feature subsets are applied to construct discriminant models for FDP, respectively. Feature subset A has the highest accuracy of 91.5 % in the original grouped cases and 88.5 % in the cross-validated grouped cases. In addition, the F1 score of feature subset A is 0.89 in the original grouped cases and 0.85 in the cross-validated grouped cases, which are the highest. Therefore, the results show that the method proposed in this paper is more effective. Furthermore, the proposed method can effectively deal with the “curse of dimensionality”, thereby improving the accuracy of FDP.

Finally, to examine the importance of the selected features in the proposed model, SHAP and PDP are introduced to enhance the interpretability and credibility of the “black box” model. SHAP is to quantify the importance of features. The analysis results suggest that managers and financial institutions in China should focus more on essential features such as NAPS, Totalprf, NRecProLos and NetprfCut. Moreover, PDPs reveal the nonlinear relationship between the features and the chance of financial distress. The results display that an increase in NAPS, Totalprf and NPwiomin decreases the probability of financial distress, while an increase in NRecProloss increases its probability. These are consistent with the prior knowledge of experts, which increases the credibility of the proposed model.

Overall, the model constructed in this paper efficiently addresses the "curse of dimensionality" and enhances the prediction accuracy. The proposed method allows managers to detect an impending deterioration of the financial situation at an early stage, alerting them to take risk management measures to steer clear of bankruptcy. In addition, it can help banks prevent the growth of non-performing loans by identifying risky customers more accurately, and reduce the loss of quality customers due to the misidentification of normal companies. Finally, it can also contribute to investors identifying companies that will be in financial distress in the future, thus mitigating financial losses. In addition, SHAP and PDP are applied to learn the importance of the selected features and the relationships between these features and the prediction results in the proposed model. Their results align with experts' prior knowledge, improving the prediction model's reliability. These make stakeholders more confident when using this "black box" model.

However, the two-stage FS method has its limitations. First, the wrapper in the second stage may yield different results. The main reason is the existence of random numbers in GA and ANN. In GA, the individuals in the initial population are randomly generated. The initial coefficients of ANN are also set randomly. Second, the number of companies in the empirical analysis is relatively small, which may affect the accuracy and credibility of the model to some extent. Third, the samples we use are audited financial data, which may have a negative impact on the stability of the model. In addition, like most research in this area, this paper only focuses on financial data in predicting the financial distress of firms and does not consider industry and regional factors. Finally, the cases analyzed are Chinese listed companies, and the constructed model is data-dependent, making it difficult to validate it in other countries and regions.

Acknowledgements

This paper was funded by the National Natural Science Foundation of China (71771066 and 72131005).

References

Ariza-Garzon, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M. J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. *IEEE Access*, 8, 64873–64890. http://doi.org/10.1109/ACCESS.2 020.2984412

Beaver, W. H., McNichols, M. F., & Rhie, J. W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies*, 10(1), 93–122. http://doi.org/ 10.1007/s11142-004-6341-9

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A review of bankruptcy prediction studies : 1930-present. *Journal of Financial Education*, 33(Winter), 1–42.

Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using joint mutual information maximisation. *Expert Systems with Application*, 42(22), 8520–8532. http://doi.org/10.1016/j.eswa.2015.07.007

Ben Jabeur, S., Sadaaoui, A., Sghaier, A., & Aloui, R. (2020). Machine learning models and cost-sensitive decision trees for bond rating prediction. *Journal of the Operational Research Society*, 71(8), 1161–1179. http://doi.org/10.1080/01 605682.2019.1581405

Ben Jabeur, S., Gharib, C., Mefteh-Wali, S., & Ben Arfi, W. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, 166, 120658. http://doi.org/10.1016/ j.techfore.2021.120658

Dai, J. H., & Xu, Q. (2013). Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Applied Soft Computing*, 13(1), 211–221. http://doi.org/10.1016/j.asoc.2012.07.029

Deng, X. L., Li, Y. Q., Weng, J., & Zhang, J. L. (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78(3), 3797–3816. http://doi.org/ 10.1007/s11042-018-6083-5

Dimitras, A.I., Zanakis, S.H., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, 90 (3), 487–513. https://doi.org/10.1016/0377-2217(95)00070-4

Du, X. D., Li, W., Ruan, S. M., & Li, L. (2020). CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection. *Applied Soft Computing Journal*, 97(A), 106758. http://doi.org/10.1016/j.asoc.202 0.106758.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. http://doi.org/10.1214/aos/1013203451

Ghosh, M., Guha, R., Sarkar, R., & Abraham, A. (2020). A wrapper-filter feature selection technique based on ant colony optimization. *Neural Computing & Applications*, 32(12), 7839–7857. http://doi.org/ 10.1007/s00521-019-04171-3

Guo, Y. M., Chung, F. L., Li, G. Z., & Zhang, L. (2019). Multi-label bioinformatics data classification with ensemble embedded feature selection. *IEEE Access*, 7, 103863–103875. http://doi.org/10.1109/ACCESS.2019.2931035.

Halim, Z., Shuhidan, S. M., & Sanusi, Z. M. (2021). Corporation financial distress prediction with deep learning: analysis of public listed companies in Malaysia. *Business Process Management Journal*, 27(4), 1163–1178. http://doi.org/10. 1108/ BPMJ-06-2020-0273

Hancer, E., Xue, B., Karaboga, D., & Zhang, M. J. (2015). A binary ABC algorithm based on advanced similarity scheme for feature selection. *Applied Soft Computing*, 36, 334–348. http://doi.org/10.1016/j.asoc.2015.07.023

Jadhav, S., He, H. M., & Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing*, 69, 541–553. http://doi.org/10.1016/j.asoc.2018.04.033

Jeong, Y. S., Shin, K. S., & Jeong, M. K. (2015). An evolutionary algorithm with the partial sequential forward floating search mutation for large-scale feature selection problems. *Journal of the Operational Research Society*, 66(4), 529–538. http://doi.org/10.1057/jors.2013.72

Jiang, K., Tang, J. X., Wang, Y. L., Qiu, C. Y., Zhang, Y. P., & Lin, C. (2020). EEG feature selection via stacked deep embedded regression with joint sparsity. *Frontiers in Neuroscience*, 14, 829. http://doi.org/10.3389/fnins.2020.00829.

Karas, M., & Reznakova, M. (2017). Predicting the bankruptcy of construction companies: A cart-based model. *Inzinerine Ekonomika-Engineering Economics*, 28(2), 145–154. http://doi.org/10.5755/j01.ee.28.2.16353

Kerschke, P., Hoos, H. H., Neumann, F., & Trautmann, H. (2019). Automated algorithm selection: survey and perspectives, *Evolutionary Computation*, 27(1), 3–45. http://doi.org/10.1162/evco_a_00242

Kim, S., Mun, B. M., & Bae, S. J. (2018). Data depth based support vector machines for predicting corporate bankruptcy. *Applied Intelligence*, 48(3), 791–804. http://doi.org/10.1007/s10489-017-1011-3

Liu, J. M., Wu, C., & Li, Y. L. (2019). Improving financial distress prediction using financial network-based information and GA-based gradient boosting method. *Computational Economics*, 53(2), 851–872. http://doi.org/10.1007/s10614-017-9768-3

Liu, Y. Q., Mu, Y., Chen, K. Y., Li, Y. M., & Guo, J. H. (2020). Daily activity feature selection in smart homes based on pearson correlation coefficient. *Neural Process Letters*, 51(2), 1771–1787. http://doi.org/10.1007/s11063-019-10185-8

Lopez-Santillan, R., Montes-Y-Gomez, M., Gonzalez-Gurrola, L. C., Ramirez-Alonso, G., & Prieto-Ordaz O. (2020). Richer document embeddings for author profiling tasks based on a heuristic search. *Information Processing and Management*, 57(4), 102227. http://doi.org/10.1016/j.ipm.2020.102227

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

Mafarja, M. M., & Mirjalili, S. (2017). Hybrid whale optimization algorithm with simulated annealing for feature selection. *Neurocomputing*, 260, 302–312. http://doi.org/10.1016/j.neucom.2017.04.053.

Malakauskas, A., & Lakstutiene, A. (2021). Financial distress prediction for small and medium enterprises using machine learning techniques. *Inzinering Ekonomika-Engineering Economics*, 32(1), 4–14. http://doi.org/10.5755/j01.ee.32.1. 27382.

Mehanovic, D., Keco, D., Kevric, J., Jukic, S., Miljkovic, A., & Masetic, Z. (2021). Feature selection using cloud-based parallel genetic algorithm for intrusion detection data classification. *Neural Computing & Applications*, 33(18), 11861–11873. http://doi.org/ 10.1007/s00521-021-05871-5

Mensah, Y. M. (1984). An examination of the stationarity of multivariate bankruptcy prediction models: a methodological study. *Journal of Accounting Research*, 22(1), 380–395. https://doi.org/10.2307/2490719

Nouri-Moghaddam, B., Ghazanfari, M., & Fathian, M. (2021). A novel multi-objective forest optimization algorithm for wrapper feature selection. *Expert Systems with Applications*, 175, 114737. http://doi.org/ 10.1016/j.eswa.2021. 114737

Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*, 174, 114765. http://doi.org/ 10.1016/j.eswa.2021. 114765

Reineking, T. (2016). Active classification using belief functions and information gain maximization. *International Journal of Approximate Reasoning*, 72, 43–54. http://doi.org/10.1016/j.ijar.2015.12.005

Sayed, G. I., Hassanien, A. E., & Azar, A. T. (2019). Feature selection via a novel chaotic crow search algorithm. *Neural Computing and Applications*, 31(1), 171–188. http://doi.org/ 10.1007/s00521-017-2988-6

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146, 102551. http://doi.org/10.1016/j.ijhcs.2020. 102551.

Solorio-Fernandez, S., Martinez-Trinidad, J. F. & Carrasco-Ochoa, J. A. (2020). A supervised filter feature selection method for mixed data based on spectral feature selection and information-theory redundancy analysis. *Pattern Recognition Letters*, 138, 321–328. http://doi.org/10.1016/j.patrec.2020.07.039

Sun, J., Li, H., Fujita, H., Fu, B. B. & Ai, W. G. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Information Fusion*, 54, 128–144. http: //doi. org/10.1016/j.inffus.2019.07.006

Tayal, A., Coleman, T. F., & Li, Y. Y. (2014). Primal explicit max margin feature selection for nonlinear support vector machines. *Pattern Recognition*, 47(6), 2153–2164. http://doi.org/10.1016/j.patcog.2014.01.003

Tsai, C. F., Sue, K. L., Hu, Y. H. & Chiu, A. (2021). Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. *Journal of Business Research*, 130, 200–209. http://doi.org/10.1016/j.jbusres.2021.03.018.

Ul Haq, A., Li, J. P., Memon, M. H., Khan, J., Ali, Z., Abbas, M., & Nazir, S. (2020). Recognition of the Parkinson's disease using a hybrid feature selection approach. *Journal of intelligent and fuzzy system*, 39(1), 1319–1339. http://doi.org/10.3233/JIFS-200075.

Yan, C. K., Ma, J. J., Luo, H. M., &Wang, J. X. (2018). A hybrid algorithm based on binary chemical reaction optimization and Tabu search for feature selection of High-Dimensional Biomedical Data. *Tsinghua Science and Technology*, 23(6), 733–743. http://doi.org/ 10.26599/TST.2018.9010101

Zhu, Y. Q., Peng, Q. L., Lin, Y. X., Zou, L., Shen, P. P., Chen, F. F., Min, M., Shen, L., Chen, J. J., & Shen, B.R. (2020). Identification of biomarker microRNAs for predicting the response colorectal cancer to neoadjuvant chemo radiotherapy based on microRNA regulatory network. *Oncotarget*, 8(2), 2233–2248. http://doi.org/10.18632/oncotarget.13659.

Authors' Biographies

Zijiao Zhang is pursuing her PhD at the School of Management, Harbin Institute of Technology, China. Her research interests are in financial distress prediction, feature selection, and optimization algorithms. The recent reseach results are published in journals Information Processing and Management, http://doi.org/10.1016/j.ipm.2022.102988 and Mathematical Biosciences and Engineering, http://doi.org/10.3934/mbe.2023924.

Shiyou Qu is a professor at the School of Management, Harbin Institute of Technology, China. He received his Ph.D. in Technical Economics and Management from Harbin Institute of Technology in 2003. He has published more than 20 peer-reviewed papers. His current research interests are in corporate finance and technological innovation theory.

Chong Wu is a professor at the School of Management, Harbin Institute of Technology, China. He received his Ph.D. in Fundamental Mathematics from Harbin Institute of Technology in 1998. He has published more than 70 peer-reviewed papers, including IEEE Transactions on Knowledge and Data Engineering, Journal of the Association for Information Science and Technology, Information Sciences, Knowledge-based Systems and Expert Systems with Applications. His current research interests are in prediction theory and applications.

The article has been reviewed.

Received in March 2022; accepted in March 2024.