Predicting Accounting Fraud in Publicly Traded Chinese Firms via A PCA-RF Method

Donger Chen

Department of Accounting, Nanjing Audit University, Nanjing, Jiangsu, China, 211815

donger.chen@outlook.com

Abstract. Financial fraud occurs from time to time and gradually becomes a worldwide problem with the expansion of the international capital markets and the rise of the information industry economy under the Internet eco-system. This paper provides a methodology for predicting financial fraud using basic financial data. The methodology is based on PCA-RF. Different from traditional methods such as logistic regression and support vector machine, we creatively proposed a PCA-RF model, first using principal component analysis to reduce the dimensionality of the data, then using grid search to optimize the random forest model, and finally directly selecting the raw financial data from the financial statements for direct analysis. We compare the analysis results with random forest and neural network methods, and the study finds that the PCA-RF model is superior for predicting domestic financial fraud in China. In this paper, we use an ensemble learning approach to introduce the PCA-RF method into the field of prediction of financial fraud for listed companies.

Keywords: fraud prediction; machine learning; Principal component analysis; random forest; combined model

1 Introduction

Financial fraud occurs from time to time and gradually becomes a worldwide problem with the expansion of the international capital markets and the rise of the information industry economy under the Internet eco-system. The internationally famous Enron case, the WorldCom fraud case and the recent financial fraud cases of Kangmei Pharmaceuticals and Zhangzidao Island in China have been exposed as concealing or delaying the disclosure of information, falsifying financial data and cheating investors, causing huge losses to stakeholders. In recent years, China has paid great attention to the quality of accounting information, internal control and auditing procedures, etc [1]. In 2019, "the new Securities Law" of the People's Republic of China was amended to show "zero tolerance" for financial fraud. In January 2022, the Supreme People's Court issued "Several Provisions on the Trial of Civil Compensation Cases for False Statements in the Securities Market" to further solidify the provisions on legal liability for financial fraud activities. In March 2022, the Ministry of Finance and In March 2022, the Ministry of Finance and the Securities and Futures Commis-

H. Wu et al. (Eds.): CSIEDE 2022, ACSR 103, pp. 739–748, 2022.
https://doi.org/10.2991/978-94-6463-108-1_82

sion (SFC) issued "the Notice on Further Improving the Effectiveness of Internal Control over Financial Reporting of Listed Companies," which requires strengthening the assessment and control of financial fraud risks [2]. The SEC issued "the Notice on Further Improving the Effectiveness of Internal Control Over Financial Reporting of Listed Companies," which requires strengthening the assessment and control of financial fraud risks. Unfortunately, many financial frauds are hard to detect. Financial statements make it difficult to identify related party transactions, illegal stock trading, and other activities. The systematic fraud of managers makes the financial statement data lose its "early warning" function, and it is difficult to avoid the fraudulent "trap" of managers simply by analyzing financial indicators.

With the development of big data and artificial intelligence technology, artificial intelligence methods represented by big data and machine learning methods in the study of financial fraud have gradually received academic attention [3]. Kotsiantis et al. used a decision tree model to predict financial fraud and found six financial indicators with high correlation [4]. Liu, Jun, and Wang, Liping constructed a neural network model for financial reporting fraud identification, which used financial indicators and equity structure indicators as feature variables and achieved a correct judgment rate of 80%, significantly better than the linear model [5]. Cecchini et al. used raw financial data directly and predicted financial fraud using a support vector machine model, and the accuracy of their prediction results was better than the prediction model based on financial indicators [6]. Ravisankar et al. compared the effectiveness of multilayer feedforward neural network, support vector machine, genetic algorithm, logistic regression, and probabilistic neural network models for predicting financial fraud [6]. Shuangjie Li and Xingxing Chen (2013) constructed a profit manipulation identification model for listed companies using a BP neural network approach and improved the overall identification rate by reducing the second type of errors in the model through a data envelopment analysis (DEA) approach [7]. BAO et al. directly used 28 raw data from financial statements and used an ensemble learning machine learning approach [8].

In summary, statistical models based on big data and machine learning are gradually being used more often in financial fraud research, which is different from traditional logistic regression models and has certain advantages. In this paper, it provides a methodology for predicting financial fraud using basic financial data. The methodology is based on PCA-RF. Different from traditional methods such as logistic regression and support vector machine, the author creatively proposed a PCA-RF model, first using principal component analysis to reduce the dimensionality of the data, then using grid search to optimize the random forest model, and finally directly selecting the raw financial data from the financial statements for direct analysis. It hopes to provide a certain insightful suggestion for this field.

2 Method

The PCA-RF model is a method of ensemble learning in machine learning. The combined PCA-RF model is a combined principal component analysis (PCA) and random forest (RF) model [9]. First, principal component analysis is used to transform the

original correlated variables into a set of linearly uncorrelated variables, called principal components. Then, the RF model is built based on the principal components. Consider a data matrix, X = $\{\chi_1, \chi_2, \ldots, \chi_n\}$ with a column-by-column zero empirical mean, i.e., the sample mean of each column has been shifted to zero.

The basic steps of PCA are as follows.

- Calculate the covariance matrix of X, Cov $= \sum_{i=1}^{n} \chi_i \chi_i^T$.
- Calculate the eigenvalues and the corresponding eigenvectors of the covariance matrix.
- The eigenvectors of the first k eigenvalues form the matrix W = $\{\omega_1, \omega_2, \ldots, \omega_n\}$. In this study, the eigenvectors with eigenvalues greater than 1 are taken.
- Therefore, the principal component solution of X can be expressed as T=XW.

Random forests are ensemble learning methods for classification, regression, and other tasks that output mean values for classification or prediction by constructing multiple decision trees at training time.

The process of constructing a random forest consists of the following 3 main steps.

- Sampling in the data set and generating a training set for each decision tree.
- generating decision trees for each training set, the process of generating decision trees does not require pruning process. The decision tree generation algorithm consists of 2 main parts: node splitting and selection of random feature variables.
- Generating a random forest.

AUC value is the area under the ROC curve, and the larger the AUC is, the better the performance of the model is. In this paper, we use AUC as a measure for grid search to optimize the random forest model.

3 Data and sample formation

3.1 Sample

The sample data in this paper is obtained from the financial statement database of Chinese listed companies in the CSMAR database, and the financial data of a total of 3,325 A-share listed companies from FY2017 to FY2022 were selected, and the sample covers a wide range of industries [10], including manufacturing, construction, agriculture, and retail. In this paper, we propose to classify listed companies with irregular data such as "fictitious profit", "false listing of assets" and "false record (misleading statement)" in the documents issued by the regulator; listed companies with ST treatment announced by the SFC, and listed companies with "negative opinion" or "disclaimer of opinion" in the annual audit report as companies with financial fraud. We classify companies with a standard unqualified opinion in the annual audit report, which are not treated as ST, do not have "fictitious profit", "misrepresented assets", "false statements", etc. and have complete financial statements as companies without financial fraud.

3.2 Data sets

When dealing with common financial problems, models based on financial ratios may be more powerful than those using raw data directly because the financial ratios determined by experts combine financial theory and, at the same time, combine the ratios of business efficiency and financial position through an intrinsic linkage, forming a complete system. However, there is less research on financial fraud, and the use of financial ratios for fraud prediction may lose its validity. Besides, converting raw accounting data into a limited number of financial ratios could mean the loss of useful predictive information (e.g.,Bao Y and Ke B [2020]) [1]. Therefore, the paper uses the raw financial statement data directly for fraud prediction.

The list of raw financial data items is selected based on Cecchini et al., Dechow et al., and BAO et al. [11]. This paper combines 28 raw financial data variables selected from the BAO et al. paper, adjusted for Chinese reporting accounts, to obtain 13 financial variables. We use the presence of fraud as the dependent variable.

The variables also need to be preprocessed before modeling. The 14 variables included in the model are divided into two types: continuous variables and categorical variables. The categorical variable is fraud, and its value is expressed as an integer, and the variable "fraud" is denoted as 1 for the presence of financial fraud and 0 for the absence of financial fraud. The rest of the financial statement data is continuous variables. In order to improve the convergence speed and accuracy of the model, the continuous variables are normalized using the Min-Max normalization method before the model is built. The variables and their types are shown in Table 1.

Min-max normalization method: $\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$

In the formula: $\hat{x}$ is the normalized data, x is the original data, x_{max} is the maximum value of the variable, x_{min} is the minimum value of the variable.

Table 1. List of Variables

Variable	Type
X1=Owner's equity, total	continuous variable
X2=Price close, annual, fiscal	continuous variable
X3=Inventories, total	continuous variable
X4=Property, plant and equipment, total	continuous variable
X5=Current liabilities, total	continuous variable
X6=Investment and advances, other	continuous variable
X7=Sales/turnover (net)	continuous variable
X8=Issue price	continuous variable
X9=Accounts payable	continuous variable
X10=Liabilities, total	continuous variable
X11=Account receivable, total	continuous variable
X12=Assets, total	continuous variable
X13=Cost of goods sold	continuous variable
X14=fraud	classified variables: yes(1)/no(0)

The random forest model can evaluate the importance of features, i.e., compare the magnitude of contribution between individual features by looking at how much contribution each feature makes to each tree in the random forest. Before building the combined PCA-RF model, we established a random forest model on the original data to analyze the importance of each variable. The results of variable importance analysis are shown in Figure 1.

According to the results, "Owner's equity, total", "Price close, annual, fiscal", "Inventories, total", and "Property, plant, and equipment, total" have a greater impact on whether financial fraud occurs. When making financial fraud predictions for listed companies, it is important to pay attention to the raw data in the financial statements and be able to prioritize when working on fraud prediction.

```
Feature ranking:
 1)X1                                0.161402
 2)X2                                0.092415
 3)X4                                0.092162
 4)X3                                0.080885
 5)X6                                0.069453
 6)X5                                0.068528
 7)X12                               0.067609
 8)X7                                0.065566
 9)X8                                0.064082
10)X10                               0.063171
11)X11                               0.059886
12)X9                                0.058719
13)X13                               0.056122
```

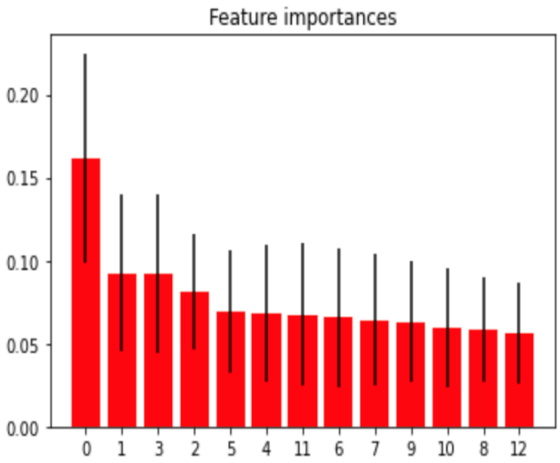

Fig. 1. Feature importances after Random Forest

4 Analysis Method

Because 13 variables are selected for prediction analysis in this paper, the dimensionality of the input variables is relatively large, and overfitting is likely to occur [12]. The basic steps of the PCA-RF combined model algorithm are as follows.

- The data set is randomly divided into a training set and a test set. 70% of the events (1477 events) are randomly selected as the training set and the remaining 30% (634 events) are randomly selected as the test set.
- Principal component analysis is performed. First, the equation of each principal component is established. Then, the data is substituted into the principal component equation and the principal component corresponding to each incident is calculated. The principal components with eigenvalues greater than 1 and a cumulative contribution of about 85% are selected as the new data set.
- RF model parameters are adjusted by grid search method to optimize the random forest model.
- RF model is established by using a training set.
- Use the test set to make predictions and get the prediction results.
- Analysis and evaluation of the results. The combined PCA-RF model was compared and analyzed with the random forest model and neural network model.

The training flow chart of the PCA-RF prediction model proposed in this paper is shown in Figure 2

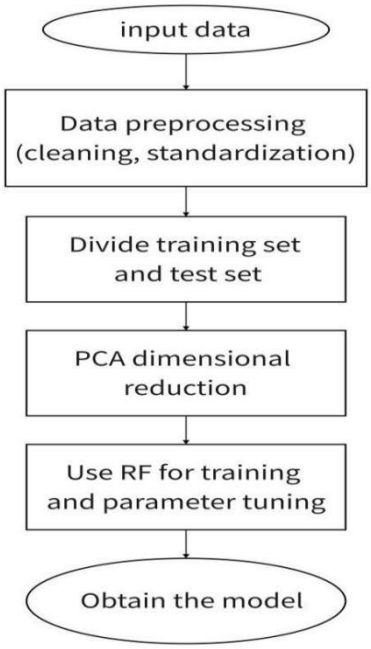

Fig. 2. .Flow chart of PCA-RF model algorithm

5 Results analysis

5.1 Results of PCA

Through principal component analysis, three principal components with eigenvalues greater than 1 were obtained. The principal component matrix is shown in the table. After the principal component analysis, the principal components with eigenvalues greater than 1 and with a cumulative contribution of about 85% are selected, which simplifies the 13-dimensional data into 3-dimensional data and improves the efficiency of the model. Table 2 and Table 3 report the results of PCA.

Table 2. Principal component eigenvalues and variance contribution rates

Principal components/correlation	Number of obs	=	2,111
	Number of comp.	=	3
	Trace	=	13
Rotation:(unrotated=principal)	Rho	=	0.8403

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	7.60934	5.76817	0.5853	0.5853
Comp2	1.84117	.368268	0.1416	0.7270
Comp3	1.4729	.665095	0.1133	0.8403
Comp4	.807803	.265576	0.0621	0.9024
Comp5	.542228	.115954	0.0417	0.9441
Comp6	.426273	.311986	0.0328	0.9769
Comp7	.114287	.0269501	0.0088	0.9857
Comp8	.0873368	.0275147	0.0067	0.9924
Comp9	.059822	.0338302	0.0046	0.9970
Comp10	.0259919	.016021	0.0020	0.9990
Comp11	.0099708	.00708814	0.0008	0.9998
Comp12	.00288267	.00288245	0.0002	1.0000
Comp13	2.12912e-07	.	0.0000	1.0000

Table 3. Principal components with eigenvalues greater than 1

Variable	Comp1	Comp2	Comp3	Unexplained
X1	0.3371	0.1692	-0.0194	.08182
X2	0.0044	0.0697	0.6992	.2708
X3	0.3287	-0.2905	0.0312	.02101
X4	0.1333	0.4380	-0.1490	.479
X5	0.3509	-0.1722	0.0175	.007851
X6	0.3169	-0.2872	0.0264	.08735
X7	0.3144	0.2840	-0.0249	.09866
X8	-0.0118	0.0864	0.6894	.2852
X9	0.3518	-0.1228	0.0227	.02983
X10	0.3550	-0.1287	0.0044	.0107

X11	0.1353	0.5457	-0.0703	.3051
X12	0.3591	-0.0638	-0.0009	.01142
X13	0.2039	0.3957	0.0724	.3877

5.2 Parameter Adjustment

This method adjusts 2 parameters of the RF algorithm, uses GridSearchCV in Python's model_selection library for grid search, and uses roc_auc _score in sklearn.metrics library for evaluation judgment, and selects the most suitable 'n_estimators', 'min_ samples_split', 'max_depth', 'criterion':['gini','entropy'] to improve the prediction accuracy of the model and prevent it from over-fitting the data.

The result of grid search is that the four metrics are 'n_ESTIMators' = 125,'min_samples_split' = 20,'max_depth' = 5, and 'criterion': ['gini','entropy'] = 'entropy'.

Table 4. Training parameters of Random forest model

Parameter	Name of parameter	Parameter values
n_ESTIMators	Maximum Iterations	125
min_samples_split	The minimum number of samples that a node can divide	20
max_depth	Maximum depth of decision tree	5

5.3 Results and Discussion

It can be concluded that the PCA-RF model has good prediction accuracy. The AUC (Area Under Curve) is the area under the ROC (receiver operating characteristic curve) curve enclosed by the coordinate axis, and the value ranges from 0.5 to 1. The closer the AUC is to 1.0, the higher the authenticity of the detection method.

To further test the performance of the combined PCA-RF model, the results of the combined model were compared with those of the RF model without PCA, and it was found that the method based on PCA and random forest classification improved the accuracy of the prediction to some extent, and the truthfulness of the detection method was higher.

In addition, the table also compares the combined PCA-RF model with the traditional methods of logistic regression, support vector machine model, and neural network model. From Table 5, it can be seen that the prediction accuracy of the PCA-RF combined model is slightly higher than the other methods, but the AUC index is significantly better than the other models. Importing sklearn.ensemble and sklearn.metrics with Python is easy to operate, and processing data is relatively fast, and the PCA-RF combination model is both easy to apply and can achieve good prediction results.

Table 5. Comparison of prediction results between different models

Method	AUC	Prediction Accuracy
Neural Network	0.840	0.946
SVM	0.826	0.950
Logistic Regression	0.847	0.939
Random Forest	0.835	0.938
PCA-RF	0.882	0.956

The prediction results show that the combined PCA-RF model has a good prediction accuracy of 95.6%, which is better than SVM and logistic regression. At the same time, the combined PCA-RF model outperforms the RF model without PCA, indicating that PCA can improve the performance of the RF model in predicting financial fraud. By comparing the prediction results of the combined PCA-RF model and the neural network model, it is found that the prediction results of the combined PCA-RF model and the neural network model are closer, but the detection method of the combined PCA-RF model is more realistic, simpler, more efficient, and easier to apply in practice.

6 Conclusion

In conclusion, by directly collecting raw financial statement data, the PCA-RF combined model can be applied to identify possible fraudulent behaviors, give early warning to stakeholders, provide strategic support for corporate decision-making, and facilitate supervision and control by regulatory units. In this paper, 13 financial statements are selected to predict financial fraud, hoping to predict fraud directly and easily from published financial statements intuitively and conveniently.

However, this paper still has some shortcomings in that it only considers the analysis of the financial data itself and does not combine macro factors such as policies, laws, and national conditions to make a comprehensive forecast. Analyzing the policy influencing factors together as variables can further improve the accuracy of the prediction. For example, whether the new Securities Law has been enacted and whether the regulators have imposed penalties for fraud can be used as categorical variables for a comprehensive analysis to make the forecast results more comprehensive and accurate. The influence of these macro factors should be taken into consideration in subsequent studies on fraud prediction, and this paper will be improved in this aspect in the future.

References

1. Bao Y., Ke B., Li B., Yu J., Zhang J.,2020, Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach [J], Journal of Accounting Research, 58(1),199~235.

2. Beaver W. H.,1966, Financial Ratios As Predictors of Failure [J], Journal of Accounting Research,4,71~111.
3. Cecchini M., Aytug H., Koehler G.J.,Pathak P.,2010,Detecting Management Fraud in Public Companies [J], Management Science,56(7),1146~1160.
4. Dechow P. M., Ge W.,Larson C.R.,Sloan R. G.,2011,Predicting Material Accounting Misstatements [J],Contemporary Accounting Research,28(1),17~82.
5. Kotsiantis S.B., ZaharakisI.D., Pintelas P.E., 2006, Machine Learning: A Review of Classi-Fication and Combining Techniques [J], Artificial Intelligence Review,26(3),159~190.
6. Ravisankar P.,Ravi V., Raghava Rao,G.,BoseI.,2011,Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques [J], Decision Support Systems,50(2),491~500.
7. Li X., Chen S.J.,2013, Research on profit manipulation of Chinese Listed Companies Based on BP neural network model and DEA model [J], Mathematical statistics and management,32(03),440~451.
8. Liu J., Wang L.P,2006, Financial Fraud Identification Model Based on Probabilistic Neural Network [J], Journal of Harbin University of Commerce (Social Science Edition), (03), 102~105.
9. Noori R., Karbassi A.R., Moghaddamnia A., Han D., Zokaei-Ashtiani M.H., Farokhnia A., Ghafari Gousheh M., 2011, Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction [J], Journal of Hydrology，401(3),177~189.
10. Gu H.Q.,2010, Research on application of support vector Machine in futures price prediction [J].Computer simulation,27(12),358~360+385.
11. Zhou W.H., Zhai X.F., Tan H.W.,2022, Research on financial Fraud Prediction Model of listed Companies based on XGBoost [J]. Research on quantitative economy and technical economy,39(07),176~196.
12. He K., Yang S.X., Gao Y.G., 2019,. Tunnel traffic accident duration prediction based on pca-rf combined model [J].Traffic information and safety,37(05),26~32.