Published in final edited form as:
Int J Econ Finance. 2015 June 25; 7(7): 178–188. doi:10.5539/ijef.v7n7p178.

Financial Fraud Detection Model: Based on Random Forest

Chengwei Liu[1], **Yixiang Chan**[1], **Syed Hasnain Alam Kazmi**[1], and **Hao Fu**[1]

[1]School of Economics and Management, Southwest Jiaotong University, Chengdu, Sichuan, China

Abstract

Business's accelerated globalization has weakened regulatory capacity of the law and scholars have been paid attention to fraud detection in recent years. In this study, we introduced Random Forest (RF) for financial fraud technique detection and detailed features selection, variables' importance measurement, partial correlation analysis and Multidimensional analysis. The results show that a combination of eight variables has the highest accuracy. The ratio of debt to equity (DEQUTY) is the most important variable in the model. Moreover, we applied four statistic methodologies, including parametric and non-parametric models to construct detection models and concluded that Random Forest has the highest accuracy and the non-parametric models have higher accuracy than non-parametric models. However, Random Forest can improve the detection efficiency significantly and have an important practical implication.

Keywords

financial fraud detection; random forest; ratio of debt to equity; partial correlation analysis; statistic methodologies; parametric models

1. Introduction

In today's globalized market, the size of the firm is larger day by day, and more arenas are evolved and in the context fraudulent behavior of enterprise leads to huge losses and brings serious damage to the investors' confidence and the people pay more attention to the topic and brand companies are focusing to target the pricing in the right manner (e.g., Breiman, 2001; Liaw & Wiener, 2002; Strobl et al., 2009; Fang et al., 2012; Hansen 1996; Kirkos et al., 2007; Alam Kazmi, 2015a, b). However, reports of fraudulent behavior detection were difficult for staff supervision. Kirkos et al. (2007) found that there are three main reasons. First, there is lack of knowledge about financial fraud, although many prior researches on the financial fraud were investigated to some conclusions, but they were far away from the practice. Second, most auditors lack experiences, which makes them difficult to detect the roots in financial fraud. Finally, corporate executives deliberately deceive and they use sophisticated tools to allow auditors not to know where to begin. Therefore, how to help the

Correspondence: Hao Fu, School of Economics and Management, Southwest Jiaotong University, Chengdu, Sichuan, China. Tel: 86-028-8760-0895. haofu12@126.com.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

government and auditors to improve the ability to detect fraud is crucial problem and this paper try to meet this demand with in depth analytical tools and techniques. We demonstrated models to make a companion with Random Forest techniques and compared the efficiency between a parameter and non-parametric models.

1.1 Models in Random Forest

Identification of financial fraud is a challenging task. A large number of scholars have established recognition model to detect it from different angles, using different methods, such as define by Hansen (1996) and Kirkos et al. (2007). Hansen (1996) applied a generalized qualitative response model (EGB2) to identify management fraud. They used data from the international accounting firm to fit the model and it showed a good predictive ability. They Found that the model can solve asymmetric errors cost from the Type II and I errors, it can effectively prevent the loss from lawsuit from making the Type II error. Kirkos et al. (2007) used a data mining method to establish three fraudulent identification model, known as Decision Tree (DT), Neural Networks (NN) and Bayesian Network (BBN) and through Ten-fold cross-validation. The results show that Bayesian network (BBN) has the highest accuracy. Ravisankar et al. (2011) used multi-layer feedback neural networks, support vector machines, probabilistic neural networks and other four methods to build models. The results showed that probabilistic neural network (PNN) performance was outstanding. Fanning and Cogger (1998) constructed model with the neural network and the independent variables included the financial ratios and qualitative variables. They also used some other traditional statistical methods to build models to make a comparison with neural networks. They concluded that neural networks have a higher accuracy than the traditional statistical methods.

1.2 Model Developments

Summers and Sweeney (1998) established a Logistic model to validate the relationship between internal transactions and fraud. They found that managers would decrease the company’s stock holdings by frequent trading when there is a fraud.

Abbott and Parker (2000) exam whether the presence of an independent audit committee effectively reduces the possibility of corporate fraud through a regression. The regression results show when the independent audit committee and corporate annual meetings held at least more than twice, the company have reported to reduce financial reporting errors. Chen et al. (2006) studied the relationship between corporate governance and ownership structure corporate financial fraud with the sample of Chinese listed companies. The results show that corporate ownership structure and board of governance characteristics are important indicators for fraud interpretation. They also found that the ratio of independent directors, board meeting frequency and the term of office of the chairperson related to fraud.

In this paper, based on prior studies, we introduced and analyzed the application of random forest in financial fraud. We construct and test the model with data from listed companies from China. Meanwhile, we construct some models to make a companion with Random Forest. We compared the efficiency between a parameter models and non-parametric models. The study follow the structure as, introduction to the research method in Section 2

including random forests, feature selection and introduction of data. The empirical analysis, including fitting, testing of random forests, and other models and presented in Section 3 and Section 4, presents the results discussions and conclusions.

2. Research Methodology

2.1 Approach for Random Forest

Random Forest is a combined classification method proposed by famous scholars (for instance, Breiman, 2001; Liaw & Wiener, 2002; Strobl et al., 2009; Fang et al., 2012) and including technique of Decision tree (DT) by Strobl et al. (2009) is the basic classifier and it establishes a large number of trees. The specific algorithm process is as follows:

Step 1: Sample “m” (the number of samples chosen) (m < M, where M is the number for the entire sample) samples randomly from all samples with a bootstrap method Liaw and Wiener (2002).

Step 2: Construct a decision tree with the extracted sample, which is no pruning.

Step 3: Repeat steps 1, 2 and build a large number of decision trees and develop decision tree classification sequence $\{h_1(X), h_2(X), \ldots, h_{\text{ntree}}(X)\}$.

Step 4: Final classification is determined by each record vote from the results of the decision tree classification.

It can be expressed as follows, which h_i is a single decision tree model trees, Y Represents the output variable (or target variable) and $I()$ is the indicator function.

$$H(x) = \arg\max_{Y} \sum_{i=1}^{n} I(h_i(x) = Y)$$

During the random forest modeling, data are from a bootstrap Liaw and Wiener, (2002) sampling so that about 1/l = 0.368 samples, which we call it out of bag that have not been drawn. Always use this part data as a test set to test the performance of the model and the estimated error rate called OOB Estimation. Breiman (2001) has proved that OOB estimate is an unbiased estimation. This internal unbiased estimate makes random forest not appear overestimated.

2.1.1 Measurement Method Estimation—The concept of variable importance is problematic to define. The measurement method from random forests is significantly different from the conventional method, which is the one of the main features of random forest. It follows the theory that, during the process random forest modeling, it will generate an OOB and an OOB estimation. In order to evaluate the importance of the variables, in a random target bag, change targets variable randomly on condition keeping the other variables form OOB constant. Then use the noise data to test the model and get another OOB estimation. These two OOB estimation have a positive relationship with the variable importance. The difference from the two OOB estimation divided by the standard deviation and the result is the variable importance. We use the variable’s importance to delete the

unimportant variables until two variables left. The process helps to choose the most outstanding performance model of Breiman (2001).

2.1.2 Ballpark Matrix of Data—Proximity between any two points is define as the ratio of the number of occurrences of the two data on the same end of the classification tree node. Establish a N × N-dimensional proximity matrix (N is the number of data points) and each element of the matrix represents the random forest each tree in two corresponding data points fall on the same end node ratio. According to the experience, dissimilar data points gathered at the end of the probability of a branch is not significantly higher than the probability of proximity data points gathered. However, the computation of the proximity of the matrix is relatively very large. The proximity can be applied to input the missing values, detect the singular values, make partial plot, clustering analysis, dimension reduction and visualization. In this study, we introduce the partial plot and dimension reduction.

Partial plot can present how a variable from the “black box” (SVM, RF, DT, NNT, classification and regression model impact the prediction in a visual way. (Zhang et al., 2014) In broad-spectrum, the classification or regression function depends on many predictors. Partial dependence of classification or regression function for a particular variable (X_j) defined as function of exception of remaining variables. In practice, usually a fixed, variable X_j, it averages other variables of the prediction function for all combinations. This process requires the training data set for each value of the variable X_j predictions from all data. However, partial plot in random forest does not apply variables from the training set. Rather, use the training data set variables X_i variables within the scope of equidistant Interval data. When analyzing just need to specify the number of segments. This is very useful when the data is large.

Multidimensional scaling analysis is used to reflect the number of examination between things similar (dissimilar) degree. Through an appropriate dimension reduction method, this will be similar (dissimilar) degree in low-dimensional space using the distance between points that out. This may help to detect those potential factors. Random forest generates a proximity matrix of data points. The value of all individuals in the proximity matrix within 0-1 and it is the distance between data points. Bipartite metric multidimensional scaling analysis plot is scatter distribution from first two principal components analysis of the proximity matrix.

2.2 CSMAR Data

In this study, CSMAR (China Stock Market & Accounting Research) database is used. As financial fraud includes many aspects, so if we classify all indicators presenting different kinds of fraud in a class, it will weaken the detection function. In this paper, we emphasis on analysis the listed companies involving manipulating profits. In order to control the external environment and industry factor as well as taking into account the difficulty of collecting data, we collect data to follow the following rules.

1). We argue that the company commits fraud in different years and its annual report meets the fraud samples selection and the annual report from the non-fraud years meet the non-fraud samples.

Europe PMC Funders Author Manuscripts

2). We select the companies were that, disclosed fraud in the annual report. At the same time exclude ST, * ST and PT companies.

3). Given the diverse industry have a significant difference to indicators; this study involves the manufacturing listed companies from 1998 to 2014. Finally, we selected 138 fraud samples and 160 non-fraud samples.

2.3 Variable Selection and Indications

Indicators in this study were chosen from preceding studies, such as of Kirkos et al. (2007) and James (2003). The researchers have found that corporate fraud and capital structure, asset composition are closely linked. In this paper, we use the ratio of debt to equity market (DEQUTY), current assets ratio (CURASS), fixed assets ratio (FIXASS) to express this relationship. Persons (2011), Feroz et al. (1991), Fanning and Cogger (1998) show that sales can help to predict corporate fraud. Accounts receivable and income ratio (ACRESAL), inventory and income ratio (INVIN) mobile asset turnover (ACURAST), fixed assets and income ratio (FASINC) are applied to represent sales. Stice (1991); Persons (2011); Fanning et al. (1998); Spathis (2002); Abbott and Parker (2000), found that corporate growth opportunities, the value of equity investments, the investment and corporate profitability can effectively help identify fraud. In this study we emphasis on Price-Earnings ratio (PE), Sales Ratio (PS), Book Value (PBV) that, represent growth opportunities and equity value of the businesses. We use Return on Invested Capital (RINCAP) and Long-Term Capital Gains (LOTCAG) to indicate the Profitability of Investments. Operating Margin (OPROSAL), Return on Assets (ROA) of the Total Net Assets Ratio (NPROTA), current assets, Net Profit Margin (NPROCA), Net Fixed Assets Ratio (NPRFAS), and Return on Equity (ROE) is engaged to represent the profitability of enterprise.

Summers and Sweeney (1998) found that the corporate debt level is an important indicator to recognize the corporate financial fraud. Therefore, this paper uses gross profit and EBIT (TPEBIT), EBIT and operating income ratio (EBITSAL), cash flow interest coverage ratio (NOCFIE), interest coverage ratio (TINEAR), long-term debt to capitalization ratio (LTDCAP), working capital and borrowing ratio (WOCAPL) to measure a corporate solvency. Ravisankar et al. (2011); James (2003) and Cohen et al. (2004), found that funding pressures, financial risk, and corporate fraud are closely related; hence, we use Financial Leverage (DFL) to measure this relationship.

We also apply the Management Expense Ratio (MANEXP), Working Capital Ratio (WORCAP), Fixed Asset Turnover (FASSTU), Mobile Asset Turnover (CURAST), and Cash Flow Ratio (OPECAF) to express companies operating capacity. We select 29 features as input variables and the descriptive statistics are shown in Table 1. We set up the dependent variable Fraud, where “1” indicate the fraud companies and “0” represent the non-fraud companies.

Europe PMC Funders Author Manuscripts

3. Analysis of the Study

3.1 Model Test Method Cross-Validation Approach

We applied the commonly used method in machine learning, model test method cross-validation approach. We divided the data into five parts, including four training model and the rest for testing model. Every part done as the test data. The mean error rate of the five tests will be a standard for impartiality that, how the model performance.

First, we will build a model with all variables, get the error, and then exclude the most unimportant variable from important measurement until only two variables left. We choose the combination with the lowest error to construct the random forest model. We analysis data using statistical software R Ver.3.0.2 and with the Random Forest software package. The parameters node number mtry was set at Squrt (M) (M is the number of independent variables) and the number of the tree is 2000. The result shown in Figure 1.

Figure 1, shows that it has the lowest error rate with the 8 independent variables and the error is 12.38%. Therefore, we will choose the 8 variables to build the model and the variables are shown in Table 2.

In order to improve the accuracy, we will find the best parameter to optimize model built and the result are shown in Figure 2. The figure shows that the best parameter of mtry is 3 and there is no impact when the tree number is more than 500. We will set these two parameter at 3 and 500, again to test the model by five-fold test and we get a 12% error rate. Consequently, the optimization of parameters has an obvious impact on the results.

3.2 DAccuracy Method Measurement

Measurement methods based on random forests, including Decease Accuracy and Decease Gini coefficient are used to measure the importance of the variables in the model. While DAccuracy method of measuring the importance of each variable on the dependent variable Fraud and Nfraud follows two kinds of companies. The results are shown in Table 3. Results show that the measured maximum prediction accuracy and Gini coefficient of 48 and 48.61, respectively, and for the same indicator–DEQUTY and its value is much higher than the other indicators. The most important indicator of the model is DEQUTY. The value of the other variables in the model is relatively not particularly prominent. The resulting values of the two measurement methods are the same category of high to low. Therefore, the Variable importance in the model has the stability.

We use the Deceasing accuracy measurement method to measure the importance of variables for each type of company. We still found that the greatest degree of importance DEQUTY variables and it is much higher than other variables. The importance of the variable does not exist, particularly prominent among other variables. We use the Mann-Whitney U test to validate the difference of the importance of variables between Nfraud and Fraud companies. The results show the importance of the two types of variables have significant differences ($p < 0.000$) and the variable importance for fraud is significantly higher than Nfraud.

3.3 Partial Dependence Plot and MDS

Partial plot represent how a variable affect the results in the “black box” classification and regression in a graphic way. In this paper, we draw for partial plots for the important variables in the model, which are DEQUTY, TPEBIT, CURAST, FASSTU, and as shown in Figure 3. Fig. a, is the partial plot of DEQUTY and its partial correlation is not a linear relationship. The interval between [0, 0.4] and [0.6, 0.8] is negatively correlated with the model prediction accuracy. A correlation is positive when interval is between 0.4 and 0.6. When the variable is greater than 0.8 there is no effect on the accuracy for the prediction and that there is no partial correlation. Fig. b, is partial plot for TPEBIT. When the variable is less than 0.9, predictive accuracy has significantly improved, greater than 0.9 and less than 1 by rapidly reducing the accuracy of the forecasts, when greater than 1 variable has no effect on prediction accuracy. That means no partial correlation. Fig. c, shows the relationship between CURSAT and dependent variables. When the variable is less than 0.5 CURSAT the prediction accuracy can be reduced. In addition, when the interval between [0.5, 1.5] it presents a significant positive correlation and it improves the detection accuracy. However, when CURSAT is greater than 1.5 also show a slow negative correlation. Fig. d, is partial plot for FASSTU. The figure shows when FASSTU is less than 0.5. The variables significantly improve the prediction accuracy but the rapid decline presents wavy when in range, when the variable is greater than about 2.7 it presents a wavy, but the impact was not significant.

Earlier, we use the partial plot to analysis the most important variables impaction on the predictive accuracy. Here we will analyze the results multi-dimensionality scaling analysis and show the two main components of the scatter plot for two main principal component analysis for the proximity matrix shown in Figure 4. From the figure shown, some samples can be distinguish significantly. Although at the bottom of the figure a small part samples cannot be distinguish, but most of the samples can be identified effectively. Therefore, better accuracy estimated.

3.4 Constructed Models for Logistic/K-Mean/DT/SVM

The Random Forest model have a high accuracy reaching 88%. In order to compare its accuracy with other methods as well as parameter models and nonparametric models. This study established several other detection models including two parametric models (Logistic and KNN) and two non-parametric models (Decision tree and Support vector machine). Their feature selection follows the following principles. We use Mann Whitney U test to validate the difference between the two type companies and we will choose the significance under 0.05 variables to construct model as well as delete the variables when the correlation ship parameter is greater than $\rho > |0.5|$. Finally, we choose seven features as the input variables AS shown in table 2. First, we use the all variables to build the Logistic model. Then removed insignificant variables stepwise regression. We set threshold at 0.5, meaning to distinguish different types by 0.5 as a boundary. All data used to fit the model to get finally get the formula shown below.

$$\ln\left(\frac{p}{1-p}\right) = -0.16 + 0.0037\mathrm{PE} - 0.35\mathrm{CURAST} + 0.53\mathrm{WORCAP} \quad (2)$$

KNN, DT and SVM three kinds of model are “black box” model, which does not specifically use the expression of a specific formula. Therefore there will not be specifically addressed in detailed. We use all data to fit and test the models and results are shown in Table 5.

The results show that the Logistic model’s identification for frauds was 42.03%, 23.18% for Nfraud, only 32.18% for total. KNN model successfully identified 88.41% of the fraudulent companies and 87.5% of non-fraud companies and the total success rate is 87.92%. DT model has an efficiency of 77.54% for fraud Company, and 83.13% for Fraud Company. 80.54% in total. SVM model for fraudulent companies identify low success rate of 66.67%, but the company’s non-fraudulent identification efficiency is much higher at 80%, the final full recognition efficiency 73.83%. From the results show that the overall success rate of recognition to identify the most efficient KNN, DT is then followed by SVM, least efficient Logistic model is only 32.18%.

3.5 Model Validation

Training set is used to test the model that will lead to partiality, because the model is in the training process and will produce a sample memory, resulting in the testing process, which will make the results of test at the higher rate. Therefore, in order to avoid biasness to affect the performance of the model, this paper uses half of the actual cross-validation of the model test method. All samples are divided into five equal portions, including four models for training, and the remaining one is used to test the model. Each used as a test set and we will get an average error rate. Since all models are run on R3.0.2 software platforms, so all aliquots data are the same, and the efficiency of the model can be a valid in to comparison, the results shown in Table 6.

From, Table 6 showing as Logistic model recognizes the overall success rate was 42.91%, the identification of Nfraud success rate of 37.5%, successfully identified fraud 49% of the sample is the lowest of all models model identification success rate. KNN model to identify the overall efficiency of 60.11%, compared with logistic model identification efficiency has been greatly improved. Recognition success rate for Nfraud and Fraud was 59% and 63%, respectively. DT model can successfully identified 66.43% of the tested samples, Nfraud and Fraud Recognition accuracy was 68.13% and 63.62%, for Nfraud recognition success rate is higher than Fraud, committing Type I error than to commit a Type II error have a higher possibility. SVM model identification of 81.88% and 78.13% of Nfraud sample of Fraud sample test set, the overall success rate reached 80.8 percent, the results of the test is relatively good. RF models are models to identify the most efficient models, the overall success rate of 88 percent recognition, which the Nfraud recognition success rate reached 85% of Fraud recognition success reached 90.71%. Possibility of committing Type II error is lower than the likelihood of committing Type I error. This is for audit firms and government departments is very important. If a normal company misjudged as having a fraud, then to his reputation, credibility a devastating blow, then the next they will face huge

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

damages, and random forests to some extent alleviated such crises. The ability to identify the strongest random forest and extrapolation ability is the best, we can see that the introduction of random forests overall financial recognition from the perspective of recognition rate can improve the recognition of financial fraud. We also found that recognition rate parameter model, which are Logistic and KNN to be significantly lower than the non-parametric models that, are DT, SVM and RF, which identify the lowest efficiency Logistic model.

4. Discussion

This paper introduce the Random Forest model to financial fraud and implied random forest method to data mining. Moreover we compared four other models including two parametric models with two non-parametric models and found that Random forest have the highest accuracy. It has been shown in comparison with other classification models such as of (Kirkos et al., 2007; Liaw & Wiener, 2002), and random forest has incomparable advantages to other models in several ways. Firstly, it has a very high recognition efficiency of random forests; in almost all of its models, it has the highest rate. Secondly, it ignore data normality assumption and handle efficiently more non-normal data fields. It can be a good deal of high-dimensional data analysis and co-linear over-fitting case that does not appear easily. Third, it can measure importance of each variable and can effectively eliminate unimportant variables. Finally, we can select the best combination of variables to build models. The partial correlation plot for each variable can effectively compensate for deficiencies in “black box” model. The results of the study found that non-parametric models have a higher accuracy than parametric models. The reason is the variance of the normal distribution of data, which will lead to a basis of the theoretical parameters of the model. In practice, the data always does not meet the requirement. We conducted on selected variables with Kolmogorov-Smirnov normality test, results in Table 1 shows that except for CURASS, all other variables are significant at the 0.05 level of significance. Most of the data does not follow a normal distribution. So the model parameters based on the normality assumption when fitting with a non-normal data lead to an inevitable recognition efficiency, but the parameters model can solve the problem perfectly. The unpatrolled advantages make it the highest detection efficiency in the financial fraud.

5. Conclusion

This investigation studies the application of the random forest in financial fraud detection. The model is fitted with Chinese listed company data and used to variable selection. Moreover, we measured all variables’ importance and made the partial plot and an analysis with multidimensional scaling. Finally, we establish four models with four statistical methods and compare the difference of models as also summarized here. Study 1: In this study, we find that the model performance most outstanding when there are eight input variables and the accuracy is 88%. The other models’ accuracy is 42.91%, 60.11%, 66.4% and 80.18%. Random forest commits the Type II error probability to be significantly lower than Type I error. The introduction of random forests, detect financial fraud significantly and improve the efficiency. Study 2: We can see from the random forest variable importance measure. The ratio of debt to equity market (DEQUTY) is the most important variable, not

only for model but also for both two kinds companies. Random Forest model pay more attention to capital formation. Study 3: We found that the parameters models have a lower accuracy than the non-parametric models. Least efficient model parameter model is Logistic with a 42.91% accuracy and the KNN is only 60.11%. Non-parametric model SVM and RF reach 80.18% and 88%. So we know from the results that parameter identification models compared to non-parameter models will have greater preeminence.

References

Abbott LJ, Parker S. Auditor selection and audit committee characteristics. Auditing: A Journal of Practice & Theory. 2000; 19(2):47–66. http://dx.doi.org/10.2308/aud.2000.19.2.47.

Alam K, Syed H. Developments in Promotion Strategies: Review on Psychological Streams of Consumers. International Journal of Marketing Studies. 2015a; 7(3):129–138. http://dx.doi.org/10.5539/ijms.v7n3p129.

Alam K, Syed H. Brand the Pricing: Critical Critique. International Journal of Marketing Studies. 2015b; 7(3):125–128. http://dx.doi.org/10.5539/ijms.v7n3p125.

Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32. http://dx.doi.org/10.1023/A:1010933404324.

Chen G, Firth M, Gao DN, Rui OM. Ownership structure, corporate governance, and fraud: Evidence from China. Journal of Corporate Finance. 2006; 12(3):424–448. http://dx.doi.org/10.1016/j.jcorpfin.2005.09.002.

Cohen JR, Krishnamoorthy G, Wright A. The corporate governance mosaic and financial reporting quality. Journal of Accounting Literature. 2004:87–152. Retrieved from https://www2.bc.edu/~cohen/Research/Research4.pdf.

Fang, KNB.; Jian-Bina, WU. A Review of Technologies on Random Forests. Statistics & Information Forum. 2011. Retrieved from http://en.cnki.com.cn/Article_en/CJFDTOTAL-TJLT201103007.htm

Fanning KM, Cogger KO. Neural network detection of management fraud using published financial data. International Journal of Intelligent Systems in Accounting, Finance & Management. 1998; 7(1):21–41. http://dx.doi.org/10.1002/(SICI)1099-1174(199803)7:1<21::AID-ISAF138>3.0.CO;2-K.

Feroz EH, Park KJ, Pastena V. The financial and market effects of the SEC's accounting and auditing enforcement releases. Journal of Accounting Research. 1991; 29:107–142. http://www.jstor.org/stable/2491006.

Hansen J, McDonald JB, Messier W Jr, Bell TB. A generalized qualitative-response model and the analysis of management fraud. Management Science. 1996; 42(7):1022–1032. http://dx.doi.org/10.1287/mnsc.42.7.1022.

James KL. The effects of internal audit structure on perceived financial statement fraud prevention. Accounting Horizons. 2003; 17(4):315–327. http://dx.doi.org/10.2308/acch.2003.17.4.315.

Kirkos E, Spathis C, Manolopoulos Y. Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications. 2007; 32(4):995–1003. http://dx.doi.org/10.1016/j.eswa.2006.02.016.

Liaw A, Wiener M. Classification and Regression by random Forest. R. News. 2002; 2(3):18–22. Retrieved from http://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf.

Persons OS. Using financial statement data to identify factors associated with fraudulent financial reporting. Journal of Applied Business Research. 2011; 11(3):38–46. Retrieved from http://cluteinstitute.com/ojs/index.php/JABR/article/view/5858/5936.

Ravisankar P, Ravi V, Raghava RG, Bose I. Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems. 2011; 50(2):491–500. http://dx.doi.org/10.1016/j.dss.2010.11.006.

Spathis CT. Detecting false financial statements using published data: Some evidence from Greece. Managerial Auditing Journal. 2002; 17(4):179–191. http://dx.doi.org/10.1108/02686900210424321.

Stice JD. Using financial and market information to identify pre-engagement factors associated with lawsuits against auditors. Accounting Review. 1991:516–533. Retrieved from http://www.jstor.org/stable/pdf/247807.pdf?acceptTC=true.

Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods. 2009; 14(4):323. http://dx.doi.org/10.1037/a0016973. [PubMed: 19968396]

Summers SL, Sweeney JT. Fraudulently misstated financial statements and insider trading: An empirical analysis. Accounting Review. 1998:131–146. Retrieved from http://www.jstor.org/stable/248345.

Zhang LW, Zhang XD, Liu SR, Sun P, Wang TL. The basic principle of random forest and its applications in ecology: A case study of Pinus yunnanensis. Acta Ecologica Sinica. 2014; 34(3): 650–659.

Europe PMC Funders Author Manuscripts

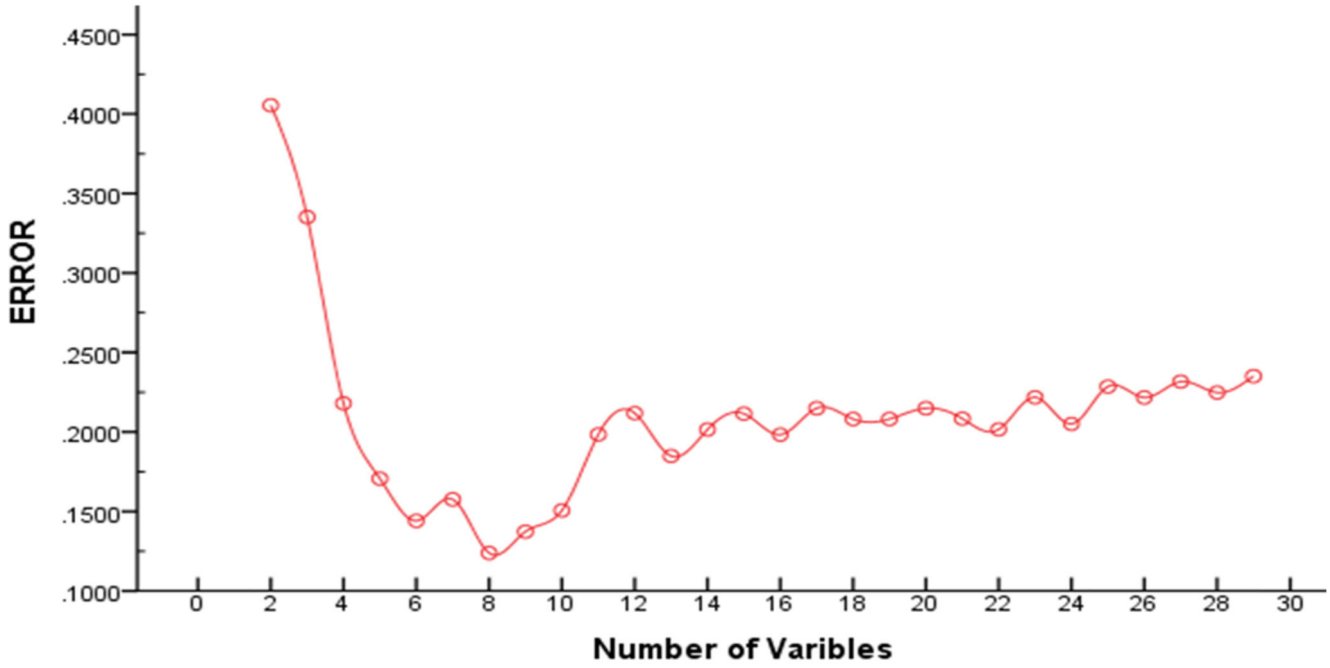

Figure 1.
RF five-fold across-test

Europe PMC Funders Author Manuscripts

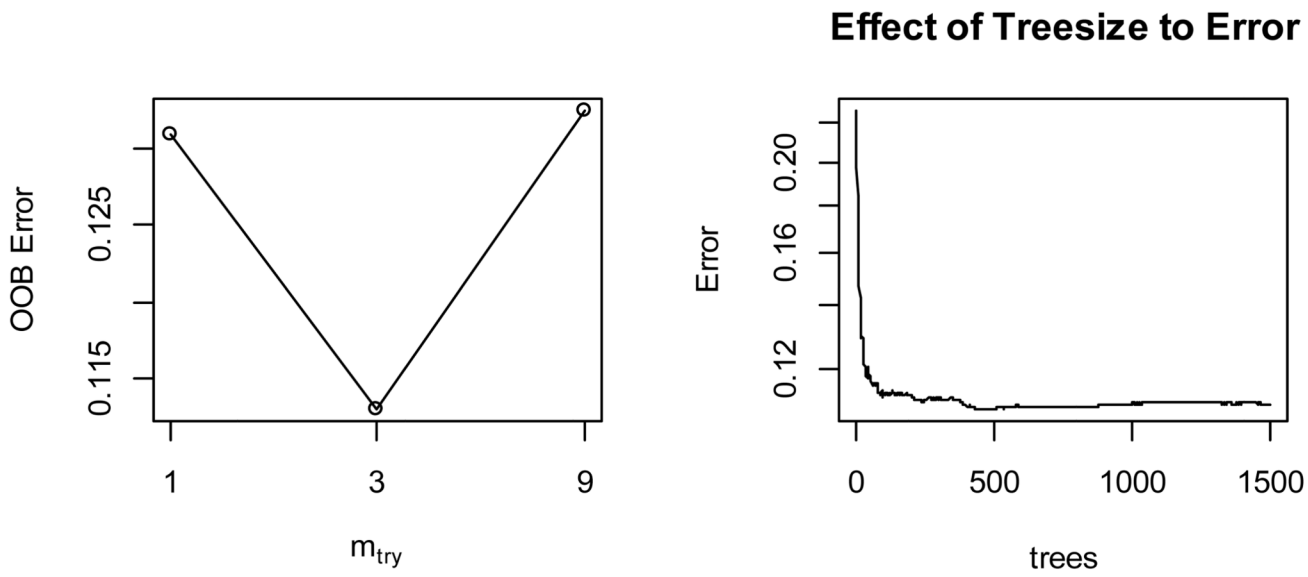

Figure 2.
Parameter optimizing

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

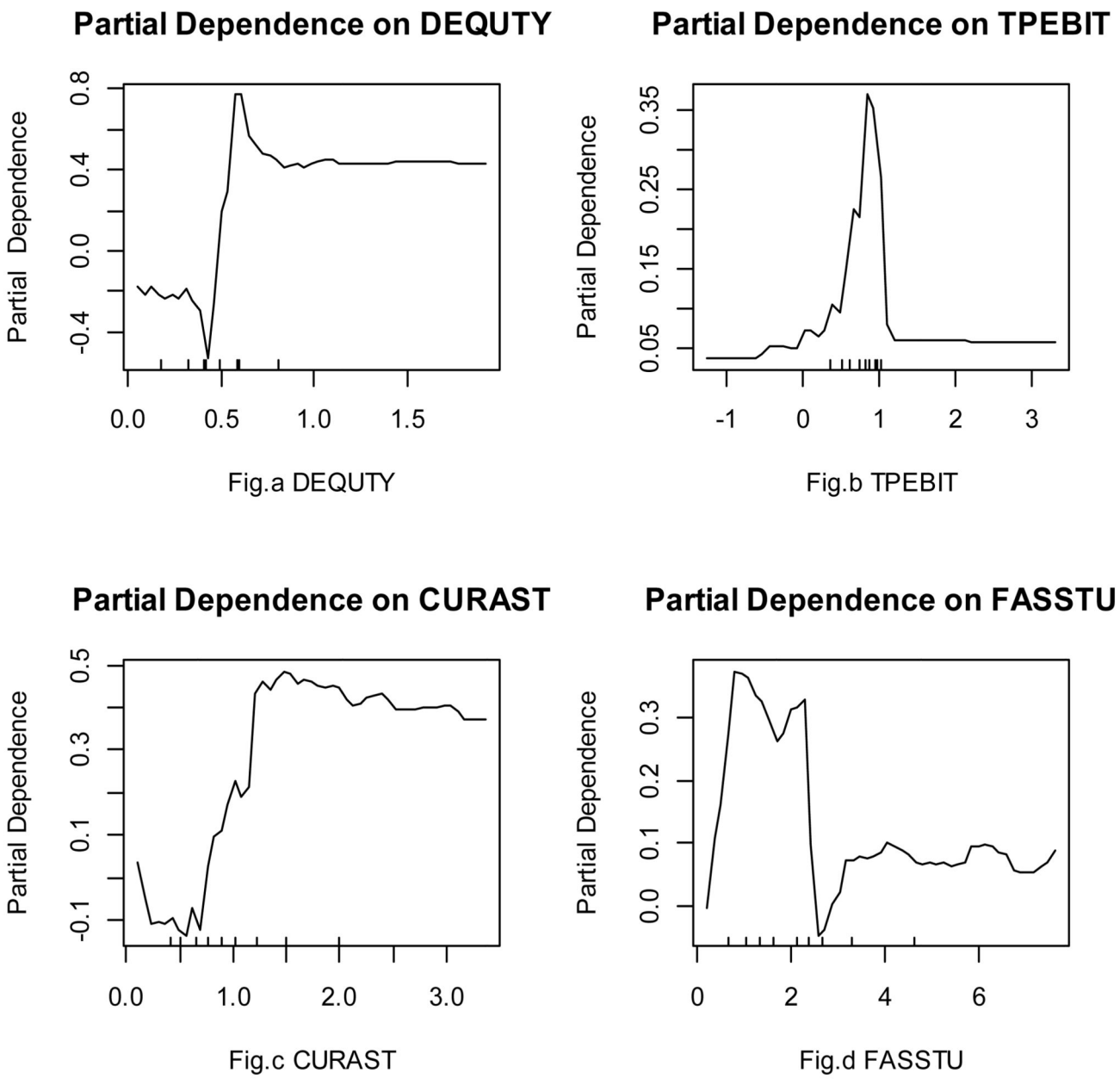

Figure 3.
Partial dependence plots for the first four important variables

Europe PMC Funders Author Manuscripts

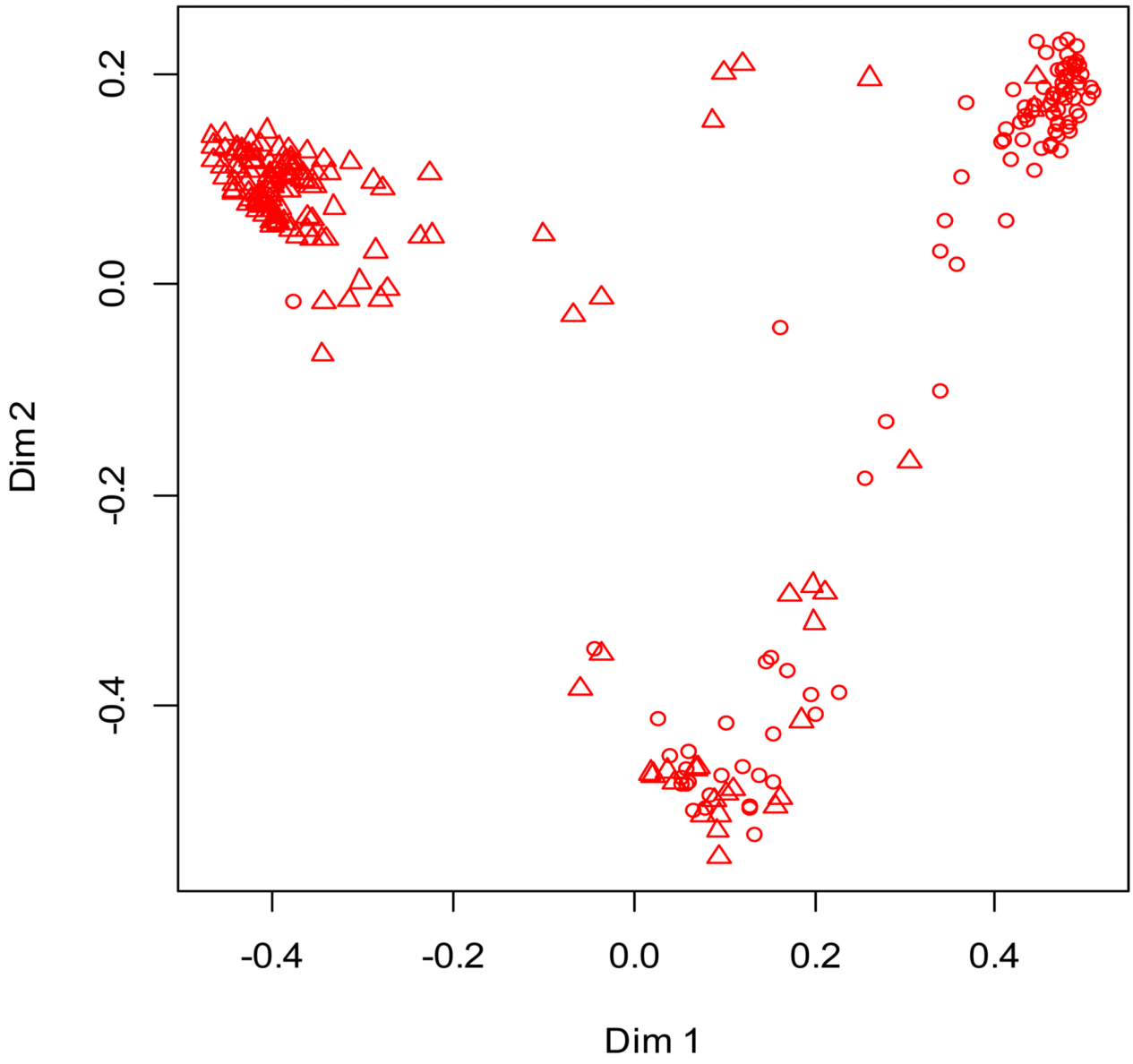

Figure 4.
Multi-dimension scaling plot

Europe PMC Funders Author Manuscripts

Table 1

Descriptive statistics

Variables	Min	Max	Mean	sd	Mann Whitney U	Normality test
OPRSAL	−0.654	0.327	0.047	0.151	.433	0.000
TPEBIT	0.023	1.355	0.756	0.252	.649	0.016
EBITSAL	−0.950	0.846	0.089	0.175	.877	0.000
ROA	−0.184	0.258	0.053	0.068	.008	0.001
NPROTA	−0.215	0.220	0.033	0.064	.025	0.000
NPROCA	−0.636	0.322	0.051	0.138	.006	0.000
NPRFAS	−0.815	2.031	0.168	0.347	.761	0.000
ROE	−0.540	0.378	0.046	0.133	.010	0.000
RINCAP	−0.223	0.554	0.084	0.117	.061	0.000
LOTCAG	−0.286	0.546	0.085	0.114	.001	0.000
MANEXP	0.008	1.019	0.116	0.130	.047	0.000
PE	6.059	296.667	83.410	64.469	.015	0.000
PBV	0.049	14.884	4.252	2.757	.761	0.000
PS	0.320	21.763	5.153	4.373	.069	0.000
NOCFIE	−6.430	63.675	7.903	11.279	.545	0.000
OPECAF	−1.137	1.273	0.112	0.246	.106	0.000
ACRESAL	0.001	1.913	0.400	0.370	.005	0.000
INVIN	0.016	1.934	0.323	0.290	.798	0.000
CURAST	0.057	8.261	1.154	0.891	.000	0.000
FASINC	0.038	4.172	0.700	0.632	.291	0.000
FASSTU	0.111	9.258	2.412	1.780	.664	0.000
DFL	−0.539	5.282	1.464	0.716	.737	0.000
WORCAP	−1.726	0.933	0.166	0.482	.002	0.002
LTDCAP	0.001	0.659	0.117	0.123	.310	0.000
WOCAPL	−1.729	23.700	2.347	4.529	.004	0.000
TINEAR	−5.108	50.194	10.331	11.739	.083	0.000
CAITAL	0.039	2.544	0.564	0.483	.051	0.000
CURASS	0.135	0.932	0.536	0.181	.003	0.440
FIXASS	0.031	0.678	0.288	0.159	.000	0.000

Europe PMC Funders Author Manuscripts

Table 2

Variables in models

Variables for random forest	Variables for other models
TPEBIT	LOTCAG
PS	MANEXP
CURAST	PE
FASSTU	ACRESAL
TINEAR	CURAST
DEQUTY	WORCAP
CURASS	LTDCAP
FIXASS	

Europe PMC Funders Author Manuscripts

Table 3

Indexed variables

Variables	DAccuracy	DGini	Nfraud	Fraud
TPEBIT	18.91	14.81	5.57	19.04
PS	14.27	12.57	7.99	12.52
CURAST	18.96	17.22	10.41	16.49
FASSTU	18.37	15.58	3.58	20.16
TINEAR	14.52	13.08	7.41	12.9
DEQUTY	48	48.61	31.97	46
CURASS	13.24	11.45	7.37	11.22
FIXASS	16.86	14.31	9.68	13.94

Europe PMC Funders Author Manuscripts

Table 5

Results of model test

Models	Fraud (%)	Nfraud (%)	Total (%)
Logistic	42.03	23.18	32.18
KNN	88.41	87.50	87.92
DT	77.54	83.13	80.54
SVM	66.67	80.00	73.83

Europe PMC Funders Author Manuscripts

Table 6

Five-fold cross-validation results

Models	Nfraud (%)	Fraud (%)	Total (%)
Logistic	37.50	49.01	42.91
KNN	59.00	63.19	60.11
DT	68.13	64.62	66.43
SVM	81.88	78.13	80.18
RF	85.16	90.71	88.00