Probabilistic modeling and inference applied to company bankruptcies

Lorenzo Rossi
University of Milan, course of Probabilistic Modeling

28/06/2022

Abstract

The paper aims to study the causes of company bankruptcies implementing probabilistic modeling algorithms such as mixed graphical models in order to study causality and to make inference on the factors that lead to bankruptcy. The conclusion is that Probabilistic modeling techniques are as performing as other Supervised Learning algorithms, generally leading to the same results.

Introduction

Understanding cause-effect relationships between variables and being able to identify the most important factors in estimating bankruptcies can yield valuable information. Usually, experimental intervention is used to find these relationships, as what decides the default of a company can differ across economic systems, environments and human actions. As such, this knowledge can be put to further use upon implementing different models for the actual prediction of bankruptcies. This paper will focus on analyzing the impact of economic and financial factors at the firm level on bankruptcy risk. Since it is very unlikely that only one factor may determine the default of a company, the choice of a framework in which it is possible to assess the interactions among variable is preferred. For this purpose probabilistic modeling algorithms represent a valid choice of framework. Such models models are used for representing complex domains, conditional independencies and joint multivariate probability distributions through graphs. The paper is structured as such: in the first part there will be a brief literature overview about company bankruptcies and theoretical background for probabilistic models. The second part is dedicated to the choice of probabilistic models and the results obtained through inference on the chosen data set. Finally, the results of this approach will be confront it with other supervised learning algorithm in order to evaluate the performances of both approaches.

Dataset

The dataset comes from the notorious Kaggle dataset Company Bankruptcy Prediction: Bankruptcy data from the Taiwan Economic Journal for the years 1999–2009. As stated by the authors, data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. However, the features selected were 30 out of the original 95. This was done by hand and taking account of the economic literature, leading to the selection of a subset of variables. In detai:

Y - Bankrupt?: Class label X1 - ROA(C) before interest and depreciation before interest: Return On Total Assets(C) X2 - ROA(A) before interest and % after tax: Return On Total Assets(A) X3 - ROA(B) before interest and depreciation after tax: Return On Total Assets(B) X4 - Operating Gross Margin: Gross Profit/Net Sales X5 - Realized Sales Gross Margin: Realized Gross Profit/Net Sales X13 - Cash flow rate: Cash Flow from Operating/Current Liabilities X15 - Tax rate (A): Effective Tax Rate X16 - Net Value Per Share (B): Book Value Per Share(B) X17 - Net Value Per Share (A): Book Value Per Share(A) X18 - Net Value Per Share (C): Book Value Per Share(C) X19 - Persistent EPS in the Last Four Seasons: EPS-Net Income X20 - Cash Flow Per Share X22 - Operating Profit Per Share (Yuan ¥): Operating Income Per Share X23 - Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share X37 - Debt ratio %: Liability/Total Assets

X38 - Net worth/Assets: Equity/Total Assets X42 - Operating profit/Paid-in capital: Operating Income/Capital X43 - Net profit before tax/Paid-in capital: Pretax Income/Capital X52 - Operating profit per person: Operation Income Per Employee X54 - Working Capital to Total Assets X57 - Cash/Total Assets X59 - Cash/Current Liability X60 - Current Liability to Assets X61 - Operating Funds to Liability X68 - Retained Earnings to Total Assets X70 - Total expense/Assets X82 - CFO to Assets X84 - Current Liability to Current Assets X86 - Net Income to Total Assets X89 - Gross Profit to Sales X95 - Equity to Liability

Literature Overview for company bankruptcies

There's a rich literature in microeconomics research about firm defaults. The aim is to develop the models based on a combination of these features, and confront them with the results obtained by the algorithms. The concept of "failure" can vary from the narrow definition of bankruptcy or permanent insolvency to simply non-achievement of goals (Cochran, 1981; Pretorius, 2009). Altman constructed a model to predict bankruptcy with a multiple discriminant analysis finding that profitability, liquidity and solvency were the most significant factors in predicting bankruptcy (Altman, 1968). Ohlson explains four different factors he found statistically significant in affecting the probability of failure: the size of the company, the state of financial structures, performance and liquidity (Ohlson, 1980). Various empirical studies (Baldwin et al., 1997) tend to support the theoretical assumption that firm failure is rarely caused by only one cause or source. The most common factor in literature and evidence is that more than one causes or variables are taken into consideration when investigating bankruptcy.

Probabilistic Modeling

Theoretical Framework

The main goal of probabilistic models is to capture conditional independence relationships between interacting random variables. Moreover, by being aware of the graph structure of a PGM, one can solve tasks such as inference.

The starting point is to consider

$$p(x_1, x_2, ..., x_n)$$

as a probability distribution that can be decomposed in

$$p(x_{1j}) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)...p(x_j|x_{1j-1})$$

through the chain rule, which expresses the probability of interceptions through conditional probabilities, much similar to the Bayes rule:

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)}$$

Two events α and β are independent given a third event $\gamma(\alpha \perp\!\!\!\perp \beta|\gamma)$ if $P(\alpha|\beta \cap \gamma) = P(\alpha|\gamma)$ and vice-versa $P(\beta|\alpha \cap \gamma) = P(\beta|\gamma)$ Which means that knowing γ makes β irrelevant for predicting α (the same is valid for α and β inverted)

Following these statements and the necessary mathematical iterations, the Conditional Density Function is met:

$$f_{1|2}(x_1|x_2) = \frac{f_{12}(x_1, x_2)}{f_2(x2)}$$

Mixed Interactions Models

The dataset in this work is a case where there are both discrete and continuous variables with 1 discrete variable and 30 continuous variables. In the literature they are called Mixed Interaction Models, which are models for qualitative and quantitative variables that combine log-linear models for discrete variables with graphical Gaussian models for continuous variables. Moreover, a MIM is called "Homogeneous" if the covariance matrix of the Gaussian variables does not depend on the values of the discrete variables. In the case for a MIM, the following density has to be considered:

$$f(i,y) = p(i)(2\pi)^{-q/2} det(\Sigma)^{-1/2} exp[-\frac{1}{2}(y-\mu(i))^T \Sigma - 1(y-\mu(i))$$

which belongs to the exponential family:

$$f(i,y) = exp[g(i) + h(i)^T y - \frac{1}{2} y^T K y]$$

where K is the concentration matrix (a symmetric positive $q \times q$ matrix), $g(i)$, $h(i)$ are the log-linear expansion of the probability $p(i)$ (Canonical parameters).For a fixed i, $g(i)$ is a value and $h(i)$ is a q vector. Note that it is possible to impose conditional independence among variables by setting their interactions to zero. The dimension of a MI-model is the number of canonical parameters composing $g(i)$ and $h(i)$ adding to the number of free elements of the covariance matrix less 1. Under M, D is asymptotically χ^2(k) where k is the difference in dimension between the saturated model and M. The minimal sufficient statistics for a generator (a, b), where $a \subseteq \delta$ and $b \subseteq \gamma$ are: marginal frequency, total of variables, sub-matrix of the sum of squares matrix.

Models and results

First of all, given the enormous disproportion between the number of bankruptcies and non-bankruptcies, the number of the former was increased through the procedure of oversampling performed with the ROSE package of R, building a database with the same number of observation for the target variable (Bankrupt), with a proportion of 1:1. Furthermore, the continuous variables were standardized.

The analysis with the graphical models was started by inspecting the relations between continuous variables in the dataset. For this purpose, the GLASSO model was implemented through the relative *glasso* package. It gives a fast technique to find the Gaussian graphical model that maximizes a log-likelihood for K which is penalized by the $L1 - norm|K|$. Note that in this first part of the analysis, the target variable was not considered in order to visualize links and dependencies among the explanatory variables. After different tests, the model that best matches the interaction of the variables with respect to economic theory without an extreme penalization is the model with the value of ρ (the one that penalizes further connections) equal to 0.4 (Fig.1). The model presents a graph with four clusters of variables: the main cluster collects all those variables that interact with net revenues and expenses; a second cluster is related only to gross revenues; another one is made by the variables related to asset values of the company; the final cluster presents the cash-flow variables. This graph is useful for understanding the main groups of variables and how the interactions within them. The target variable will now be included again in the analysis.

The next two algorithm for graphical representation of the interactions among variables are the *minForest()* algorithm from *gRapHD* and *stepw()* from *gRim* packages respectively. The first implements the minForest model which returns the tree or forest that minimizes the $-2log - likelihood$, AIC, or BIC. From the direct connections of the target variable in the plot (Fig.2), it is possible to observe the link between Tax Rate, Persistent Earning per Share (EPS) and the Net Value of the assets of the company. The *stepw()* algorithm perform stepwise selection, and it shows many more variables linked to the target (Fig.3). A part from the previous variables, now it the target variable presents connections with Debt, Return On Total Assets and Net Profits. All of these interactions find proof in the economic theory and in the aforementioned literature.

Finally, the *mgm()* algorithm from the its own package was implemented using nodewise regression. The k parameter was set equal to three, and a cross-validation (CV) with ten folds was considered. The result (Fig.4) maintains the same connections as the previous plot but it highlights four specific variables: Debt Ratio, Net Worth assets, Working Capital and Cash On Total Assets.

Also in this case, the connected variables can fall into the categories of factors that the literature has shown to be relevant in predicting company bankruptcy, as they are all linked to solvency, profitability and liquidity. The previous model brought solid results. The next step is to evaluate numerically the performance of the previous algorithm in terms of confusion matrix and ROC curve. In the next plots it is possible to check the number of the false positives and negatives estimated by the algorithm in the confusion matrix as well as an accuracy score. The MGM obtained an accuracy score of the 84% (Table 1). The MGM algorithm obtained a moderately high accuracy and created fair graphical representation of the interactions among variables.

Comparison with other learning algorithms

The next part of the analysis will be to confront its result with other "classic" Machine Learning algorithms. For this comparison the following models were chosen:

- Logistic Regression
- Random Forest Classifier
- Boosting

Logistic Regression

The regression output (Table 2) reveals the high significance of certain variable to be the ones highlighted by the graphical model connections as well. As expected by economic literature, many of them affect either positively or negatively the probability of a bankruptcy (i.e. higher debt risks to lead to bankruptcy, while higher net worth assets lowers that possibility etc.). However, the model obtain a slightly lower accuracy than the MGM (Table 1).

Random Forest

The RF model was trained using 1000 trees. The plot (Fig.5) shows the relevance of the features in determining the classification task. It is clear that the first two variables outclass the other for importance in explaining the classification output. The two are followed by Debt Ratio and Return On Total Assets. However, most of the relevant variables are not changed if compared to the previous algorithms. On the contrary, certain ones maintain constant their high influence on the target variable (i.e. Net Worth Assets and Debt Ratio). The confusion matrix for the RF shows that the model obtains an higher accuracy (almost 90%) than the MGM with less false negative predictions (Table 1).

Boosting

Boosting is another approach to improve the predictions resulting from a decision tree. Like bagging and random forests, it is a general approach that can be applied to many statistical learning methods for regression or classification. Boosting builds lots of smaller trees. Unlike random forests, each new tree in boosting tries to patch up the deficiencies of the current ensemble. It's a sequential process in which each next model which is generated is added so as to improve a bit from the previous model. The output (Fig.6 and Table 3) is similar to the RF and it shows how much each variables "explains" the variance of the target variable (i.e. the percentage of the variance that changes at different levels of the explanatory variable): the first three variables remain the most influential, while the other variables in the output show slightly different levels of importance. The accuracy of the Boosting is higher than MGM and Random Forest (Table 1).

Conclusions

This paper focused on analyzing the impact of economic and financial factors that determine company bankruptcies through a probabilistic modeling framework. The results show that net revenue, asset values, level of debt, and productivity are the main factors that mostly determine the occurrence of this event. Algorithms that allow for graphical representation were implemented in order to find and visualize the connections and the independencies of the target variable (Bankruptcy). The MGM model was used to make inference, discover the most relevant explanatory variables

and make prediction, obtaining a moderately high accuracy. The results were confronted with the accuracy of other Machine Learning algorithms, reaching the conclusion that the MGM performed as well as other models (Table 1 and Fig.7). If the MGM gave a clear visual representation of the interactions and the links between Bankruptcy and other variables, the other models tried to show which variables had the most influence on the target. This represents two different approaches with which making inference.

Bibliography

https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction

Altman, E. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4), p.589.

Altman, E. I., & Hotchkiss, E. (2010). Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt (Vol. 289). Hoboken, NJ: John Wiley & Sons.

Avenhuis, J (2013) Testing the generalizability of the bankruptcy prediction models of Altman, Ohlson and Zmijewski for Dutch listed companies. Netherlands: University of Twente

Cochran, A. B. (1981). Small business mortality rates: a review of the literature. Journal of Small Business Management, 19(4), 50–59.

Pretorius, M. (2009). Defining business decline, failure and turnaround: a content analysis. Southern African Journal of Entrepreneurship and Small Business Management, 2(1), 1–16.

Ohlson, J. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18(1), p.109

Appendix

Tables

Table 1: Models

Indicators	MGM	Logit	Random Forest	Boosting
Accuracy	0.8415	0.8377	0.891	0.8939
95% CI	(0.8326, 0.8501)	(0.8209, 0.8534)	(0.8766, 0.9041)	(0.8797, 0.9069)
No Information Rate	0.555	0.5051	0.5051	0.5051
P-Value [Acc > NIR]	< 2.2e-16	<2e-16	< 2.2e-16	<2e-16
Kappa	0.6826	0.6753	0.782	0.7878
Mcnemar's Test P-Value	< 2.2e-16	0.9562	0.002067	0.5871
Sensitivity	0.8114	0.8383	0.8693	0.8906
Specificity	0.8791	0.8370	0.9130	0.8972
Pos Pred Value	0.8937	0.8400	0.9108	0.8984
Neg Pred Value	0.7883	0.8353	0.8725	0.8893
Prevalence	0.5559	0.5051	0.5051	0.5051
Detection Rate	0.4511	0.4235	0.4391	0.4499
Detection Prevalence	0.5048	0.5042	0.4822	0.5007
Balanced Accuracy	0.8453	0.8376	0.8912	0.8939

Table 2: Logistic Regression Estimates				
Coefficients	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.994378	0.063044	-31.635	< 2e-16 ***
ROA.C..before.interest.and.depreciation.before.interest	-0.187664	0.035968	-5.218	1.81e-07 ***
ROA.A..before.interest.and...after.tax	-0.091707	0.031266	-2.933	0.003356 **
ROA.B..before.interest.and.depreciation.after.tax	-0.172696	0.033943	-5.088	3.62e-07 ***
Operating.Gross.Margin	0.011738	0.042679	0.275	0.783297
Realized.Sales.Gross.Margin	0.039153	0.042537	0.920	0.357348
Cash.flow.rate	-0.036411	0.041412	-0.879	0.379267
Tax.rate..A.	-0.050766	0.029401	-1.727	0.084227
Net.Value.Per.Share..B.	-0.110229	0.051668	-2.133	0.032891 *
Net.Value.Per.Share..A.	-0.150188	0.052368	-2.868	0.004131 **
Net.Value.Per.Share..C.	-0.163826	0.051240	-3.197	0.001388 **
Persistent.EPS.in.the.Last.Four.Seasons	-0.267095	0.045489	-5.872	4.32e-09 ***
Cash.Flow.Per.Share	-0.139830	0.043221	-3.235	0.001215 **
Operating.Profit.Per.Share..Yuan.Â..	-0.147633	0.051732	-2.854	0.004320 **
Per.Share.Net.profit.before.tax..Yuan.Â..	-0.226875	0.047207	-4.806	1.54e-06 ***
Debt.ratio..	0.334091	0.038555	8.665	< 2e-16 ***
Net.worth.Assets	-0.451235	0.039632	-11.386	< 2e-16 ***
Operating.profit.Paid.in.capital	-0.206815	0.052507	-3.939	8.19e-05 ***
Net.profit.before.tax.Paid.in.capital	-0.197898	0.046751	-4.233	2.31e-05 ***
Operating.profit.per.person	-0.096408	0.041001	-2.351	0.018706 *
Working.Capital.to.Total.Assets	-0.248542	0.034814	-7.139	9.39e-13 ***
Cash.Total.Assets	-0.421486	0.047734	-8.830	< 2e-16 ***
Current.Liability.to.Assets	0.115648	0.031864	3.629	0.000284 ***
Operating.Funds.to.Liability	-0.003648	0.039242	-0.093	0.925938
Retained.Earnings.to.Total.Assets	-0.040097	0.031685	-1.265	0.205692
Total.expense.Assets	-0.076020	0.022847	-3.327	0.000877 ***
CFO.to.Assets	0.039336	0.034601	1.137	0.255609
Current.Liability.to.Current.Assets	0.085559	0.027614	3.098	0.001946 **
Net.Income.to.Total.Assets	-0.060468	0.026802	-2.256	0.024066 *
Gross.Profit.to.Sales	-0.024991	0.042859	-0.583	0.559817
Equity.to.Liability	0.093894	0.027836	3.373	0.000743 ***
Signif. codes: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; 0.1 ' '				

Table 3: Boosting: influence on Variance	
Variable	rel.inf
Net.worth.Assets	19.26205824
Net.Income.to.Total.Assets	14.45098592
Debt.ratio..	7.21390807
Cash.Total.Assets	6.49575171
ROA.B..before.interest.and.depreciation.after.tax	6.37390688
Current.Liability.to.Current.Assets	5.55060702
Total.expense.Assets	5.51866230
ROA.A..before.interest.and...after.tax	5.43980197
ROA.C..before.interest.and.depreciation.before.interest	5.28973574
Operating.Profit.Per.Share..Yuan.Â..	3.69537791
Operating.profit.Paid.in.capital	3.03984054
Retained.Earnings.to.Total.Assets	2.21416630
Persistent.EPS.in.the.Last.Four.Seasons	2.05142004
Net.Value.Per.Share..C.	1.60513111
Net.Value.Per.Share..A.	1.55766746
Net.Value.Per.Share..B.	1.51436637
Net.profit.before.tax.Paid.in.capital	1.36887330
Working.Capital.to.Total.Assets	1.36305971
Tax.rate..A.	1.03998436
Operating.profit.per.person	0.81853681
Per.Share.Net.profit.before.tax..Yuan.Â..	0.78854754
Cash.flow.rate	0.72699808
Equity.to.Liability	0.57075199
Current.Liability.to.Assets	0.45206189
Cash.Flow.Per.Share	0.43720829
CFO.to.Assets	0.39632424
Operating.Funds.to.Liability	0.26898213
Operating.Gross.Margin	0.26026484
Realized.Sales.Gross.Margin	0.15271522
Gross.Profit.to.Sales	0.08230403

Images

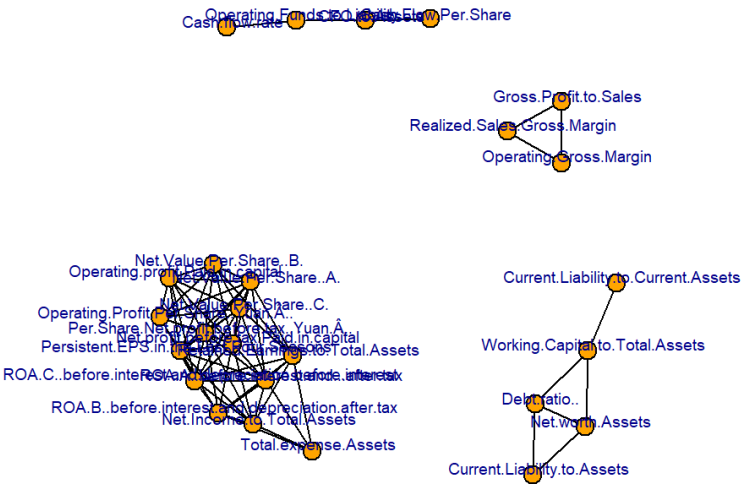


Figure 1: GLASSO: $\rho = 0.4$

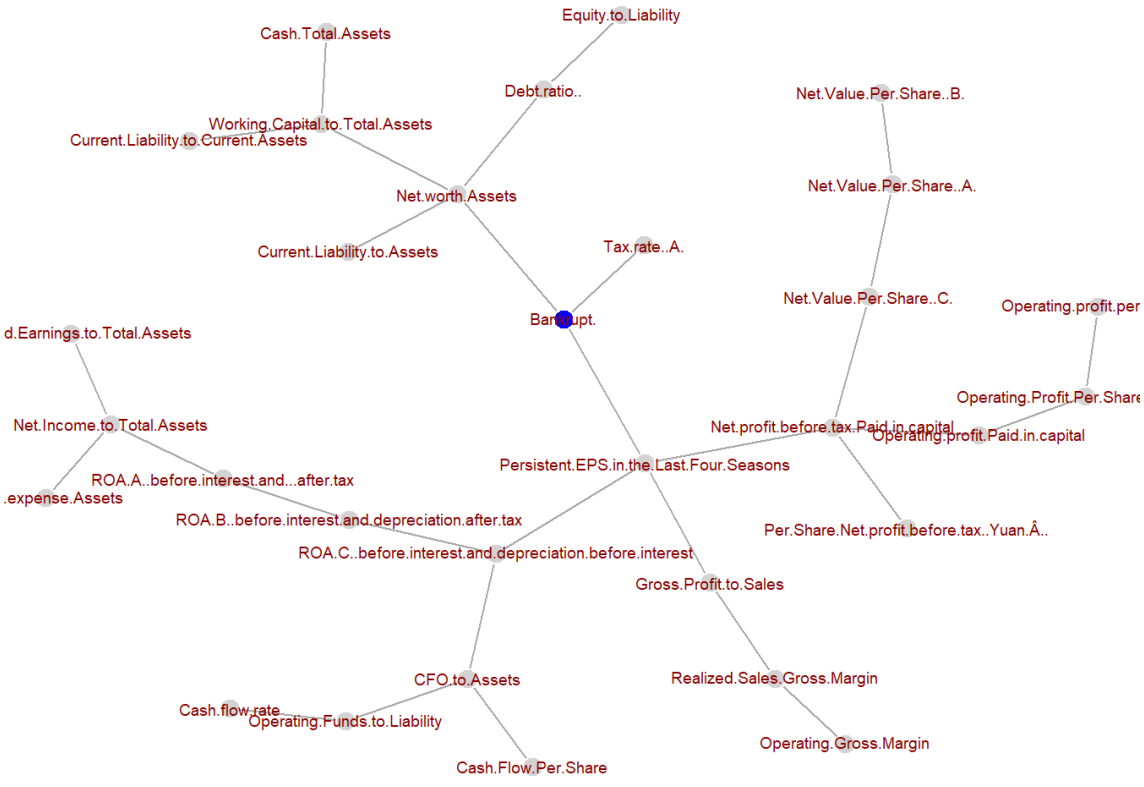


Figure 2: minForest: plot

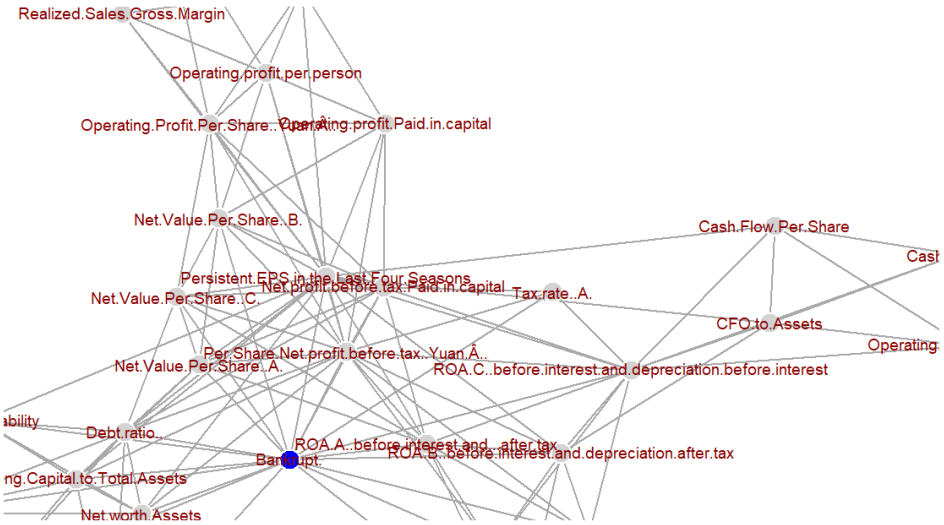


Figure 3: Stepwise selection: plot

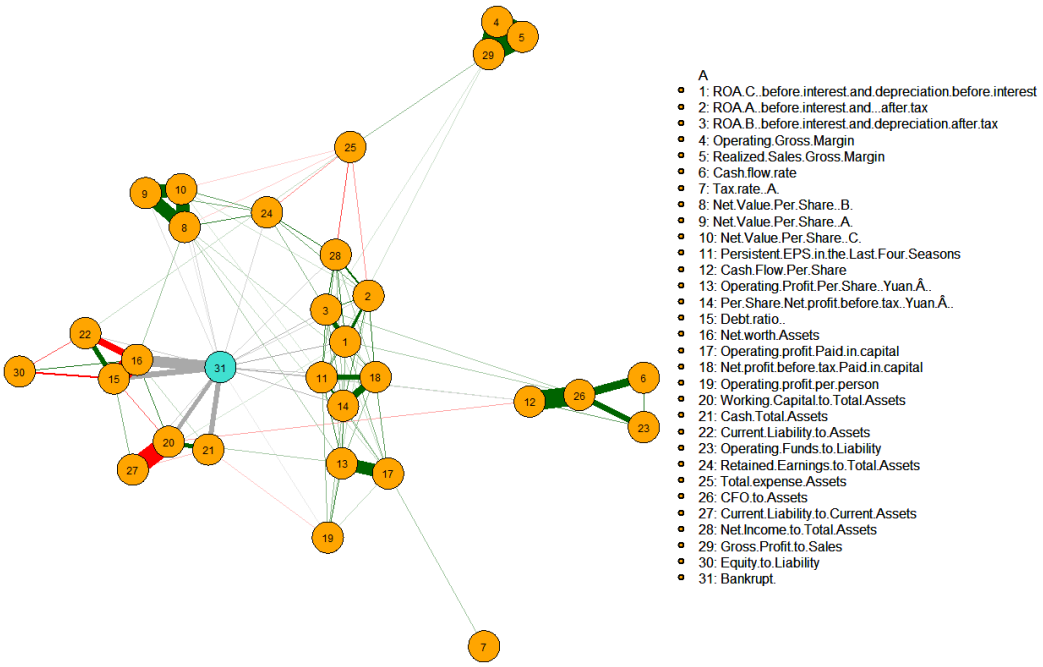


Figure 4: MGM: plot

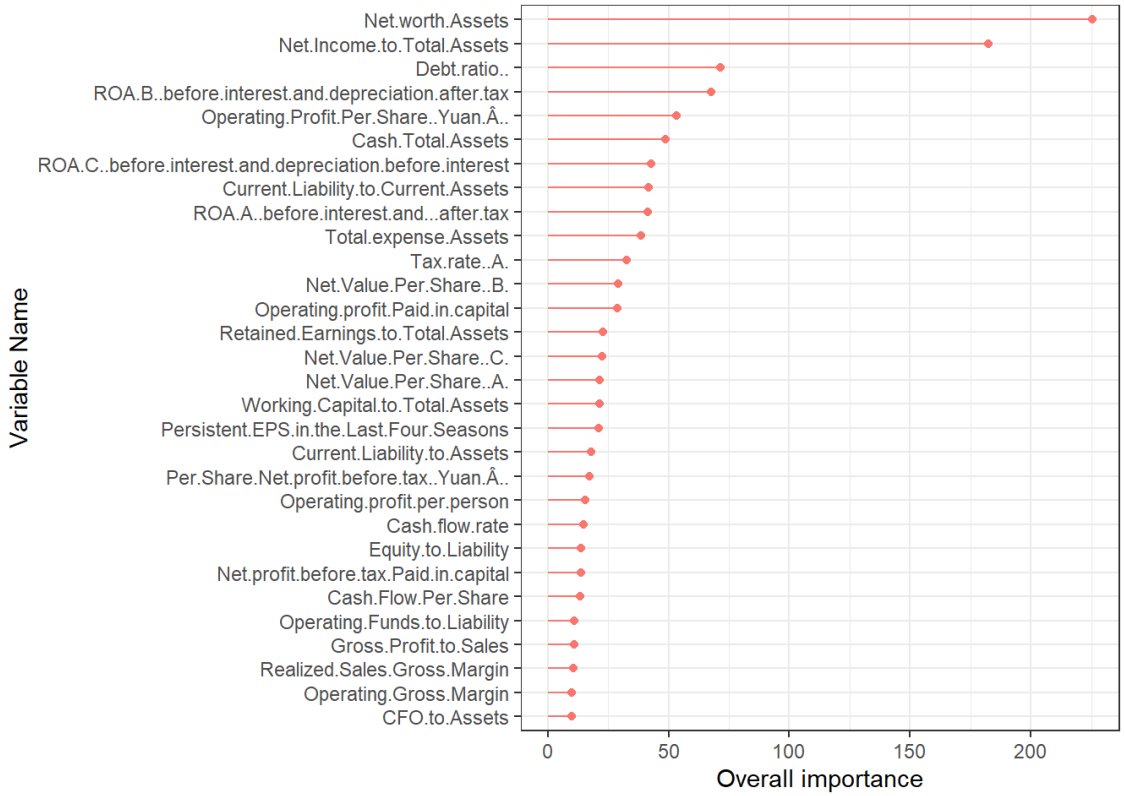


Figure 5: Random Forest: most relevant variables

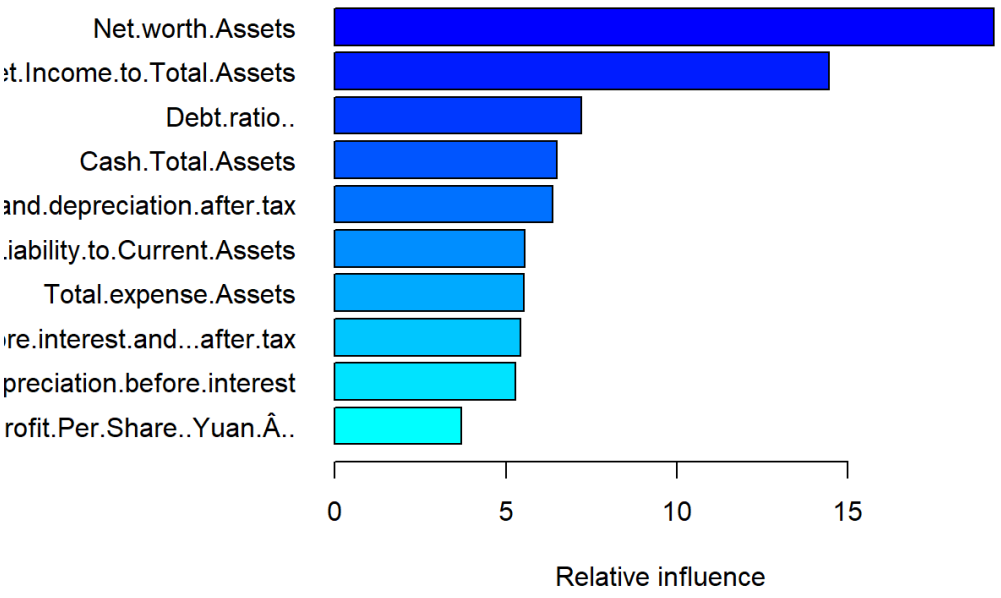
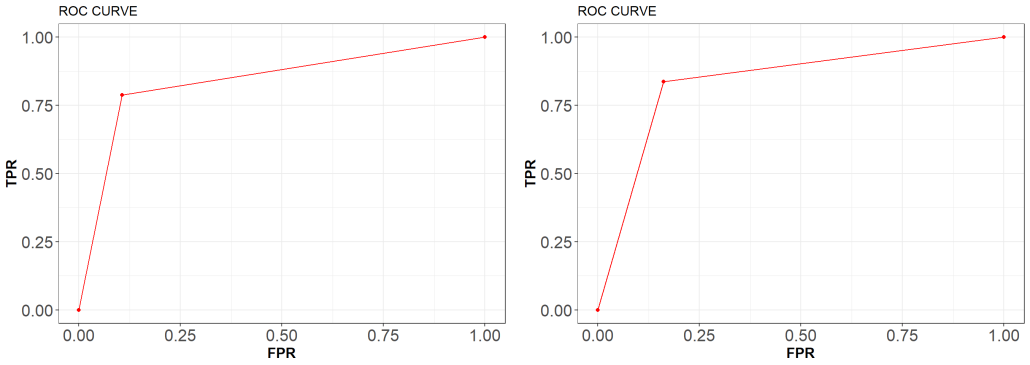
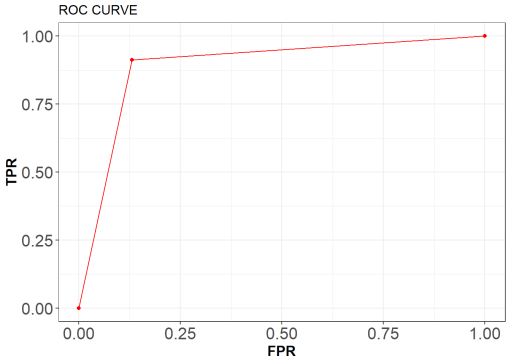


Figure 6: Boosting: influence on Variance

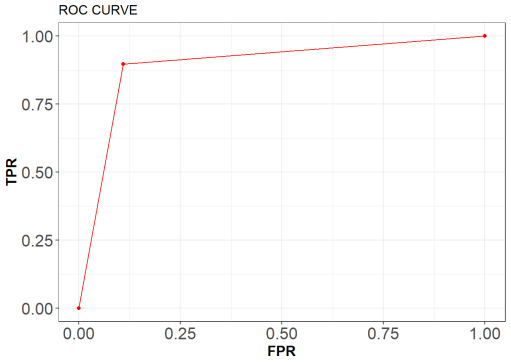


(a) MGM

(b) Logit



(c) Random Forest



(d) Boosting

Figure 7: ROC curves of the models