A two-stage Bayesian network model for corporate bankruptcy prediction

Yi Cao* Xiaoquan Liu† Jia Zhai‡ Shan Hua§

June, 2020

Abstract

We develop a Bayesian network (LASSO-BN) model for firm bankruptcy prediction. We select financial ratios via the Least Absolute Shrinkage Selection Operator (LASSO), establish the BN topology, and estimate model parameters. Our empirical results, based on 32,344 US firms from 1961-2018, show that the LASSO-BN model outperforms most alternative methods except the deep neural network. Crucially, the model provides a clear interpretation of its internal functionality by describing the logic of how conditional default probabilities are obtained from selected variables. Thus our model represents a major step towards interpretable machine learning models with strong performance and is relevant to investors and policymakers.

JEL code: C63, F47.

Keywords: Bayesian network; LASSO; Accounting ratios; Sensitivity analysis; Interpretability analysis.

*Management Science and Business Economics Group, Business School, University of Edinburgh, 29 Buccleuch Place, Edinburgh EH8 9JS, UK. Email: jason.caoyi@gmail.com.

†Corresponding author. University of Nottingham Business School China, University of Nottingham Ningbo, Ningbo 315100, P. R. China. Email: xiaoquan.liu@nottingham.edu.cn. Phone: +86 574 88180000 ext 8207.

‡Business School, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu Province, China 215123. Email: jia.zhai1982@gmail.com..

§Surrey Business School, University of Surrey, Guildford, Surrey GU2 7XH, UK. Email: s.hua@surrey.ac.uk.

1 Introduction

Corporate bankruptcy is a serious issue in the financial market due to its damaging economic and social consequences. As a result, the academic community, financial industry, and regulators are keen to explore reasons behind and ways to predict and prevent it. In the literature, early studies such as Altman (1968), Ohlson (1980), and Zmijewski (1984) document that accounting ratios and stock market data contain valuable information for assessing firm financial health.

More recently, forecasting firm default probability has attracted a lot of attention in the financial technology literature as state-of-the-art computational methods allow us to develop models that evaluate default prediction with great precision (see Chen et al., 2019; Goldstein et al., 2019, for example). These include the logit model (Tian et al., 2015), the support vector machine model (Liang et al., 2016), the random forest (Chandra et al., 2009), and the deep neural network (Cerchiello et al., 2017). Empirical evidence suggests that default forecasting performance can be improved by selecting the most relevant variables via the least absolute shrinkage and selection operator (LASSO) (Tian et al., 2015); or including new heterogeneous features such as textual information (Mai et al., 2019); or employing complicated deep neural network models (Cerchiello et al., 2017), which consist of a number of layers, each armed with numerous hidden neurons, and exhibit strong capability in capturing the relationship between input variables and output bankruptcy forecasts.

Our paper is motivated by this strand of literature but its contribution lies in developing an *interpretable* machine learning model that not only performs well empirically but also reveals the mechanism through which bankruptcy forecasts are obtained from input variables, i.e., it paints a clear picture of model internal functionality. Our paper thus addresses a growing call for model interpretability in an age when increasingly sophisticated machine learning models and big data make the decision making process obscure. Kim and Doshi-Velez (2017) indicate that opening the black box is not about understanding *all bits and bytes of the model* but instead knowing the *logic of the internal functionality for the downstream conclusions*; whereas Mittelstadt et al. (2018) acknowledge the need for this but expresses concern that, with complicated internal states and millions of interdependent values, the black box is difficult to open up.

In this paper, we adopt the Bayesian network model, a powerful machine learning tool in handling uncertainty and multi-faceted relationship with a combination of domain knowledge and data-driven modeling (Liu et al., 2018). It has enjoyed great success in the healthcare diagnosis area in predicting the survival of the Alzhemer's disease, heart disease, breast cancer, and so forth

(Liu et al., 2018; Lu et al., 2016; Seixas et al., 2014). To the best of our knowledge, the Bayesian network has not been implemented in default probability prediction, an area similar in nature to that of healthcare diagnosis, making the Bayesian network an appropriate method for our purpose.

Methodologically, we perform the least absolute shrinkage and selection operator (LASSO) in the first stage to select the most relevant accounting and financial variables (Tian et al., 2015). In the second stage, we construct the Bayesian network structure from selected variables and estimate parameters for the conditional probability via the expectation-maximization (EM) algorithm. The same selected variables are also used in alternative models including the logistic regression, the decision tree, the support vector machine, and the deep neural network model in the empirical analyses. Our data contain quarterly COMPUSTAT accounting and financial information from January 1961 to August 2018 with 31 variables of 32,344 firms with more than 1.5 million firm-quarter observations in total.

Our empirical analyses reveal that the Bayesian network model achieves the second most accurate forecasts and is only outperformed by the complex deep neural network model with three hidden layers. More importantly, once we identify the dependence structure of the Bayesian network, we are able to explain clearly the way that the model arrives at conditional probability for default, and how the default probability varies upon changes with input variables. In this way, the Bayesian network is able to address *what-if* questions of an *ad-hoc* scenario, such as what could a firm do differently to achieve a better health status. This allows us to construct bankruptcy probability surface by changing input variables in company financial statements. In other words, we are able to gauge the sensitivity of conditional default probabilities with respect to variations in input variables.

Hence, our paper makes three contributions to the literature. First, as far as we are aware, this is the first study that balances the performance and interpretability of a machine learning model in predicting firm bankruptcy probability, as the existing data science literature is yet to embrace the interpretability issue. Second, given the clear internal functionality of the Bayesian network model, we are able to draw probability surfaces of variables of interest and perform sensitivity and scenario analyses to address *what-if* questions such as how bankruptcy probabilities change with regard to a particular input variable. We believe that this is the first *ad-hoc* scenario analysis in bankruptcy prediction. Finally, we offer solid empirical evidence that the Bayesian network model is a promising tool for predicting conditional default probability with precision. Our paper showcases a meaningful application of this powerful method and is relevant to investors, portfolio managers, and regulators. It also points to a promising avenue to which the Bayesian network can make substantial contribution.

The rest of this paper proceeds as follows. We undertake a review of relevant strands of the literature in Section 2. The two-stage Bayesian network model and other methodological issues are discussed in Section 3. In Section 4, we introduce the data, perform empirical analyses, and undertake robustness check. Finally, Section 5 concludes.

2 Literature review

In this section, we review relevant studies that focus on the use of machine learning models in predicting corporate default, on the Bayesian network model, and on the LASSO method.

Machine learning models

Since the seminal work of Altman (1968), Ohlson (1980), and Zmijewski (1984), predicting corporate bankruptcy has been a topical issue in the literature for a long time. Mai et al. (2019) conduct a recent review of this area and note that methodologically many studies focus on machine learning models due to their estimation precision. In Table 1, we provide a partial summary of studies in the past three years, all of which feature a model in the machine learning family including the logistic regression, decision tree, random forest, support vector machine, and deep neural network. We further classify them into two groups: the interpretable and non-interpretable ones according to Mittelstadt et al. (2018). Only the ***simple*** logistic regression and tree-based models, serving as benchmarks, can be considered interpretable models.

It is worth noting that Gogas et al. (2018) develop a geometric interpretable model and use it as a stress testing tool to visualize the classification space with two variables as well as the linear decision boundary. By calculating the distance of the data to the decision boundary and simulating certain scenarios, the tool provides an effective interpretation for the model results and partially answers the ***what-if*** question of changing the critical variable values. However, this study assumes the ***fail*** and ***alive*** cases are linearly separable by two selected variables, which is contrary to findings in most papers in Table 1.

Bayesian network model

The Bayesian network is able to capture the relationship and probability distribution to enhance the ontology inference capability in the diagnosis of a variety of diseases. Hence, it is often implemented in the healthcare diagnoses for medical ontology probabilistic inference and achieved via the *K2* greedy algorithm. Delen et al. (2019) perform the Bayesian network with the elastic net variable selection method in understanding and predicting prominent variables that determine student attrition and achieve an accuracy as high as 84%. However, no comparison is conducted between the Bayesian network and alternative approaches. Dag et al. (2016) use the Bayesian network to predict heart transplant survival. They adopt different selection methods to generate a set of potential predictors with medically relevant variables and construct the Bayesian network from selected predictors. The Bayesian network not only achieves similar predictive performance compared with the best-performing approaches in the literatures but also provides an interactive relation among the predictors and the conditional survival probability. Meanwhile, the Bayesian network is implemented in project management (Hu et al., 2013; Yet et al., 2016), cyber-security assessment (Zhang et al., 2018), and stock index forecasting (Malagrino et al., 2018). In a pioneer study, Sun and Shenoy (2007) use a Naïve Bayesian network to assess the bankruptcy. The bankruptcy predictors are selected by a heuristic method and a Naïve Bayesian network is constructed based on these predictors. However, the Naïve Bayesian network does not contain any topology or hierarchy logic among the predictors as it considers parallel impacts of all predictors to the output.

LASSO

Introduced by Tibshirani Tibshirani (1996), the LASSO is a powerful method for variable selection widely adopted in economics and finance. It is successfully implemented in the literature for predicting stock returns using intraday NYSE data (Chinco et al., 2019), corporate bankruptcy (Amendola et al., 2011; Tian et al., 2015), corporate bond recovery rate (Nazemi and Fabozzi, 2018), and macroeconomic time series (Bai and Ng, 2008; Kim and Swanson, 2014). Tian et al. (2015) apply the LASSO to forecast corporate bankruptcy in the US and achieve strong out-of-sample performance, whereas Rapach et al. (2013) show that the LASSO outperforms a backward or forward stepwise regression.

3 Bayesian network with Lasso

In this section, we first outline the Bayesian network with a simple illustration. We then introduce the LASSO selection method in our two-stage Bayesian network model. Alternative machine learning models are also discussed.

3.1 The Bayesian network

A Bayesian network is a directed graph that encodes the latent probabilistic relationship between variables of interest in a reasoning representation problem (Heckerman et al., 1995; Lauritzen, 1995). The representation usually starts from the domain knowledge, constructs a *prior network*, and combines it with the observed data to learn a new Bayesian network (Heckerman et al., 1995). In this framework, a variable is termed a node, vertice, or point. The nodes are connected by directed arrows indicating probabilistic dependencies.

To illustrate, we assume that bank failure is caused by two variables: Total Capital (C) and Risk-adjusted Capital Ratio (R) shown in Figure 1. The two arrows starting from the Total Capital (C) and the Risk-adjusted Capital Ratio R, respectively, to Bank Failure (F) suggest that F is dependent on R and C. Meanwhile, the arrow pointing from C to R indicates that the Risk-adjusted Ratio also depends on the Total Capital. In this example, Total Capital (C) and Risk-adjusted Capital Ratio (R) are *parent* variables of Bank Failure (F), and Total Capital (C) is also a *parent* of Risk-adjusted Capital Ratio (R).

The joint probability of Bank Failure (F), Total Capital (C), and Risk-Adjusted Capital Ratio (R) can be expressed as follows:

$$P(F, C, R) = P(R|F, C)P(C|F)P(F). \tag{1}$$

We are usually interested in addressing the following question: Given an observed Risk-Adjusted Capital Ratio (R), what is the probability of Bank Failure (F)? The answer can be evaluated by the conditional probability as follows:

$$P(F|R) = \frac{P(F, R)}{P(R)} = \frac{\sum_C P(F, C, R)}{\sum_C \sum_F P(F, C, R)}. \tag{2}$$

In general, a Bayesian network model can be summarized as follows. Suppose we have a domain of discrete variables $U = x_1, \ldots, x_n$, and a set of cases $D = C_1, \ldots, C_m$. Our main interest is in determining the joint probability distribution $p(C_{m+1}|D, \ )$, which is the probability of a new case C_{m+1} given the set of past observations D and current information . In Bayesian network models, we do not intend to recover the complete distribution. Instead, we assume that the distribution for data is generated from a latent structural network B_s with a number of unknown parameters.

Hence, the probability $p(C_{m+1}|D, \xi)$ with the structural network B_s can be expressed as follows:

$$p(C_{m+1}|D, \xi) = \sum_{all B_s} p(C_{m+1}|D, B_s, \xi) p(B_s|D, \xi). \tag{3}$$

The structural network B_s reflects our belief of the variables and the relationship between them based on domain knowledge. In most cases, however, B_s is unknown even though variables U and case observations D are available. Two methods are usually adopted to construct B_s: the domain expert heuristic method employed by Chakraborty et al. (2016); and the statistical structure learning algorithm used by Heckerman et al. (1995) and Liu et al. (2018). In this paper, we implement the structural learning method due to the lack of domain knowledge as in most practical cases.

To implement the statistical structure learning algorithm, our first step is *structure learning*, i.e., we identify the interactive relation between variables, specify the topology of the framework in order to construct a Bayesian network. Once we obtain the network, we determine the parameters of the network and define the joint probability representing the statistical behavior of observed data in *parameter learning* (Heckerman et al., 1995; Lauritzen, 1995; Liu et al., 2018).

Structure learning

Structure learning can be performed primarily in three ways: the search-score method, the constraint-learning method, and the dynamic programming based method. Among them, the search-score method is suitable for problems with large volume of data (Daly et al., 2011; Heckerman et al., 1995). In our paper, we construct our Bayesian network model with the popular *K2* search-score algorithm (Cooper and Herskovits, 1992; Feng et al., 2014; Garvey et al., 2015).

Specifically, we assume a domain of n discrete variables $U = x_1, \ldots, x_n$ with an ordering of variables, a set of cases $D = C_1, \ldots, C_m$, and an upper limit u on the number of parents a variable may have. The algorithm heuristically searches for the most appropriate belief-network structure based on D. In the initial stage, an empty π_i is created as a set of parents of variables x_i, $i = 1 \ldots n$. A function $pred(x_i)$ is defined to represent a set of variables preceding x_i. For each variable x_i in U, we calculate the score $P_{old} = f(i, \pi_i)$ as follows:

$$f(i, \pi_i) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} \alpha_{ijk}!, \tag{4}$$

where $q_i = |\phi_i|$, and ϕ_i is a list of all possible parents of x_i in D; $r_i = |V_i|$, and V_i lists all possible values of the variable x_i; $_{ijk}$ is the number of cases in D in which the variable x_i is instantiated with its kth value, and the parents of x_i in $_i$ are instantiated with the jth instantiation in $?_i$; $N_{ij} = \sum_{k=1}^{r_i} {}_{ijk}$ is the number of instances in the data where the parents of x_i in $_i$ are instantiated with the jth instantiation in $?_i$.

The $f(i;\ _i)$ is considered the probability of the case set D given that the parents of x_i are $_i$. When the number of variables in $_i$ is less than u, the variables x_m of $pred(x_i)$ will be iteratively added to $_i$. The probabilistic score is updated if x_m is added to the set $_i$. In this way, the $K2$ algorithm finds a network structure B_s with variables in U that each node in B_s exhibits at most u parents, such that the achieved probabilistic score metric is larger than a pre-defined real value of p.

Parameter learning

The parameters of the Bayesian network, $_{ijk}$, is the conditional probability distribution of the node X_i in U taking the kth value with its parent node $_i$ taking the jth value as follows:

$$_{ijk} = p(X_i = x_k j\ _i = j): \tag{5}$$

The parameters can be determined by the expectation-maximization (EM) algorithm, a well-known approach for estimating the maximum likelihood of the model with latent structure (Dempster et al., 1977; Green, 1990; Lauritzen, 1995). Green (1990) introduces an EM algorithm for estimating the penalized likelihood, which exhibits a more efficient convergence rate than the traditional EM algorithm. Following Green (1990) and Lauritzen (1995), we consider a log-likelihood function given the observed data as follows:

$$Q(\ ' j\ ) = E\ \log f(X j\ ')jy; \tag{6}$$

where X is the learned variable based on the complete data with the density function f, and y is the observed data. The EM algorithm features a recursive process of two steps: First, the expectation step (E-step) of fixing and calculating the expected value of $E\ log f(X j\ ')jy$; Second, the maximization step (M-step) of finding values that maximize likelihood $Q(\ ' j\ )$. At the E-step, we add a penalty $J(\ ')$ to the log-likelihood function following Green (1990) as follows:

$$Q^*(\ ' j\ ) = Q(\ ' j\ )\quad J(\ '); \tag{7}$$

where $J(\theta')$ is a function proportional to a prior density. The M-step maximizes the penalized log-likelihood function.

3.2 LASSO selection method

Following Tian et al. (2015), we estimate the LASSO parameters by minimizing the negative log-likelihood of discrete hazard function with a penalty for the sum of absolute value of covariate parameters. The discrete hazard function is given as follows:

$$P(Y_{i,t+N} = 1 | Y_{i,t+N-1} = 0, X_{i,t}) = \frac{e^{\beta_0 + \beta' X_{i,t}}}{1 + e^{\beta_0 + \beta' X_{i,t}}}, \tag{8}$$

where $X_{i,t}$ is a vector of time-varying predictive variables observed for quarter t, and i is the firm index. The variable $Y_{i,t+N}$ is the default label, which is equal to one if firm i files for bankruptcy protection at $t+N$ given that it survives $N-1$ quarters from time t to $t+N-1$. The negative log-likelihood function with a penalty of sum of the absolute value of the covariate parameters is specified as follows:

$$-\sum_{i=1}^{n}\left[Y_{i,t+N}(\beta_0 + \beta' X_{i,t}) + \log(1 + \exp(\beta_0 + \beta' X_{i,t}))\right] - \lambda \sum_{k=1}^{p} |\beta_k|, \tag{9}$$

where n is the number of firms and p is the number of predictive variables in the hazard model. Following Tibshirani (1996), we employ a ten-folder cross-validation for parameter estimation.

4 Data and empirical analyses

In this section, we first introduce our data. In the empirical analyses, we begin with discussing accounting and financial variables selected by the LASSO. We then make comparison of bankruptcy prediction accuracy between alternative models and highlight the interpretability of the models. Finally, we perform a subsample analysis to check the robustness of the baseline results with respect to the sample period.

4.1 Data

We use quarterly COMPUSTAT data from January 1961 to August 2018 for 31 accounting variables as candidate variables for bankruptcy prediction following the existing literature (see Amendola et al., 2011; Bharath and Shumway, 2008; Campbell et al., 2008; Ding et al., 2012; Liang et al., 2016; Mai et al., 2019, for example). When constructing the accounting-based predictors, we align a firm's fiscal year with the calendar year to ensure that the accounting information is observable to investors at the time of prediction. Because we use quarterly data, we lag all variables by a quarter. Furthermore, we remove variables at the top and bottom one percentile following Tian et al. (2015). Our final dataset contains 1,563,010 firm-quarter observations for 32,344 firms.[1]

The descriptive statistics of accounting variables is shown in Table 2. As some variables are scaled (such as current assets over current liabilities ACTLCT) whereas others are in monetary terms (such as total asset TASSET), the descriptive statistics varies to a large extent. Our bankruptcy indicator is based on the *Reason for Deletion* variable *dlrsn* in the COMPUSTAT. A firm is defaulted if it is de-listed from the stock exchange due to liquidation or bankruptcy and the default indicator is one; otherwise the indictor is zero. In total, we identify 16,924 bankruptcy and liquidation filings over the sample period. In Figure 2 we demonstrate the occurrence of default by year. We observe two clear peaks during 1982 to 1991 and from 2007 to 2008 of the Great Recession.

4.2 Alternative models

Among the most popular bankruptcy prediction models reviewed in Liang et al. (2016) and Lin et al. (2012), we employ the logistic regression (LR) and decision-tree (Tree), which are simple and interpretable models. We also implement the support vector machine (SVM) which is of modest complexity. The deep neural network (DNN) is selected as a complex model and we follow a standard specification DNN(50,30,20) with three hidden layers of 50, 30, and 20 hidden neurons, respectively (Goodfellow et al., 2016). Thus we include four models in addition to the Bayesian network model to assess their interpretability and prediction accuracy. They are applied to the same variables selected by the LASSO model.

All models are implemented via *R* packages. The logistic regression is based on the *ISLR* package; the classification and regression trees (CART) model of Huang (2014) is based on the *tree* package; the support vector machine (SVM) with radial kernel is based on the *e1071* package; and the deep neural network is based on the *H2o* package with H2o cloud computing backend. To train the CART and SVM models, a 10-folder cross-validation is applied to achieve a stable and optimal selection of parameters.

[1] We use the first 70% of data as the training set and remaining 30% as testing set but our empirical results remain qualitatively the same if we split the data 50:50 for training and testing sets. These results are available upon request from the authors.

Furthermore, the deep neural network is trained by the stochastic gradient descent algorithm with epochs of 200. To avoid overfitting, the traditional L2 regularization is applied to penalize the weights. The dropout method of Srivastava et al. (2014) is performed to randomly omit a subset of hidden neurons at each iteration of the training process. An early stopping suggested in Bengio (2012) is also implemented for monitoring the performance of the validation and training set and for stopping the training early when the performance of the training set keeps improving but validation set stops. The Bayesian network model is implemented by the *bnlearn* package; and the LASSO is performed by the *glmnet* package.

4.3 LASSO selection results

We identify the most relevant predictors via a 10-folder cross-validation LASSO to optimize the coefficient for each predictive variable. Selected variables are summarized in Table 3 Panel A. As we can see, 16 variables exhibit non-zero coefficients out of 31 potential predictive variables. This is more than the number of variables identified in Tian et al. (2015).

The selected variables mainly concern a firm's leverage, liquidity, profitability, and market based variables. First of all, the leverage ratio of total liability over total assets (LTAT) is the most influential with the largest coefficient. This is in line with Campbell et al. (2008), Ohlson (1980), and Zmijewski (1984). The other leverage ratio chosen by the LASSO is a book leverage measure of total debts over total assets (FAT), also chosen in Tian et al. (2015), which contains information about future default risk. We notice that the market leverage measure of total liabilities over the sum of market equity and total liabilities (LTMTA), which is heavily influenced by stock prices, is not selected. Hence, the book leverage ratio may convey more information than the market leverage measure.

The relevance of market based variables is eloquently argued in Campbell et al. (2008), which suggest that the logarithmic market capitalization (RSIZE) and logarithmic stock price (PRICE) are important. As Tian et al. (2015) point out, the information conveyed in PRICE is forward looking, whereas RSIZE reflects the true value of a firm. These two variables exhibit the second and third largest coefficient in the LASSO selection results.

Six liquidity ratios are chosen, including growth of inventories over inventories (INVCHINVT), working capital over total assets (WCAPAT), current liabilities over total liabilities (LCTLT), current assets over current liabilities (ACTLCT), current liabilities over total assets (LCTAT), and cash and short-term investment over total assets (CASHAT). A lack of liquidity is more likely to increase default risk rather than causing bankruptcy directly. As a result, the selection of ACTLCT (current ratio) and WCAPAT (working capital turnover) is consistently with previous research (see Chava and Jarrow, 2004; Ohlson, 1980; Shumway, 2001, for example). Furthermore, the inventory variable (INVCHINVT), the percentage of current liability (LCTLT), the current liability coverage (LCTAT), and a cash and short-term investment variable (CASHAT) all capture different aspect of a firm's liquidity.

Also essential in predicting bankruptcy are profitability ratios. Retained earnings over total assets (REAT) receives the most attention. This choice is consistent with Altman (1968), Chava and Jarrow (2004), and Shumway (2001), all of which show the impact of cumulative profitability on reducing the bankruptcy probability. The other four profitability measures, retained earnings over current liabilities (RELCT), net income over total assets (NIAT), net income over the total of market equity and total liabilities (NIMTA), and operating income over sales (OIADPSALE), imply that cumulative and current period profitability help reduce bankruptcy risk, but to a lesser degree.

In addition to these major accounting and financial ratios, the LASSO also picks sales over total assets (SALEAT): the higher the sales turnover, the lower the bankruptcy risk (Altman, 1968; Shumway, 2001).

4.4 Prediction accuracy

To provide a comprehensive analysis, our bankruptcy prediction horizon ranges from one to 12 quarters ahead. We use two popular measures: the accuracy ratio (ACCU) and the area under the *Receiver Operating Characteristic (ROC)* curve (AUC), to evaluate the performance of alternative models following Liang et al. (2016), Mai et al. (2019), and Tian et al. (2015). The ACCU is based on the cumulative accuracy profile (CAP) that measures the percentage of true bankrupt firms included if choosing a different percentage of observations using the sorted forecasted probabilities generated by a used model (Engelmann et al., 2003; Mai et al., 2019). The baseline model assigns class labels randomly. The accuracy ratio of a forecasting model is the difference in the area between the CAP of the model and that of a baseline model. The AUC is an equally popular measure of the overall performance of a model. It is calculated by the ROC curve, which shows the capability of a model balancing the false positive rate and the true positive rate. The area under the ROC curve provides a measure of the capability of the overall performance and the corresponding robustness of the model.

In the empirical analyses, we examine the forecasting performance of alternative models based on two groups of prediction variables. In Group 1, we follow the LASSO result reported in Table 3 and use all 16 selected variables; whereas in Group 2, we only use the ten top-ranked variables. This is because the absolute value of LASSO coefficients for variables ranked from 11th to 16th is lower than 1E-5, much closer to zero than the top 10 ranked variables. Furthermore, a model with fewer variables affords better interpretation as only the most relevant variables are included.

Panels A and B in Table 4 contains the ACCU and AUC for bankruptcy predictions over one to 12 quarters ahead generated by the Bayesian network model and alternative ones via, respectively, Group 1 and Group 2 variables. In Panel A, we note that the AUC values of all models are above 0.76 over forecasting horizons of up to one year. The DNN(50,30,20) and the Bayesian network model perform the best at 0.9003 and 0.8951, respectively, over the one-quarter horizon. Over longer forecasting horizons, the prediction accuracies gradually decrease; whereas across the models the prediction accuracy increases from the more traditional models such as the logistic regression to the state-of-the-art large scaled neural network. The DNN(50,30,20) exhibits the best performance in terms of AUC and ACCU while the Bayesian network model comes second. It is interesting to note that the Bayesian network model easily beats the other three less sophisticated ones. The logistic regression and decision tree model exhibit the lowest accuracy and smallest ACCU and AUC.

In Panel B, with fewer variables and less information to draw upon, the prediction accuracy drops for all models. However, we still find that, similar to results in Panel A, increasing model complexity leads to more accurate bankruptcy prediction. Interestingly, over longer forecasting horizons the Bayesian network seems to make a better use of the information content of variables and tends to outperform the DNN(50,30,20). For example, over 10 to 12 quarters ahead, the AUC for the Bayesian network model is 0.7416, 0.7289, and 0.7259, respectively, and the corresponding AUC for the DNN(50,30,20) is marginally lower at 0.7414, 0.7268, and 0.7257. Similar patterns hold for the ACCU over the longest forecasting horizons.

4.5 Model interpretability

Interpretability refers to the transparency of a model's internal function and the degree of human comprehensibility (Doshi-Velez and Kim, 2017; Mittelstadt et al., 2018). Recently, Bastani et al. (2017) and Ribeiro et al. (2016) pursue two classes of *approximate* models such as linear models and the decision tree type of models precisely because of model interpretability. In Table 4, approximate models such as the logistic regression and decision, whose internal functions can be easily interpreted by their structures and corresponding parameters, fare poorly empirically. Meanwhile, more complex models such as the support vector machine and deep neural network model do not offer easily observable or comprehensible structures but perform well empirically. For the Bayesian network model, it exhibits smaller scale and simpler structure than the DNN model with comparable forecasting accuracy. More importantly, it sheds light on interpreting the internal reasoning logic with a structural network. Below we scrutinize model interpretability in detail based on their one-year ahead forecasting performance.

Logistic regression

It is fairly easy to interpret the logistic regression model by looking at variable coefficients and their statistical significance in Table 5. For example, the ratio between firm net income over market equity and total liabilities (NIMTA) and the scaled market capitalization (RSIZE) exhibit the largest coefficient at 0:283 and 0:171, respectively, and are both highly significant at the 1% level. Hence, these two variables are the most influential in determining future bankruptcy than other variables; and they suggest that the larger the relative net income and market capitalization of a firm, the lower the probability of default.

Decision tree

Figure 3 shows the decision tree for forecasting bankruptcy over the next year using all 16 LASSO identified variables. We observe that the model constructs the decision tree with only 12 variables, including OIADPSALE, INVTSALE, QALCT, RELCT, LTAT, PRICE, FAT, CASHMTA, REAT, NIAT, TSALE, and APSALE. The decision tree provides a graphical and self-explainable interpretation of the internal function. However, this simplistic structure yields an average AUC of 0.7581 and accuracy ratio of 0.7431, which is among the lowest across alternative models.

Although the structure and coefficients of the above two models offer a transparent functionality and can be relatively easily comprehended in their entirety, the models show a lack of accuracy in handling the prediction problem thus making them a *local approximation* and represent a partial or a *slice* of the entire problem (Mittelstadt et al., 2018).

Bayesian network

Figure 4 illustrates two structures of the Bayesian network for forecasting bankruptcy four quarters ahead based on Group 1 and Group 2 variables. In Figure 4(a), the complex interleaved arrows show the inter-dependence of 16 variables and *dlrsn*, the state of bankruptcy. We see that bankruptcy is directly determined by only eight crucial variables: ACTLCT, CASHAT, FAT, NIMTA, PRICE, RSIZE, WCAPAT, and LTAT. They cover key aspects of current asset, income, cash flow, liability, and market capitalization of a firm. Figure 4(b), meanwhile, shows a simpler structure with a clearer relation for the decision making process determined by fewer variables including ACTLCT, FAT, PRICE, RSIZE, WCAPAT and LTAT. It thus provides a simpler logic reasoning of the internal functionality. However, as results in Table 4 Panel B suggest, the simpler structure compromises on the forecasting accuracy. Hence, a natural trade-off exists between the simplicity and interpretability of the model and its forecasting performance.

Furthermore, the structure in Figure 4(a) can be interpreted as the conditional probability of *Pr*(dlsrn *j* ACTLCT, CASHAT, FAT, NIMTA, PRICE, RSIZE, WCAPAT, LTAT), which not only provides a binary answer of *true* or *false* to the future bankruptcy problem but also yields a probability of it. Each variables is affected by others through a conditional probability, such as *Pr*(FAT *j* CASHAT, LTAT, REAT, WCAPAT, SALEAT), where LTAT, the ratio between total liabilities and total assets, indirectly influences the probability of bankruptcy via its impact on FAT.

The conditional probability of *Pr*(dlsrn*jScenario*) is well suited to address *what-if* questions via a scenario analysis. In Table 6, the Bayesian network model generates a default probability of 0.6867 based on real data. This suggests that the firm under scrutiny is highly likely to default in the future. Hence, we perform a scenario analysis to see what happens when key variable values change. We consider two different scenarios: In the first scenario, we assume that a firm's financial health deteriorates with decreasing net income, cash flow and current asset, and increasing debt and liability; whereas in the second scenario, the financial status improves with higher income, cash flow and current asset and lower debt and liability. Specially, in the first scenario, ACTLCT, CASHAT, and NIMTA are reduced to be less than 1, 0.01, and 0.0001, respectively, while FAT and LTAT are increased to be over 0.8 and 1, respectively. As we expect, the bankruptcy probability shoots up to 0.9789, showing an extremely risky situation that, with less asset, less income but more debt, the firm is almost surely going to default. In the second scenario, a financially healthy firm with more asset, more income and less debt exhibits an extremely low bankruptcy probability of 0.0465. The scenario analysis not only points to the direction of firm survival but also quantifies the default probability given specific variable values, making this analytical tool very helpful for stakeholders in- and outside the firm.

We are also able to perform sensitivity analysis of bankruptcy prediction with respect to particular variables of interest. Based on the information in Table 6, we select three pairs of influential variables and generate probability surfaces to capture the impact of these variables on bankruptcy probability in Figure 5. Figure 5(a) illustrates the bankruptcy probability conditional upon the ratio between total liabilities and total assets (LTAT) and the ratio between current assets to current liabilities (ACTLCT). With zero current asset (ACTLCT=0) and a high total liability ratio (LTAT=7 10^4), the bankruptcy probability increases to as high as 0.8. If the total liability ratio remains at LTAT=7 10^4 but current asset ratio (ACTLCT) increases from 0 to the third quartile at 2.32, the bankruptcy probability decreases slowly with tiny magnitude. This shows that a massive liability is hugely detrimental to firm solvency even with large current assets.

Figure 5(b) exhibits bankruptcy probability changes conditional on the ratio of debts to total assets (FAT) and the ratio of cash to total assets (CASHAT). The pattern is similar to that in Figure 5(a) and indicates that when FAT is as high as 2 10^4, the bankruptcy probability remains high at 0.7 even when CASHAT reaches the third quartile at 0.69. The probability drops only to 0.5 even when CASHAT reaches an incredibly high level of 10000. Meanwhile, if the FAT is drops to the 3^{rd} quartile at 0.268, the probability decreases almost linearly as CASHAT increases. This quantifies and highlights the importance of the debt ratio for the financial health of a firm and reveals that maintaining an appropriate level of debt is an effective way of avoiding future bankruptcy.

In Figure 5(c), we note a similar pattern that when total liabilities over total assets (LTAT) is at a high level of 7 10^4, the bankruptcy probability reaches 0.8 and decreases ever so slowly even when the ratio of net income to the total liability (NIMTA) grows to 15, an extremely high level for net income. However, if LTAT stays at the third quartile at 0.75, the probability decreases to 0.5 and further decreases to lower value than 0.4 as NIMTA increases.

To summarize, Figure 5 clearly reflects *post x* interpretation based on the model. It addresses the link between conditional probabilities and the final outcome suggested by the model, and captures inferences from the structural conditional probability. The scenario and sensitivity analyses described above are of great use to investors and policymakers as they offer detailed explanation in terms of the theoretical and empirical functionality of the model.

4.6 Robustness check

The empirical analyses so far are based on a long sample period from January 1961 to August 2018, which experiences different business cycles and a number of financial crises. As a robustness check, we evaluate the LASSO selection and Bayesian network model again using a shorter and more recent sample period starting from just before the Great recession in March 2007 to the end of the sample period in August 2018. This is motivated by Figure 2 which shows a recent wave of firm defaults and it would be interesting to see which variables the LASSO identified and how well they predict bankruptcy. In total we have 423,012 firm-quarter observations for the robustness test.

Table 3 Panel B shows the selected variables based on the shorter sample period. We notice a large overlap between the selected variables from the whole sample and the subsample. For example, the log market capitalization (RSIZE) and the log stock price (PRICE) remain the second and third most important variables. Among all 16 variables with non-zero coefficient, nine of them overlap with those in Panel A. The new list of selected variables cover quick assets (QALCT), cash (CASHMTA), inventory (INVTSALE), liability (LCTSALE) and sales (TSALE) that are very similar to variables of current asset (ACTLCT), cash (CASHAT), inventory (INVCHINVT), liabilities (LCTAT), and sales (SALEAT) in Panel A. The new selection exhibits more focus on the liability (LTMTA), accounts payable (APSALE), and total asset (TASSET), all of which are essential for firm survival in turbulent market conditions.

We use all selected 16 variables (Group 1) and the top 11 variables (Group 2) to predict firm bankruptcy and summarize the results in Table 7. The bankruptcy forecasting results are generally in line with those in Table 4: First, the forecasting performance generally improves with model sophistication; second, the Bayesian network is only outperformed by the most sophisticated deep neural network model; third, results with fewer variables are less accurate.

5 Conclusion

Despite the advancement of sophisticated algorithm that well describe data, the financial industry is faced with a new call for formulating interpretable machine learning models that are not only powerful in performance but also comprehensible to investors to allow them to make informed investment decisions. In this paper, motivated by successful applications of the Bayesian network in the healthcare diagnosis area, we modify the Bayesian network and implement it for the purpose of predicting firm bankruptcy probability. We first select relevant variables by the LASSO, construct the Bayesian network with selected variables, and estimate model parameters via the EM algorithm. We use quarterly COMPUSTAT data from January 1961 to August 2018 for the empirical analyses and show that the Bayesian network model performs very well and is outperformed only by the complicated deep neural network with three hidden layers. Furthermore, the topology of the Bayesian network exhibits a clear representation of its internal functionality based on conditional probability inferences. It offers scenario and sensitivity analyses of individual variables on bankruptcy probability making it easily understandable by the general investment community. This underlines the contribution of our paper to the literature.

While we consider our study an important step towards integrating superior forecasting performance with model interpretability, we recognize that the model can be improved along different dimensions in the future. For example, the topology of the Bayesian network can be made dynamically updated over time for more flexibility. Furthermore, most machine learning models forecast bankruptcy probability at time $t+1$ based on data at time t, whereas traditional Hazard models include data at all *healthy* time to predict bankruptcy probability. Thus a machine learning model can be enhanced by a Hazard model in bringing more promising results with an interpretable structure.

Acknowledgement

The data that support the findings of this study are available from the COMPUSTAT through the Wharton Research Data Services (WRDS). Restrictions apply to the availability of these data, which were used under licence for this study. Sample data are available from the authors with the permission of COMPUSTAT database and WRDS.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance 23*, 589–609.

Amendola, A., M. Restaino, and L. Sensini (2011). Variable selection in default risk models. *Journal of Risk Model Validation 5*, 3–19.

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*, 304–317.

Bastani, O., C. Kim, and H. Bastani (2017). Interpretability via model extraction. *arXiv: 1706.09773*.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pp. 437–478. Springer.

Bharath, S. T. and T. Shumway (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies 21*, 1339–1369.

Campbell, J. Y., J. Hilscher, and J. Szilagyi (2008). In search of distress risk. *Journal of Finance 63*, 2899–2939.

Cerchiello, P., G. Nicola, S. Ronnqvist, and P. Sarlin (2017). Deep learning bank distress from news and numerical financial data. *arXiv: 1706.09627*.

Chakraborty, S., K. Mengersen, C. Fidge, L. Ma, and D. Lassen (2016). A Bayesian network-based customer satisfaction model: A tool for management decisions in railway transport. *Decision Analytics 3*, 4–28.

Chandra, D. K., V. Ravi, and I. Bose (2009). Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications 36*, 4830–4837.

Chava, S. and R. A. Jarrow (2004). Bankruptcy prediction with industry effects. *Review of Finance 8*, 537–569.

Chen, M. A., Q. Wu, and B. Yang (2019). How valuable is fintech innovation? *The Review of Financial Studies 32*(5), 2062–2106.

Chinco, A., A. D. Clark-Joseph, and M. Ye (2019). Sparse signals in the cross-section of returns. *Journal of Finance 74*, 449–492.

Cooper, G. F. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning 9*, 309–347.

Dag, A., K. Topuz, A. Oztekin, S. Bulur, and F. M. Megahed (2016). A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival. *Decision Support Systems 86*, 1–12.

Daly, R., Q. Shen, and S. Aitken (2011). Learning Bayesian networks: Approaches and issues. *Knowledge Engineering Review 26*, 99–157.

Delen, D., K. Topuz, and E. Eryarsoy (2019). Development of a Bayesian belief network-based DSS for predicting and understanding freshmen student attrition. *forthcoming* European Journal of Operational Research.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*, 1–38.

Ding, A. A., S. Tian, Y. Yu, and H. Guo (2012). A class of discrete transformation survival models with application to default probability prediction. *Journal of the American Statistical Association 107*, 990–1003.

Doshi-Velez, F. and B. Kim (2017). Towards a rigorous science of interpretable machine learning. *arXiv: 1702.08608*.

du Jardin, P. (2015). Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research 242*, 286–303.

du Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research 254*, 236–252.

du Jardin, P. (2018). Failure pattern-based ensembles applied to bankruptcy forecasting. *Decision Support Systems 107*, 64–77.

Engelmann, B., E. Hayden, and D. Tasche (2003). Measuring the discriminative power of rating systems. Discussion paper, Series 2: Banking and Financial Supervision, Deutsche Bundesbank.

Feng, N., H. J. Wang, and M. Li (2014). A security risk analysis model for information systems: Causal relationships of risk factors and vulnerability propagation analysis. *Information Sciences 256*, 57–73.

Garvey, M. D., S. Carnovale, and S. Yeniyurt (2015). An analytical framework for supply network risk propagation: A Bayesian network approach. *European Journal of Operational Research 243*, 618–627.

Geng, R., I. Bose, and X. Chen (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research 241*, 236–247.

Gogas, P., T. Papadimitriou, and A. Agrapetidou (2018). Forecasting bank failures and stress testing: A machine learning approach. *International Journal of Forecasting 34*, 440–455.

Goldstein, I., W. Jiang, and G. A. Karolyi (2019). To fintech and beyond. *The Review of Financial Studies 32*(5), 1647–1661.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press.

Green, P. J. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological) 52*, 443–452.

Heckerman, D., D. Geiger, and D. M. Chickering (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning 20*, 197–243.

Hu, Y., X. Zhang, E. Ngai, R. Cai, and M. Liu (2013). Software project risk analysis using Bayesian networks with causality constraints. *Decision Support Systems 56*, 439–449.

Huang, J. Z. (2014). An introduction to statistical learning: With applications in r by gareth james, trevor hastie, robert tibshirani, daniela witten.

Kim, B. and F. Doshi-Velez (2017). Interpretable machine learning: The fuss, the concrete and the questions. In *34th International Conference on Machine Learning (ICML 2017)*, Volume Tutorials Session B.

Kim, H. H. and N. R. Swanson (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics 178*, 352–367.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis 190*, 191–203.

Liang, D., C.-C. Lu, C.-F. Tsai, and G.-A. Shih (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research 252*, 561–572.

Lin, W.-Y., Y.-H. Hu, and C.-F. Tsai (2012). Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42*, 421–436.

Liu, S., J. Zeng, H. Gong, H. Yang, J. Zhai, Y. Cao, J. Liu, Y. Luo, Y. Li, and L. Maguire (2018). Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach. *Computers in Biology and Medicine 92*, 168–175.

Lu, Y., Z. Cheng, Y. Zhao, X. Chang, C. Chan, Y. Bai, and N. Cheng (2016). Efficacy and safety of long-term treatment with statins for coronary heart disease: A Bayesian network meta-analysis. *Atherosclerosis 254*, 215–227.

Mai, F., S. Tian, C. Lee, and L. Ma (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research 274*, 743–758.

Malagrino, L. S., N. T. Roman, and A. M. Monteiro (2018). Forecasting stock market index daily direction: A Bayesian network approach. *Expert Systems with Applications 105*, 11–22.

Matin, R., C. Hansen, C. Hansen, and P. Mølgaard (2018). Predicting distresses using deep learning of text segments in annual reports. *arXiv: 1811.05270*.

Mittelstadt, B., C. Russell, and S. Wachter (2018). Explaining explanations in AI. *arXiv: 1811.01439*.

Nazemi, A. and F. J. Fabozzi (2018). Macroeconomic variable selection for creditor recovery rates. *Journal of Banking & Finance 89*, 14–25.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research 18*, 109–131.

Rapach, D. E., J. K. Strauss, and G. Zhou (2013). International stock return predictability: What is the role of the United States? *Journal of Finance 68*, 1633–1662.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM.

Seixas, F. L., B. Zadrozny, J. Laks, A. Conci, and D. C. M. Saade (2014). A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. *Computers in Biology and Medicine 51*, 140–158.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business 74*, 101–124.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research 15*, 1929–1958.

Sun, L. and P. P. Shenoy (2007). Using Bayesian networks for bankruptcy prediction: Some methodological issues. *European Journal of Operational Research 180*, 738–753.

Tian, S., Y. Yu, and H. Guo (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance 52*, 89–100.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 267–288.

Yet, B., A. Constantinou, N. Fenton, M. Neil, E. Luedeling, and K. Shepherd (2016). A Bayesian network framework for project cost, benefit and risk analysis with an agricultural development case study. *Expert Systems with Applications 60*, 141–155.

Zhang, Q., C. Zhou, Y.-C. Tian, N. Xiong, Y. Qin, and B. Hu (2018). A fuzzy probability Bayesian network approach for dynamic cybersecurity risk assessment in industrial control systems. *IEEE Transactions on Industrial Informatics 14*, 2497–2506.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research 22*, 59–82.

Table 1: A partial summary of relevant literature in bankruptcy prediction

This table summarizes recent studies on bankruptcy prediction based on machine learning models. The models are categorized into interpretable or non-interpretable ones.

Study	Data and sample	Non-interpretable Model	Interpretable Model
du Jardin (2018)	bureau Van Dijk, Diane, 2006-2014	ensemble-based model and self-organizing map (SOM)	-
du Jardin (2015)	Bureau van Dijk, 2003-2012	Multilayer perceptron (MLP) or Neural network Survival analysis (SA), Self-organizing map (SOM)	Discriminant analysis (DA), Logit Decision tree (DT)
Geng et al. (2015)	China Security Market Accounting Research (CSMAR), 2001-2008	Support vector machine (SVM), Neural network (NN)	Decision tree (DT)
du Jardin (2016)	Bureau van Dijk, 2003-2012	Multilayer perceptron (MLP) or Neural network, Bagging, boosting, profile-based models (PBMs)	Discriminant analysis (DA), Logit Decision tree (DT)
Liang et al. (2016)	Taiwan Economic Journal, 1999-2009	Support vector machine (SVM), k nearest neighbour (kNN), Naïve Bayes (NB), Multilayer perceptron (MLP) or Neural network	Classification and regression (CART)
Cerchiello et al. (2017)	Reuter online archive, 2007-2016	Deep neural network (DNN)	-
Gogas et al. (2018)	Collected US banks 2007-2013	Support vector machine (SVM)	-
Mai et al. (2019)	CRSP, 1994-2014	Convolutional recurrent neural network (CNN), Support vector machine (SVM), Random forest (RF)	Logit
Matin et al. (2018)	Danish Central Business Register, 2013-2016	Convolutional recurrent neural network (CNN), Long Short-Term Memory (LSTM)	Logit

Table 2: Summary statistics of input accounting variables

This table summarizes the minimum (Min), first quantile (1st Qu), median, mean, third quantile (3rd Qu), maximum (Max) and standard deviation (Std) of input accounting variables that are used for predicting corporate bankruptcy probability. Variable description is provided in the last column. The sample period is from January 1961 to August 2018.

	Min	1st Qu	Median	Mean	3rd Qu	Max	Std	Description
ACTLCT	0.0465	1.7472	2.2236	2.1325	2.2236	30.92	0.8031	Current Assets/ Current Liabilities
APSALE	0.0034	0.2308	0.3404	2.3789	0.4431	70.13	8.8062	Accounts Payable / Sales
CASHAT	0.0006	0.0309	0.0686	0.0826	0.0686	0.9271	0.0916	Cash and Short-term Investment / Total Assets
CASHMTA	0.0001	0.0143	0.0332	0.0445	0.0332	0.5713	0.0602	Cash and Short-term Investment / (Market Equity + Total Liabilities)
CHAT	0.0007	0.0615	0.1537	0.2611	0.1586	4.1804	0.4654	Cash / Total Assets
CHLCT	0.0079	0.4497	1.1090	1.3314	1.1090	66.74	1.9674	Cash / Current liabilities
EBITSALE	-0.0069	0.0005	0.0014	0.0026	0.0021	0.3248	0.0071	Earnings before Interest and Tax / Sales
FAT	0.0031	0.1300	0.1457	0.1767	0.2003	3.9578	0.1544	Total Debts / Total Assets
INVCHINVT	-1.3420	-0.0200	-0.0158	-0.0385	-0.0158	0.7186	0.1266	Growth of Inventories / Inventories
INVTSALE	0.0115	0.3289	0.5250	0.7353	0.7617	29.90	0.9205	Inventories / Sales
LCTAT	0.0024	0.1386	0.1386	0.1939	0.2403	0.8574	0.1204	Current Liabilities / Total Assets
LCTLT	0.0058	0.3057	0.3057	0.3929	0.5042	1.0000	0.2270	Current Liabilities / Total Liabilities
LCTSALE	0.0578	0.6394	0.7890	1.1478	1.0979	16.41	1.2716	Current Liabilities / Sales
LTAT	0.0927	0.4534	0.4534	0.5332	0.6300	1.2915	0.1735	Total Liabilities / Total Assets
LTMTA	0.0075	0.2195	0.2195	0.3261	0.4470	0.9674	0.2464	Total Liabilities /(Market Equity + Total Liabilities)
TASSET	5.7470	156.2	156.3	580.1	419.7	13061	1199	Total Asset
TSALE	1.2140	19.45	27.46	110.6	83.28	1916	226.5	Total SALE
NIAT	-0.1894	0.0029	0.0039	0.0073	0.0138	0.1018	0.0172	Net Income / Total Assets
NIMTA	-0.0542	0.0012	0.0019	0.0046	0.0069	0.0516	0.0100	Net Income / (Market Equity + Total Liabilities)
NISALE	-0.9243	0.0205	0.0224	0.0361	0.0673	0.3965	0.0867	Net Income / Sales
OIADPAT	-0.0750	0.0105	0.0105	0.0174	0.0252	0.1258	0.0211	Operating Income / Total Assets
OIADPSALE	-0.3453	0.0522	0.0596	0.0978	0.1327	0.6382	0.1158	Operating Income / Sales
PRICE	-1.0940	2.2460	2.4840	2.5340	3.0620	4.5380	0.8492	log(price)
QALCT	-2.6810	1.0460	1.5580	1.4060	1.5580	5.8190	0.6859	Quick Assets / Current Liabilities
REAT	-2.0272	0.0343	0.0343	0.0890	0.1841	0.7528	0.2421	Retained Earnings / Total Assets
RELCT	-6.8921	0.2473	0.2473	0.6179	1.0180	7.5856	1.3707	Retained Earnings / Current Liabilities
RSIZE	3.3080	5.5080	5.5080	5.5590	5.5080	8.3310	0.4272	Log(Market Capitalization)
SALEAT	0.0135	0.1467	0.1757	0.2496	0.3314	1.2160	0.1899	Sales / Total Assets
SEQAT	-0.0395	0.3280	0.3280	0.3778	0.4715	0.8442	0.1715	Equity / Total Assets
SIGMA	0.0000	0.0047	0.0047	0.0047	0.0047	0.0108	0.0087	Volatility
WCAPAT	-0.1373	0.0841	0.0841	0.1552	0.2357	0.6480	0.1555	Working Capital/Total Assets

Table 3: Explanatory variables selected by the LASSO

This table summarizes the explanatory variables as selected by the LASSO. In Panel A the full sample period is from January 1961 to August 2018; in Panel B the subsample is from March 2007 to August 2018. ***, ** and * denote significance at the 1%, 5%, and 10% level respectively. The z-statistics are reported in parenthese.

	Panel A. Full sample		*Panel B. Subsample*	
#	Variable	Coefficient	Variable	Coefficient
1	**LTAT**	**0.2830 (10.929)*** **	**LTMTA**	**0.2999 (14.077)*** **
2	**RSIZE**	**-0.0442 (51.000)*** **	**RSIZE**	**-0.1131 (11.261)*** **
3	**PRICE**	**-0.0109 (32.696)*** **	**PRICE**	**-0.0813 (12.441)*** **
4	**INVCHINVT**	**6.6716E-05 (9.549)*** **	**NIMTA**	**-0.0311 (11.706)*** **
5	**REAT**	**-2.3784E-05 (5.154)*** **	**CASHMTA**	**-0.0135 (11.682)*** **
6	**WCAPAT**	**-1.9306E-05 (7.015)*** **	**LCTLT**	**2.3160E-04 (12.167)*** **
7	**LCTLT**	**1.7622E-05 (5.492)*** **	**OIADPSALE**	**-1.9897E-04 (10.987)*** **
8	**FAT**	**1.2234E-05 (6.388)*** **	**LCTSALE**	**4.2607E-05 (11.137)*** **
9	**ACTLCT**	**-1.2230E-05 (11.022)*** **	**QALCT**	**-2.5710E-05 (7.100)*** **
10	**RELCT**	**-1.1713E-05 (11.133)*** **	**LTAT**	**1.8029E-05 (10.729)*** **
11	NIAT	-8.9378E-06 (2.007)**	**RELCT**	**-1.1955E-05 (7.418)*** **
12	LCTAT	8.2558E-06 (2.166)**	TSALE	-9.3633E-06 (2.753)***
13	SALEAT	-6.0518E-06 (2.123)**	FAT	8.4814E-06 (1.716)*
14	CASHAT	-4.9206E-06 (1.326)*	INVTSALE	-7.2972E-06 (1.022)*
15	NIMTA	-3.0785E-06 (1.299)*	APSALE	-1.5559E-06 (1.533)*
16	OIADPSALE	-5.1861E-07 (1.965)**	TASSET	-1.0079E-07 (2.492)**
17	LCTSALE	0.0000	ACTLCT	0.0000
18	EBITSALE	0.0000	CASHAT	0.0000
19	QALCT	0.0000	CHAT	0.0000
20	APSALE	0.0000	CHLCT	0.0000
21	CASHMTA	0.0000	EBITSALE	0.0000
22	CHAT	0.0000	INVCHINVT	0.0000
23	CHLCT	0.0000	LCTAT	0.0000
24	INVTSALE	0.0000	NIAT	0.0000
25	LTMTA	0.0000	NISALE	0.0000
26	OIADPAT	0.0000	OIADPAT	0.0000
27	SEQAT	0.0000	REAT	0.0000
28	SIGMA	0.0000	SALEAT	0.0000
29	TASSET	0.0000	SEQAT	0.0000
30	TSALE	0.0000	SIGMA	0.0000
31	NISALE	0.0000	WCAPAT	0.0000

Table 4: Bankruptcy forecasting performance across different models

In this table, the bankruptcy prediction accuracy measures ACCU and AUC are obtained via the logistic regression (LR), decision tree (Tree), support vector machine (SVM), deep neural network (DNN(50,30,20)), and Bayesian network (BN) over forecasting horizons of one to 12 quarters ahead. Panels A and B use 16 (Group 1) and 10 (Group 2) most influential variables, respectively. The sample period is from January 1961 to August 2018.

	LR		Tree		SVM		DNN (50,30,20)		BN	
	ACCU	AUC	ACCU	AUC	ACCU	AUC	ACCU	AUC	ACCU	AUC
Panel A: Group 1 variables										
1Q	0.7661	0.7866	0.7944	0.8186	0.8344	0.8804	0.8569	0.9003	0.8506	0.8951
2Q	0.7654	0.7750	0.7871	0.7892	0.8119	0.8615	0.8458	0.8890	0.8331	0.8821
3Q	0.7619	0.7785	0.7787	0.7938	0.8082	0.8534	0.8376	0.8728	0.8240	0.8669
4Q	0.7630	0.7703	0.7686	0.7946	0.8325	0.8654	0.8190	0.8686	0.8123	0.8632
5Q	0.7648	0.7827	0.7547	0.7846	0.8042	0.8392	0.8166	0.8674	0.8140	0.8597
6Q	0.7620	0.7850	0.7465	0.7642	0.7922	0.8238	0.7934	0.8532	0.7891	0.8487
7Q	0.7587	0.7635	0.7404	0.7428	0.7880	0.8139	0.7959	0.8456	0.7907	0.8415
8Q	0.7538	0.7633	0.7338	0.7444	0.7691	0.7908	0.7807	0.8360	0.7728	0.8326
9Q	0.7546	0.7694	0.7351	0.7603	0.8319	0.8413	0.7783	0.8230	0.7758	0.8222
10Q	0.7608	0.7644	0.7406	0.7434	0.7523	0.7602	0.7564	0.8191	0.7575	0.8168
11Q	0.7080	0.7144	0.6685	0.6915	0.7446	0.7446	0.7812	0.8088	0.7669	0.8055
12Q	0.6878	0.6899	0.6692	0.6698	0.7028	0.6929	0.7494	0.8115	0.7525	0.8097
Avg	0.7506	0.7619	0.7431	0.7581	0.7893	0.8139	0.8009	0.8496	0.7949	0.8454
Panel B: Group 2 variables										
1Q	0.6863	0.7029	0.6803	0.6828	0.7420	0.7820	0.7689	0.8264	0.7631	0.8205
2Q	0.6853	0.6952	0.6795	0.7053	0.7232	0.7632	0.7605	0.8132	0.7571	0.8099
3Q	0.6626	0.6730	0.6570	0.6865	0.7066	0.7466	0.7518	0.7977	0.7429	0.7912
4Q	0.6618	0.6838	0.6568	0.6770	0.6899	0.7299	0.7461	0.7868	0.7341	0.7858
5Q	0.6637	0.6862	0.6601	0.6679	0.6766	0.7166	0.7151	0.7784	0.7032	0.7781
6Q	0.6604	0.6610	0.6580	0.6766	0.6616	0.7016	0.6836	0.7613	0.6838	0.7611
7Q	0.5577	0.5807	0.5563	0.5653	0.6489	0.6889	0.6793	0.7607	0.6658	0.7629
8Q	0.5529	0.5772	0.5488	0.5618	0.6274	0.6674	0.6703	0.7552	0.6845	0.7552
9Q	0.5534	0.5782	0.5501	0.5504	0.6125	0.6525	0.6552	0.7589	0.6612	0.7559
10Q	0.5591	0.5823	0.5556	0.5639	0.5999	0.6399	0.6526	0.7414	0.6506	0.7416
11Q	0.5063	0.5313	0.5035	0.5125	0.5854	0.6254	0.6124	0.7268	0.6292	0.7289
12Q	0.5065	0.5363	0.5042	0.5143	0.5683	0.6083	0.6221	0.7257	0.6238	0.7259
Avg	0.6047	0.6240	0.6009	0.6137	0.6535	0.6935	0.6932	0.7694	0.6916	0.7681

Table 5: Logistic regression coefficients over the one-year bankruptcy forecasting horizon

This table summarizes estimated coefficients and their statistical significance obtained from the logistic regression in predicting one-year ahead firm bankruptcy. The sample period is from January 1961 to August 2018.

#		Estimate	Std. error	z value	$\Pr(> \|z\|)$
0	Intercept	-0.2230	0.017	-12.82	$<$ 2e-16
1	ACTLCT	5.84E-05	1.00E-05	5.8400	5.21E-09
2	CASHAT	-2.03E-04	7.86E-05	2.5760	0.0100
3	FAT	3.13E-04	4.87E-05	-6.4350	1.23E-10
4	INVCHINVT	0.0036	6.53E-04	5.4480	5.10E-08
5	NIAT	-5.33E-04	9.22E-05	-5.7780	7.54E-09
6	NIMTA	-0.2830	0.0380	-7.4440	9.78E-14
7	PRICE	-0.0258	0.0023	-11.39	$<$ 2e-16
8	RELCT	-5.64E-05	1.64E-05	3.4360	0.0006
9	RSIZE	-0.1710	0.0035	-48.74	$<$ 2e-16
10	SALEAT	-1.16E-04	3.20E-05	-3.6240	0.0003
11	WCAPAT	-4.91E-04	8.16E-05	6.0130	1.82E-09
12	REAT	-5.07E-05	1.21E-05	-4.1790	2.93E-05
13	LTAT	3.40E-04	1.12E-04	-3.0380	0.0024
14	LCTAT	1.55E-04	6.13E-05	2.5340	0.0113
15	LCTLT	1.56E-04	2.81E-05	5.5440	2.96E-08
16	OIADPSALE	-2.83E-04	8.10E-05	3.4950	0.0005
	AIC: 579901				

Table 6: Scenario analysis of bankruptcy probability

Based on information in Figure 4, this table reports bankruptcy probability based on variable values and two *ad-hoc* scenarios when key variable values are changed leading to different bankruptcy probability.

Variables	Data	Scenario 1	Scenario 2
ACTLCT	1.7580	< 1:0	> 2:0
CASHAT	0.0398	< 0:03	> 0:5
FAT	0.1379	> 0:2	< 0:1
NIMTA	0.0010	< 0:0001	> 0:1
LTAT	0.5893	> 0:6	< 0:1
PRICE	1.3545	1.3545	1.3545
RSIZE	5.5077	5.5077	5.5077
WCAPAT	0.2028	0.2028	0.2028
INVCHINVT	-0.0038	-0.0038	-0.0038
NIAT	0.0016	0.0016	0.0016
RELCT	0.7156	0.7156	0.7156
SALEAT	0.3343	0.3343	0.3343
REAT	0.1915	0.1915	0.1915
LCTAT	0.2676	0.2676	0.2676
Probability	0.6867	0.9789	0.0465

Table 7: Robustness test: Bankruptcy forecasting performance across different models over more recent sample period

In this table, the bankruptcy prediction accuracy measures ACCU and AUC are obtained via the logistic regression (LR), decision tree (Tree), support vector machine (SVM), deep neural network (DNN(50,30,20)), and Bayesian network (BN) over forecasting horizons of one to 12 quarters ahead. Panels A and B use 16 (Group 1) and 10 (Group 2) most influential variables, respectively. The sample period is from March 2007 to August 2018.

	LR		Tree		SVM		DNN (50,30,20)		BN	
	ACCU	AUC	ACCU	AUC	ACCU	AUC	ACCU	AUC	ACCU	AUC
Panel A: Robust Group 1 variables										
1Q	0.7470	0.7730	0.7892	0.8155	0.7886	0.8332	0.8103	0.8554	0.7975	0.8890
2Q	0.7188	0.7272	0.7753	0.7736	0.8071	0.8472	0.8256	0.8695	0.7872	0.8167
3Q	0.7435	0.7580	0.7426	0.7514	0.7865	0.8193	0.8372	0.8333	0.8008	0.8490
4Q	0.7602	0.7586	0.7682	0.7522	0.8272	0.8538	0.7802	0.8299	0.7625	0.8566
5Q	0.7521	0.7657	0.7351	0.7661	0.7848	0.8016	0.8005	0.8309	0.7615	0.8273
6Q	0.7584	0.7803	0.7169	0.7592	0.7673	0.8222	0.7858	0.8204	0.7760	0.8386
7Q	0.7476	0.7423	0.7177	0.7287	0.7467	0.7810	0.7633	0.8317	0.7395	0.8292
8Q	0.7047	0.7398	0.7241	0.7129	0.7451	0.7683	0.7773	0.7923	0.7152	0.7866
9Q	0.7153	0.7329	0.7152	0.7107	0.7826	0.8059	0.7496	0.7960	0.7498	0.7567
10Q	0.7157	0.7311	0.7078	0.6936	0.7205	0.7193	0.7222	0.8172	0.7518	0.8002
11Q	0.6925	0.6708	0.6366	0.6446	0.7266	0.7094	0.7595	0.7922	0.7643	0.7870
12Q	0.6738	0.6590	0.6308	0.6351	0.6862	0.6584	0.7204	0.7927	0.7510	0.7573
Avg	0.7275	0.7366	0.7216	0.7286	0.7641	0.7850	0.7776	0.8218	0.7631	0.8162
Panel B: Robust Group 2 variables										
1Q	0.6510	0.6599	0.6472	0.6620	0.7291	0.7706	0.7293	0.7894	0.7608	0.8049
2Q	0.6491	0.6898	0.6730	0.6811	0.6837	0.7571	0.7131	0.8127	0.7200	0.7572
3Q	0.6471	0.6415	0.6148	0.6745	0.6704	0.7074	0.7253	0.7891	0.7378	0.7727
4Q	0.6263	0.6545	0.6471	0.6582	0.6424	0.7012	0.6980	0.7680	0.6926	0.7842
5Q	0.6242	0.6392	0.6481	0.6393	0.6505	0.7113	0.6919	0.7438	0.6684	0.7256
6Q	0.6225	0.6404	0.6273	0.6480	0.6360	0.6567	0.6507	0.7269	0.6527	0.7265
7Q	0.5264	0.5444	0.5080	0.5321	0.6406	0.6512	0.6676	0.7277	0.6139	0.7457
8Q	0.5300	0.5734	0.5028	0.5457	0.6176	0.6613	0.6682	0.7409	0.6399	0.7223
9Q	0.5217	0.5719	0.5286	0.5055	0.5968	0.6046	0.6307	0.7356	0.6088	0.7327
10Q	0.5205	0.5789	0.5511	0.5256	0.5717	0.6175	0.6039	0.7385	0.6460	0.7063
11Q	0.4793	0.5279	0.4879	0.4733	0.5389	0.5957	0.5944	0.7256	0.6268	0.7107
12Q	0.4813	0.4965	0.4903	0.4971	0.5466	0.5748	0.5728	0.6946	0.6052	0.6795
Avg	0.5733	0.6015	0.5772	0.5869	0.6270	0.6674	0.6622	0.7494	0.6644	0.7390

Figure 1: A simple illustration of the Bayesian network

This figure illustrates a simple Bayesian network example in which bank failure depends on two variables: the Risk-Adjusted Capital Ratio (R) and the Total Capital (C), and the former variable also directly depends on the latter.

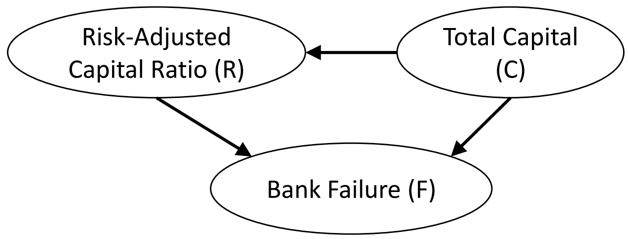


Figure 2: Bankruptcy and liquidation between 1991-2018

This figure shows the number of corporate bankruptcy filing each year from 1961 to 2018. The bankruptcy is shown by the *dlrsn* in the COMPUSTAT, which shows the reason a firm becomes inactive. We consider a firm is in default when *dlrsn* shows *bankruptcy* or *liquidation* in our study.

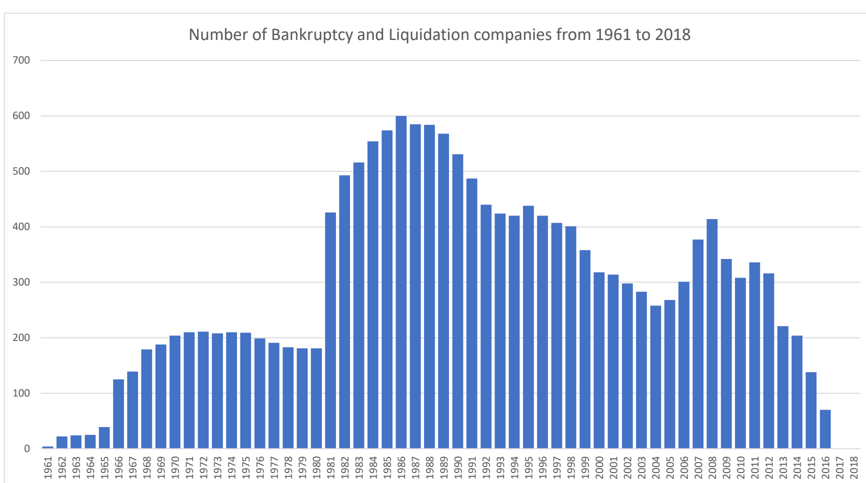


Figure 3: The decision tree model

This figure shows the decision tree model for forecasting bankruptcy four quarters ahead constructed from all 16 variables. The sample period is from January 1961 to August 2018.

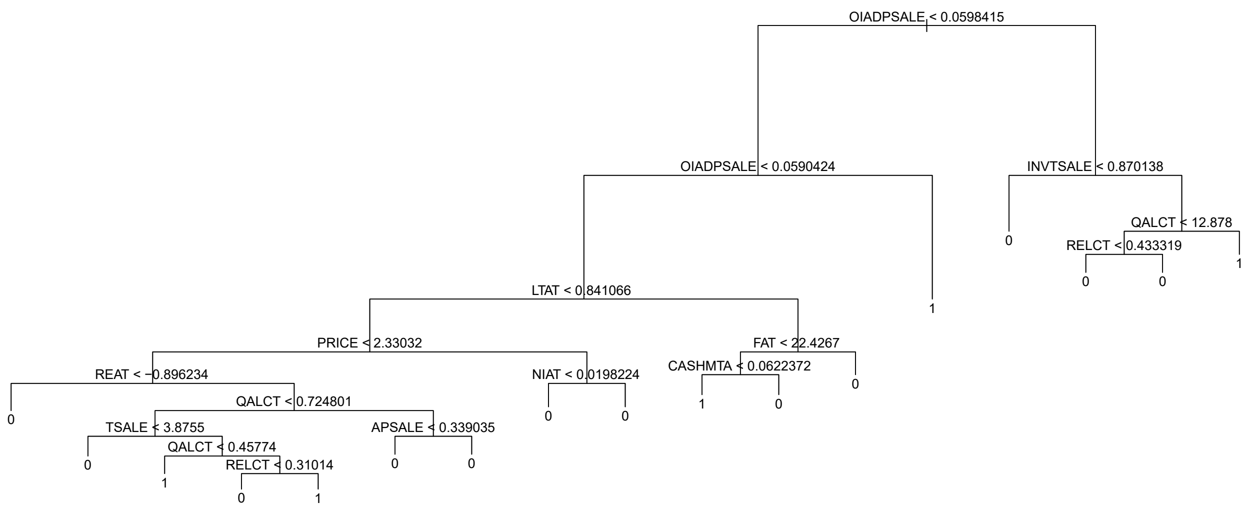
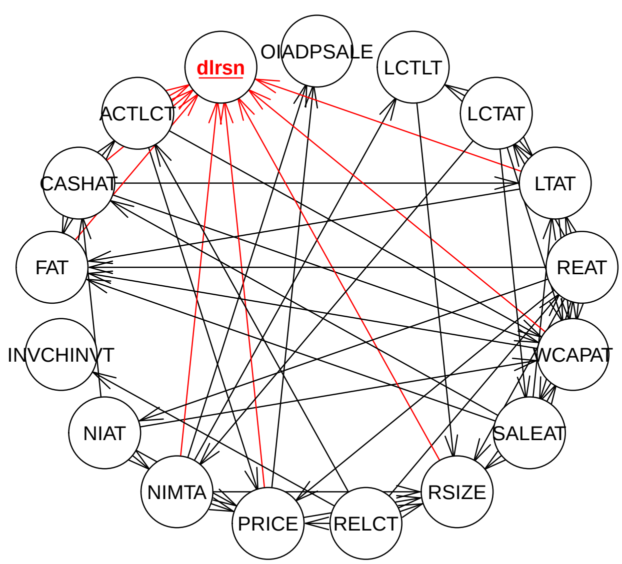
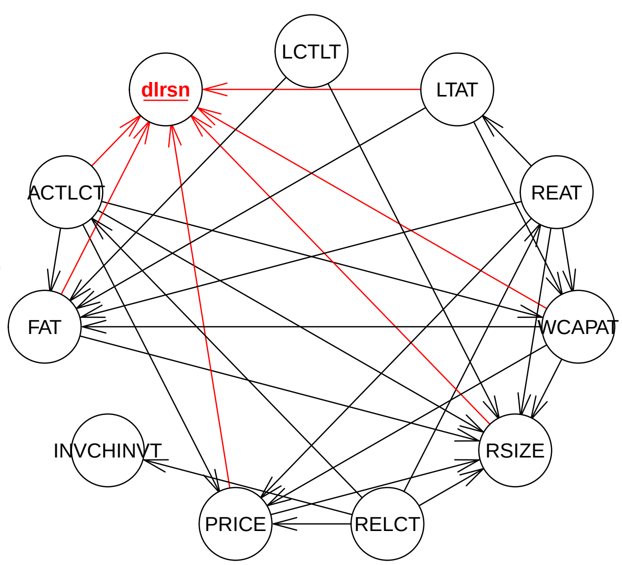


Figure 4: The Bayesian network model

This figure shows the structure of the Bayesian network for forecasting bankruptcy four quarters ahead constructed by (a) 16 variables in Group 1 and (b) 10 variables in Group 2. The sample period is from January 1961 to August 2018.



(a)



(b)