Bankruptcy Prediction of Privately Held SMEs Using Feature Selection Methods[⋆]

Florentina Paraschiv[a,b], Markus Schmid[c], Ranik Raaen Wahlstrøm[b]

[a]*Zeppelin University, Department of Finance, Am Seemooser Horn 20, Friedrichshafen 88045, Germany*
[b]*NTNU Business School, Norwegian University of Science and Technology, 7491 Trondheim, Norway*
[c]*Swiss Finance Institute (SFI) and Swiss Institute of Banking and Finance, University of St. Gallen, Unterer Graben 21, 9000 St. Gallen, Switzerland*

Abstract

In this paper, we test alternative feature selection methods for bankruptcy prediction and illustrate their superiority versus popular models used in the literature. We test these methods using a comprehensive dataset of more than one million financial statements covering the entire universe of privately held Norwegian SMEs in 2006-2017. Our methods can choose among 155 accounting-based input variables derived from prior literature. We find that the input variables chosen by an embedded least absolute shrinkage and selection operator (LASSO) method yield the best in-sample fit and out-of-sample performance. We show in a simulation, which mimics a real-world competitive credit market, that using LASSO to choose bankruptcy predictors improves credit risk pricing and decision making, resulting in significantly higher bank profits. Finally, we show that model performance can be further improved by running feature selection methods on sub-sets of the company universe, such as for example within-industry.

JEL classification codes: G33, G17, M41, C25

Keywords: Bankruptcy prediction, Feature selection methods, LASSO, Deep learning, Bank profitability

[⋆]We appreciate many insightful comments from Edward Altman, Taha Ahsin, Barbara Będowska-Sójka, Oliver Binz, Christian Brownlees, Luis Ceballos, Steven Shu-Hsiu Chen, Cathy Yi-Hsuan Chen, Jo Eidsvik, Christian Ewald, Tristan Fitzgerald, Stein-Erik Fleten, Roland Füss, Jairaj Gupta, Wolfgang Karl Härdle, Rüdiger Kiesel, Maria Lavrutich, Brian Jonghwan Lee, Codruta Mare, Daniel Mittendorf, Jens Perch Nielsen, Arild Brandrud Næss, Steven Ongena, Markus Pelger, Franziska Peter, Gabriele Sabato, Dirk Schiereck, Kristina Šutienė, Paul Söderlind, Shaonan Tian, Aljoša Valentinčič, Yan Yu, and Alan Zhang, and seminar and conference participants at the 44th Annual Congress of the European Accounting Association, the Eastern Finance Association (EFA) 2022 Annual Meeting in Washington, DC, the 34th Australasian Finance and Banking Conference, the Financial Management Association (FMA) 2021 Annual Meeting in Denver, the STAT OF ML 2021 conference in Prague, the COST FinAI Annual Meeting in Bucharest, the 2021 NTNU Business School Conference, a seminar at Wiser-funding Ltd. London, the 2021 Winter Research Conference on Machine Learning and Business at the University of Miami, the 2nd Yushan Conference, the 4th Shanghai-Edinburgh Fintech Conference and the 6th Fintech International Conference in Shanghai, the 2020 NTNU Business School Conference, the Digital Workshop in Banking and Finance at the NTNU Business School, the internal seminar at the Department of Industrial Economics and Technology Management at NTNU, the 2020 FIBE Conference in Bergen, the Workshop on Banking and Finance at the Faculty of Economics and Management at NTNU, and the 2018 NTNU Business School Conference. We acknowledge research support by COST Action "Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry" (FinAI) CA19130 funded by the Horizon 2020 Framework Programme of the EU.

1. Introduction

Bankruptcy prediction models are important for banks to facilitate the calculation of risk-weighted assets and to assess lending rates. Further, bankruptcy prediction is important for regulators as a key element for the analysis of financial markets and for the on-site supervision of banks (Bernhardsen and Larsen, 2007). There is a vast academic literature on bankruptcy prediction (e.g., Altman, 1968; Taffler, 1984; Zmijewski, 1984; Härdle et al., 2009; Tian et al., 2015). However, the choice of input variables used for bankruptcy prediction is often ad-hoc and still actively debated among both academics and practitioners.[1]

In this paper, we use alternative feature selection methods to derive discrete hazard models for bankruptcy prediction from a comprehensive set of 155 input variables used in previous research. We apply three different feature selection methods: a filter method using the Pearson correlation coefficient as criterion, a wrapper method using the improved floating forward selection algorithm, and an embedded least absolute shrinkage and selection operator (LASSO) method. We combine these feature selection methods with state-of-the art statistical and machine learning estimation techniques for bankruptcy prediction: logistic regression (LR) and a deep artificial neural network (DNN). We then compare the variable sets chosen by the different feature selection methods against each other and against benchmark sets of popular bankruptcy prediction models (Altman, 1968; Taffler, 1984; Zmijewski, 1984; Altman and Sabato, 2007; Bernhardsen and Larsen, 2007).

We apply our feature selection-based bankruptcy prediction models to more than one million financial statements which form the universe of all privately held small and medium-sized enterprises (SMEs) in Norway over the time period 2006-2017. Most prior studies on bankruptcy prediction are limited to larger and listed companies (e.g., Campbell et al., 2008; Tian et al., 2015; Liang et al., 2016) and only very few focus on private companies and SMEs (e.g., Altman and Sabato, 2007). However, SMEs represent a large part of the economy, and loans to SMEs represent a significant proportion of the portfolios of most banks (Blöchlinger and Leippold, 2006). Moreover, lending to SMEs has a strong positive effect on bank profitability and contributes to portfolio diversification, even if it is generally riskier than lending to larger companies (Dietsch and Petey, 2004; Altman and Sabato, 2007). SMEs and newly founded firms rely to a large extent on bank credit (Carbó-Valverde et al., 2009; Robb and Robinson, 2012), which often substitutes expensive trade credit (Fisman and Love, 2003). Still, access to external financing, especially to bank financing, is an important growth con-

[1] See, e.g., Fitzgerald (2009) and Toplensky (2020).

straint for SMEs (Beck and Demirguc-Kunt, 2006; Beck et al., 2008). One probable reason is that bankruptcy prediction is arguably more difficult for privately held SMEs than for public firms because reporting requirements are significantly lower and due to a lack of market-based information. In fact, studies on bankruptcy prediction of private SMEs have to rely exclusively on financial statement information. In our study, we use an extensive set of accounting-based predictors, including the combined set of all variables used in a large number of prior studies. Notwithstanding difficulties in bankruptcy predictions of SMEs, privately held SMEs are particularly well-suited to test bankruptcy prediction models, as bankruptcies are significantly more frequent among SMEs than among large public corporations.[2]

Our empirical results show that feature selection methods produce input variables for predicting bankruptcy that are superior to using variable sets drawn from prior research. When comparing different feature selection methods, our results show that the best in-sample fit and out-of-sample performance are achieved when input variables are chosen by the LASSO and wrapper methods. However, we find that the former yields variables which are much more stable over time. In addition, the LASSO method is computationally less demanding. Hence, the LASSO method overall shows the best performance in predicting bankruptcies of SMEs. Our findings are robust to either using the machine learning estimation technique DNN or the statistical estimation technique LR, as well as across different time periods. Finally, when comparing the estimation techniques, for a given feature selection method, we find that DNN is generally superior to LR, but improvements achieved by DNN over LR are statistically insignificant. These findings apply to both in-sample fit and out-of-sample performance when using common metrics for evaluating bankruptcy prediction models, including the area under the receiver operating characteristic curve (AUC), Brier score, and the Akaike information criterion (AIC).

In the second part of the paper, we investigate the implications of various bankruptcy prediction models on credit risk pricing and decision making, ultimately affecting bank profitability, by mimicking a real-world competitive credit market in a simulation following the framework of Blöchlinger and Leippold (2006). In contrast to commonly used metrics for model evaluation, our simulation accounts for the cost asymmetry between erroneously predicting a bankruptcy event that did not happen versus erroneously not predicting a bankruptcy that materialized. We show that the input variables chosen by the LASSO method result in a lower

[2]In our sample, 1.72% of the 1,002,325 observations are bankruptcies. In contrast, samples including public corporations include much less bankruptcies. For example, Campbell et al. (2008) report an average rate of bankruptcies of 0.52% in their sample of companies over the years 1963 to 1998, while the rate of annual financial statements defined as bankrupt by Tian and Yu (2017) is 0.19% in their study of British, French, German, and Japanese companies from 1998 to 2012.

proportion of bankrupt borrowers in the loan portfolio compared to input variables included in popular bankruptcy prediction models or variables chosen by the filter and wrapper methods. Moreover, we show that the LASSO method yields the highest bank profit over the sample period, independent of different configurations of the simulation. Specifically, using the set of variables selected by the LASSO method results in an average decrease in the proportion of bankrupt borrowers in the loan portfolio by 74% and 25% and an average increase in bank profits by 49% and 84%, respectively, compared to the variable sets yielding the second and third highest bank profits. This analysis highlights the importance of carefully choosing the bankruptcy prediction model as even small differences in model performance, as determined by commonly used metrics for model evaluation, such as for instance AUC, can significantly impact bank profitability. Further, we show in a second simulation that even higher bank profitability can be achieved if the LASSO method is specified at the industry-level.

Our paper is not the first to analyze the potential value added by applying feature selection methods and machine learning techniques to bankruptcy prediction models. Our main contribution to the prior literature on bankruptcy prediction is to compare the in- and out-of-sample predictive power of different feature selection methods in combination with state-of-the art statistical and machine learning estimation techniques for bankruptcy prediction, and to compare them to popular benchmark models, such as Altman (1968) or Altman and Sabato (2007), among others. Moreover, we assess the stability of different feature selection methods. Finally, besides using standard metrics of model evaluation, we conduct a simulation study, which mimics a real-world competitive credit market, to shed light on how the use of different bankruptcy prediction models affects banks' market shares, credit spreads, and profitability.

A paper related to ours is Liang et al. (2016) who use feature selection methods to reduce irrelevant variables when assessing the predictive power of financial ratios versus non-financial corporate governance indicators in samples of listed Taiwanese and Chinese firms. The authors conclude that the model's in-sample performance is slightly better when using a wrapper feature selection method compared to using filter methods or no feature selection method at all. However, the authors do not consider LASSO or other embedded methods, which we show to be superior to any other feature selection method. Moreover, they do not compare the performance of models relying on feature selection methods to the performance of popular bankruptcy prediction models proposed in prior literature. Finally, they do not analyze the stability of variables chosen by feature selection methods. Another paper closely related to ours is Härdle et al. (2009) who explore the suitability of smooth support vector machines (SSVM) as an estimation technique for predicting bankruptcy. They also investigate the effect of different feature selection methods on the predictive power of the models. However, this analysis is conditioned on using SSVM. Additional papers related to ours are Tian et al. (2015) and Tian and Yu (2017). However, they do not compare different feature selection methods, but compare one, LASSO, to popular bankruptcy prediction models proposed in prior literature and find that, not surprisingly, LASSO is superior. Moreover, the focus of both these papers is different from ours. Both studies use datasets of public corporations and Tian et al. (2015) focus on the predictive power of accounting-based versus market-based predictor variables. Tian and Yu (2017) apply their models to an international dataset of listed firms to compare differences in chosen predictor variables across different countries. Finally, a study that analyzes SMEs is Altman and Sabato (2007) who derive a model for bankruptcy prediction that specifically targets SMEs. To this end, they use a sample of about 2,000 U.S. SMEs. However, they do not use feature selection methods to choose the predictors, but rely on a forward stepwise selection procedure to select five predictors out of a set of 17 popular accounting-based variables used in prior literature.

Our paper is also related to the recent literature in financial economics that uses machine learning estimation techniques to select predictive variables more generally. Gu et al. (2020), for example, compare the ability of alternative machine learning techniques to identify factors that predict asset prices. More generally, machine learning techniques are increasingly used to tame the factor zoo (see, e.g., Bryzgalova et al., 2021; DeMiguel et al., 2020; Feng et al., 2020; Freyberger et al., 2020; Kozak et al., 2020), to derive conditioning information in asset pricing models (Chen et al., 2021), or to form mean-variance efficient portfolios (e.g., Rossi, 2018; Bryzgalova et al., 2021; Chen et al., 2021; Cong et al., 2020). Chinco et al. (2019) use the LASSO method to predict rolling one-minute-ahead return forecasts and find that it significantly improves out-of-sample fit by identifying predictors that are associated with economically meaningful events.[3] The LASSO method is also applied to other areas in finance than asset pricing. Hautsch et al. (2014) and Ogneva et al. (2019), for example, use LASSO for the selection of relevant variables for measuring systemic risk. Han and Kong (2020) use machine learning estimation techniques to study the serial dependence of commodity futures returns and assess in- and out-of-sample predictability. They find that out-of-sample forecasts based on LASSO are superior to more complex machine learning techniques, such as for instance regression trees. Machine learning techniques are also applied in real estate research. For instance, Sadhwani et al. (2021) use DNNs to estimate mortgage prepayments, delinquencies, and foreclosures. Finally, machine learning techniques are also applied in text analysis.

[3] For a recent survey of asset pricing research using machine learning estimation techniques for factor selection, see Weigand (2019).

For instance, Calomiris and Mamaysky (2019) and Adämmer and Schüssler (2020) use contents from newspaper articles to predict stock returns. The former study employs the LASSO framework to explore dynamic changes in coefficients, while the latter shows that the LASSO framework leads to improved forecasts. Garcia et al. (2021) and Kelly et al. (2021) use machine learning algorithms to classify the sentiment in congressional speeches and news articles as well as in conference call scripts, respectively. Azimi and Agrawal (2021) use recurrent neural networks to measure sentiment in 10K filings and find that sentiment measured using machine learn techniques predicts abnormal returns. Binz et al. (2022), Cao and You (2021), and Chen et al. (2022) apply machine learning techniques to forecast corporate earnings. Zhang (2021) applies natural language processing models and neural networks to predict mutual fund performance using the textual information disclosed in mutual fund shareholder letters.

2. Data and variables

2.1. Sample

In this study, we analyze all unconsolidated annual financial statements of the universe of privately held Norwegian limited liability SMEs from the accounting years 2006-2017. Dates of all filings for bankruptcy by these companies in the years 2006-2020 are provided by the Norwegian government agency Brønnøysund Register Centre. Following the EU recommendation 2003/361, we define a company as SME if its turnover is not more than EUR 50 millions or total assets are not more than EUR 43 millions.[4] As very small firms usually have no or very limited debt on their books, with capital mostly put up by the owners, we restrict our sample to firms with total assets above NOK 500,000. By doing so, we follow the central bank of Norway in the development of their so-called "SEBRA model", a bankruptcy prediction model widely used by Norwegian authorities (Bernhardsen and Larsen, 2007). Note that such small firms are also prone to outlier and data errors problems.

Following prior literature (e.g., Mansi et al., 2012), we furthermore exclude financial statements of companies in certain industries. Specifically, we exclude financial statements of companies in the industries *'Financial and insurance activities'*, *'Real estate activities'*, *'Public administration and defence'*, *'Electricity and gas supply'*, and *'Water supply, sewerage, waste'*.[5]

[4]A somewhat similar definition is found in Basel III in which an SME is a company with sales not larger than EUR 50 millions. The EU recommendation 2003/361 defines a company as SME based on, firstly, the number of employees and, secondly, either turnover or total assets. We rely on the latter classification since the number of employees is not available in our data.

[5]We categorize financial statements in the dataset into industries according to the highest level, i.e., sections, in the Norwegian Standard Industrial Classification (SIC2007). This standard is based on UN's ISIC Rev. 4 and EU's NACE Rev. 2. See www.ssb.no/nace for an overview.

We also exclude financial statements of companies which have not reported any industry including companies for investment and holding purposes only. Our final sample includes 1,002,325 financial statements, published by 192,953 companies, that pass these filters.

We categorize the financial statements as bankrupt or non-bankrupt by following the definition of the central bank of Norway in the development of their SEBRA model (Bernhardsen and Larsen, 2007). In particular, we categorize a financial statement as bankrupt if *i)* it is the last financial statement of the underlying company and *ii)* the company has filed for bankruptcy within three years from the balance sheet date of this last financial statement. Our classification results in 17,290 (or 1.72%) of the 1,002,325 financial statements in our sample being eventually categorized as bankrupt. We categorize all other financial statements as non-bankrupt.[6]

2.2. Input variables

From each financial statement, we retrieve values for 155 input variables, which our feature selection methods can choose from. As our sample includes private SMEs, for which no market data is available, we are forced to restrict the set of predictors to accounting-based variables.[7] Our choice of 155 input variables is based on prior bankruptcy studies, including Altman (1968), Ohlson (1980), Taffler (1984), Zmijewski (1984), Shumway (2001), Altman and Sabato (2007), Campbell et al. (2008), Härdle et al. (2009), Tian et al. (2015), Tian and Yu (2017), Kumar and Ravi (2007), Liang et al. (2016), and the SEBRA model (Bernhardsen and Larsen, 2007).

Earlier studies on bankruptcy prediction predominantly consider accounting-based input variables collected from financial statements data (e.g., Beaver, 1966; Altman, 1968; Ohlson, 1980; Taffler, 1984; Zmijewski, 1984). More recent research, that restricts the sample to public firms, started to add market-based input variables to the set of bankruptcy predictors (e.g., Shumway, 2001; Chava and Jarrow, 2004; Hillegeist et al., 2004; Campbell et al., 2008; Tian et al., 2015; Blöchlinger and Leippold, 2018). However, even when analyzing public firms, accounting-based models work well and have been shown to even outperform more recent and larger models including market-based variables. For instance, Reisz and Perlich (2007)

[6]Note that our sample period does not include meaningful variation in the business cycle as Norway was only mildly affected by both the financial crisis of 2007-2008 and the subsequent European debt crisis. For instance, the 12-month decline of quarterly real GDP, from the fourth quarter of 2008 to the third quarter of 2009, was 1.2% in Norway. In comparison, real GDP declined by 3.3%, 4.2%, and 3.8% in the US, EU, and OECD, respectively. These data are retrieved from the OECD at: https://stats.oecd.org/index.aspx?queryid=350. Hence, we abstain from performing business cycle related tests of our bankruptcy prediction models.

[7]As in the revisions of the Altman (1968) model (see Altman, 2018; Altman et al., 2019), in some of the financial ratios we use in our study, we replace the market value of equity – as used in the original study – by the book value of equity.

find that the accounting-based models of Altman (1968, 1993) outperform three market-based models when predicting bankruptcy over a one year period. Agarwal and Taffler (2008) find that the accounting-based model of Taffler (1984) produces a significant economic benefit over the market-based models of Bharath and Shumway (2008) and Hillegeist et al. (2004) when taking into account misclassification costs and loan pricing considerations as proposed by Stein (2005) and Blöchlinger and Leippold (2006).

All 155 input variables are listed in Table A.1 in the appendix. 151 of the variables are financial ratios, two are dummy variables, and the other two variables are the company age in years and total assets as a measure of company size.[8] Data on the complete set of 155 input variables is available in all financial statements in our sample. Hence, we do not have to deal with missing values. To mitigate the effect of outliers, we follow prior literature and winsorize all financial ratios at their 1st and 99th percentiles within each accounting year (e.g., Shumway, 2001; Chava and Jarrow, 2004; Tian et al., 2015; Tian and Yu, 2017).[9] Finally, we log-transform the variables 'age in years' and 'total assets'.

2.3. Benchmark bankruptcy prediction models

As we analyze private SMEs, the set of input variables that our feature selection methods can choose from are necessarily confined to accounting-based variables. Hence, we are confined to using purely accounting-based bankruptcy prediction models as benchmarks for our own, feature selection-based models. Certainly among the most popular models, and still widely used today, are the classic models of Altman (1968), Ohlson (1980), and Zmijewski (1984).[10] The output of these models is also employed in other areas of finance research, such as for instance in asset pricing studies to predict stock returns (Light et al., 2017; Feng et al., 2020), to measure the probability of debt renegotiations (Chang et al., 2019), or to determine gains of firms that acquire assets from distressed companies (Meier and Servaes, 2019). Further studies employ these models to assess the association between bankruptcy risk and operating leverage (Chen et al., 2019) or between bankruptcy risk and worker unionization (Campello et al., 2017). Moreover, Blöchlinger and Leippold (2018) combine the model of Altman (1968) and the market-based distance-to-default model of Merton (1974) and show that this combina-

[8] The company age in years is measured as the number of years between the balance sheet date for the financial statement and the date of incorporation of the underlying company.

[9] If the denominator of a financial ratio is equal to zero, we set the value of the variable to zero, if the numerator is equal to zero as well. If the numerator is negative or positive, we set the ratio to the 1st or 99th percentiles of the distribution of the ratio, respectively.

[10] E.g., Begley et al. (1996); Chava and Jarrow (2004); Mansi et al. (2012); Tian et al. (2015); Tian and Yu (2017).

tion outperforms Standard & Poor's ratings and any of the two models individually.

The model of Altman (1968) contains five accounting-based input variables categorized into the main aspects of a company's financial profile: liquidity, profitability, leverage, solvency (coverage), and activity. However, Altman (1968) finds that the input variable measuring activity is insignificant. He shows that this variable is highly industry-sensitive and thus excludes this variable in a revised version of the model that targets private companies across all industries (Altman, 2018; Altman et al., 2019). Similar to the revised version of Altman's (1968) model, the nine variables in the model of Ohlson (1980) measure liquidity, profitability, leverage, and solvency, in addition to company size weighted by gross national income. The model of Zmijewski (1984) contains three accounting-based input variables measuring liquidity, profitability, and leverage.

Another widely-used model is that of Taffler (1984) which contains four accounting-based input variables which measure profitability, working capital position, financial risk, and liquidity. In contrast to the models of Altman (1968), Ohlson (1980), and Zmijewski (1984), which were all specified using listed U.S. industrial companies, Taffler (1984) applies his model to listed U.K. industrial companies. Agarwal and Taffler (2007, 2008) show that the model still performs well using more recent data.

A well-known and widely used bankruptcy prediction model that specifically targets private SMEs is the model of Altman and Sabato (2007). They apply a forward stepwise selection procedure to choose predictors from a set of 17 input variables found to be the most successful in prior literature. Similar to the Altman (1968) model, the resulting model contains five accounting-based input variables categorized into the five main aspects of a company's financial profile.

Our final benchmark model is the SEBRA model by the central bank of Norway (Eklund et al., 2001; Bernhardsen and Larsen, 2007). The model was calibrated using a sample of Norwegian limited liability companies and contains eight accounting-based input variables meant to capture firms' liquidity, profitability, solvency, age, and size. The SEBRA model is widely used in Norway, for instance by the Financial Supervisory Authority of Norway for analyzing the financial markets and conducting on-site supervision of banks.

3. Methodology

3.1. Test setting

We predict one-year-ahead bankruptcy, which is the most common prediction horizon used by both practitioners and academics (Hillegeist et al., 2004; Tian et al., 2015; Tian and Yu, 2017). Moreover, Basel III focuses on one-year-ahead forecast horizons for corporate and bank exposures. In conducting our bankruptcy prediction analysis, we use a rolling window and a forward validation procedure (Kaastra and Boyd, 1996; Keles et al., 2016). Further, we follow Shumway (2001) by using discrete hazard models which is also the convention in recent literature.[11] This is in contrast to static models typically applied in earlier studies.[12] Static models use only a single firm-year observation for each company in the sample and thus may produce biased estimates as they ignore that companies change over time. Discrete hazard models, on the other hand, consider the hazard of bankruptcy over multiple years. They are therefore preferable as they adjust for period at risk, incorporate time-varying predictors, and utilize much more data which typically yield better predictions.

Specifically, we select input variables, train models, and evaluate their in-sample fit using all financial statements over four consecutive accounting years. We then evaluate the models' out-of-sample modeling performance using all financial statements in the subsequent accounting year. The first training period is 2006-2009 and the subsequent out-of-sample evaluation period is 2010. We then move forward the entire analysis by one year and train or models in the years 2007-2010 and perform out-of-sample tests in the year 2011. The last training period is 2013-2016 and the final out-of-sample test period is 2017. This test setting results in eight consecutive test periods and is illustrated in Figure 1.

[Figure 1 about here.]

3.2. Estimation techniques

Besides to selecting input variables, the other major model choice to be made in any bankruptcy prediction model is that of the estimation technique. Early studies typically use discriminant analysis (e.g., Altman, 1968; Altman et al., 1977). However, discriminant analysis relies on restrictive assumptions (Joy and Tollefson, 1975; Deakin, 1976; Eisenbeis, 1977). Consequently, Martin (1977) and Ohlson (1980) estimate their models with logistic regressions (LR) which rely on less restrictive assumptions and produce more intuitive output.

Recently, bankruptcy prediction models started to rely on machine learning techniques for model estimation. Among these, artificial neural networks are the most widely used technique in bankruptcy prediction (Kumar and Ravi, 2007). Artificial neural networks have been shown to perform well in various machine learning applications (Kumar and Ravi, 2007; Bianchi et al.,

[11]E.g., Chava and Jarrow (2004); Campbell et al. (2008); Bauer and Agarwal (2014); Tian et al. (2015); Tian and Yu (2017).

[12]E.g., Altman (1968); Ohlson (1980); Zmijewski (1984).

2021). The main benefit of artificial neural networks is that they capture multivariate non-linear relations without making any assumptions with respect to data or errors. As our study focuses on feature selection methods, comparisons of estimation techniques are beyond the scope of this study.[13] Instead, we apply the most popular estimation technique by using a deep artificial neural network (DNN), and use LR for comparison reasons.

When using LR as estimation technique, the vector of predicted probabilities for bankruptcy $\hat{\mathbf{y}} = \{\hat{y}_n\}_{n=1,\dots,N} \in [0,1]^N$ is given by

$$\hat{\mathbf{y}} = \iota \oslash (\iota + \exp(-\mathbf{X}\mathbf{w} - \iota w_0)) \tag{1}$$

where $\mathbf{X} = \{x_{(n,i)}\}_{n=1,\dots,N,i=1,\dots,I}$ is a matrix of values for input variables i derived from the financial statements n, $\mathbf{w} = \{w_i\}_{i=1,\dots,I}$ is a vector of coefficients, w_0 is the intercept coefficient, ι is an $N \times 1$ vector of ones, and $\oslash$ denotes Hadamard (element-wise) division. For ease of notation, we drop the time indices. We estimate the coefficients $\mathbf{w}$ and w_0 by minimizing the negative of the log-likelihood function $\ell(\mathbf{w}, w_0)$ given by

$$\ell(\mathbf{w}, w_0) = \sum_{n=1}^{N} [\mathbf{y} \odot (\mathbf{X}\mathbf{w} + \iota w_0) - \log(\iota + \exp(\mathbf{X}\mathbf{w} + \iota w_0))] \tag{2}$$

where $\mathbf{y} = \{y_n\}_{n=1,\dots,N} \in \{0,1\}^N$ is the vector of actual classifications of non-bankrupt (0) or bankrupt (1), and $\odot$ denotes the Hadamard (element-wise) product. For determining the significance of estimated LR coefficients, we use the Wald statistic (Hosmer et al., 2013) to derive their z-scores.

We use a feedforward DNN with one input layer, two hidden layers, and one output layer, as illustrated in Figure 2. The input layer consists of the input variables. The output layer consists of one node representing the output of the DNN, with values between 0 (non-bankrupt) and 1 (bankrupt). The hidden layers each contain five nodes. Figure 3 illustrates the computation of the values of the nodes in the hidden and output layers, respectively. Further details on the DNN and how we train it is provided in Appendix B. LR and DNN share several characteristics. Indeed, the former is a special case of the latter with no hidden layers and the logistic transfer function $\phi(a) = \iota \oslash (\iota + \exp(-a))$ between the input and output layers. When compared to LR, our DNN have much more flexibility.

[Figure 2 about here.]

[13] We refer to Jones et al. (2015, 2017) for an empirical evaluation of different estimation techniques used for bankruptcy prediction and in related applications.

[Figure 3 about here.]

3.3. Feature selection methods

Feature selection methods attempt to select a subset of relevant variables, or features, for use in an empirical model.[14] The appeal of feature selection methods is that they are generally applicable, help to eliminate irrelevant variables, help to provide a better understanding of the data structure, help to reduce computational requirements, and can avoid overfit. Moreover, they can improve the prediction performance by reducing the high dimensionality problem (Chandrashekar and Sahin, 2014; Tian et al., 2015; Chinco et al., 2019; DeMiguel et al., 2020; Freyberger et al., 2020; Gu et al., 2020; Kozak et al., 2020). Studies of bankruptcy prediction recently started to employ feature selection methods (e.g., Härdle et al., 2009; Tian et al., 2015; Liang et al., 2016; Tian and Yu, 2017).

Feature selection methods are often categorized into three types of methods (e.g., Chandrashekar and Sahin, 2014): filter, wrapper, and embedded methods. We use three feature selection methods, one within each of these three groups.

3.3.1. The filter feature selection method

Filter methods select the input variables that are ranked highest according to a predefined criterion that measures the relationship between the single input variables and what is to be modeled. We use as criterion for each input variable the Pearson correlation coefficient between the predictor variables' values and the actual classifications of non-bankrupt (0) or bankrupt (1) across all financial statements used for training.

Filter methods are computationally light, avoid overfitting and do not rely on potentially biased estimation techniques (Chandrashekar and Sahin, 2014). However, filter methods come with several disadvantages. First, they do not consider relations between the input variables. Second, when using correlations as selection criterion, filter methods consider only linear relationships. Finally, filter methods may yield suboptimal fit and prediction power as they ignore the biases of the estimation technique, e.g., LR or DNN, and thus the induced structure of the final model, i.e., the selected variables combined with the estimation technique (John et al., 1994). These shortcomings of filter methods are neither shared by the wrapper nor the embedded methods and can thus be overcome by using any of these techniques.

[14] An alternative to feature selection methods are feature extraction methods, such as factor analysis or principal component analysis (PCA). However, such methods do not produce subsets of the original variables but new variables from functions of the original variables. Feature selection methods are therefore better suited if economic interpretation of the model and its predictions are a central aspect of the analysis.

3.3.2. *The wrapper feature selection method*

Wrapper methods select input variable sets heuristically (John et al., 1994; Kohavi and John, 1997). They start with a variable set which either is empty or contains all variables available for selection. From this initial set, one or more variables are added or removed iteratively in accordance with the steps of an algorithm. For each of these steps, a model is created for each possible new variable set by combining it with a predefined estimation technique. These models are then ranked based on their performance on the data in accordance with a predefined evaluation metric. Among all possible new variable sets, the one which yields the highest ranking model is selected and used as the initial set for the next step of the algorithm, which is once again adding or removing one or more variables. When a predetermined number of variables in the set is reached, the algorithm stops and the current variable set is the one finally selected by the wrapper method.

In our study, we use the improved forward floating selection algorithm proposed by Nakariyakul and Casasent (2009). Moreover, we apply the same estimation techniques that we use later when training the final model, i.e., LR and DNN. Finally, we evaluate all possible new variable sets in each step of the algorithm based on in-sample-fit.[15] This ensures that the predictive power that we report for the variable sets selected by the wrapper method are indeed out-of-sample.

Wrapper methods have been criticized for allowing locally optimal solutions and for being prone to overfitting (Kohavi and John, 1997). To mitigate these issues, we cross-validate our results across several subsets of the data for the years 2010-2017, use an extended algorithm, and investigate the robustness of our findings across two different estimation techniques. Another critique of wrapper methods is that they are computationally intensive (Chandrashekar and Sahin, 2014). However, all these shortcomings are relieved by embedded methods.

3.3.3. *The embedded LASSO feature selection method*

The embedded LASSO method, popularized by Tibshirani (1996), incorporates feature selection as part of training. When applying this method, we estimate the coefficients $\mathbf{w}$ and w_0 by minimizing

$$-\ell(\mathbf{w}, w_0) + \lambda\|\mathbf{w}\|_1 \tag{3}$$

[15] In particular, we evaluate using in-sample AUC, as described in Section 3.4.

in the training data set, where $\ell(\mathbf{w}, w_0)$ is the log-likelihood given in Equation (2), $||\mathbf{w}||_1$ is the l_1-norm of $\mathbf{w}$ (i.e., $\sum_{i=1}^{I}|w_i|$), and λ is a positive tuning parameter.[16] If λ is sufficiently large, the penalty term $\lambda||\mathbf{w}||_1$ becomes so dominant that all the estimated coefficients in $\mathbf{w}$ become zero. We select input variables with the LASSO method by iteratively re-estimating $\mathbf{w}$ and w_0 for different values of λ given by the path proposed in Friedman et al. (2010): First, λ is set to λ_L which is the highest value possible that results in at least one of the values in $\mathbf{w}$ being non-zero. Second, λ is reduced following a geometric sequence between λ_L and $\lambda_L \times 10^{-5}$ of size 5×10^3 which causes gradually more values in $\mathbf{w}$ to become non-zero. Finally, the reduction of λ stops when the number of non-zero values in $\mathbf{w}$ match a predetermined number of variables in the variable set. The input variables associated with these non-zero values in $\mathbf{w}$ are those selected by the LASSO method.[17]

3.4. Evaluation metrics

We evaluate model performance across our alternative bankruptcy prediction models using several metrics. First, we evaluate both in-sample fit and out-of-sample performance using AUC (i.e., area under the receiver operating characteristic curve). This metric has been widely used for evaluating bankruptcy prediction models (e.g., Bernhardsen and Larsen, 2007; Tian et al., 2015; Tian and Yu, 2017; Nagel and Purnanandam, 2019). AUC represents the area under the plot of the true positive rate against the false positive rate across all observations when varying the discrimination threshold (Hosmer et al., 2013).[18] AUC $\in [0.5, 1]$, where a higher value indicates a higher explanatory power of the model. Following Hosmer et al. (2013), we consider AUC $\in [0.7, 0.8)$ acceptable, AUC $\in [0.8, 0.9)$ excellent, and AUC ≥ 0.9 outstanding.

Second, following Tian et al. (2015) and Tian and Yu (2017), we evaluate model performance using the Brier (1950) score and the AIC (Akaike, 1974). The Brier score can be thought of as a cost function. Specifically, the Brier score measures the mean squared difference between *i)* the predicted probability assigned to the possible outcomes for a set of predictors and

[16] In order to compare the input variables on an equal scale, we standardize the input variables in the training set to have zero mean and a variance of one before performing the minimization in Equation (3).

[17] In unreported analyzes, we alternatively perform feature selection using the embedded elastic net method. Zou and Hastie (2005) introduce this method to address shortcomings of the LASSO method, such as unsatisfactory results when $I \gg N$ and the tendency of selecting only one input variable among any group of highly correlated input variables. The elastic net replaces $||\mathbf{w}||_1$ in Equation (3) with $\frac{1-\alpha}{2}||\mathbf{w}||_2^2 + \alpha||\mathbf{w}||_1 = \sum_{i=1}^{I}\left(\frac{1-\alpha}{2}w_i^2 + \alpha|w_i|\right)$, where $\alpha \in [0, 1]$ is predetermined. The elastic net equals the LASSO method when $\alpha = 1$. We perform feature selection using the elastic net with different values of α. However, we find that the LASSO method yields better performance across all model evaluation metrics. Details and results are available upon request.

[18] In our setting, the discrimination thresholds vary between 0 and 1, which is the output interval of our models.

ii) the actual outcome. Hence, the lower the Brier score for a set of predictors, the better is the prediction calibrated. AIC estimates the relative amount of information lost by a given model. The less information a model loses, the higher the quality of the model. AIC considers the trade-off between the goodness of fit of a model and the simplicity of the model. Specifically, AIC takes into account the tradeoff between bias and variance by introducing a penalty term proportional to the number of input variables. AIC is defined as: $\text{AIC} = -2 \times \ell(\mathbf{w}, w_0) + 2 \times I$. As with the Brier score, a lower AIC value indicates a better prediction.[19]

Third, we evaluate the in-sample fit of LR models using McFadden's (1974) pseudo-R squared (R^2), as done among others by Campbell et al. (2008) and Tian et al. (2015). It is given by $R^2 = 1 - \frac{\ell(\mathbf{w},w_0)}{\ell(w_0)} \in [0, 1]$ where $\ell(w_0)$ is the log likelihood of a model containing no input variables but only the intercept coefficient w_0. Hence, higher values of the pseudo-R squared indicate a better model performance.

Finally, we evaluate the out-of-sample performance of our models using decile rankings. Decile rankings are widely applied to evaluate the model performance of bankruptcy prediction models (e.g., Shumway, 2001; Chava and Jarrow, 2004; Tian et al., 2015; Tian and Yu, 2017). To do so, we group all financial statements used for evaluating our models into deciles based on the estimated probability of bankruptcy. Financial statements that are predicted by the model to have the highest probability of bankruptcy are sorted into the first decile and statements predicted to have the lowest probability of bankruptcy are sorted into decile 10. Model performance is then evaluated by considering the percentage of financial statements that went bankrupt within each decile.

We use these evaluation metrics to rank the feature selection methods based on the explanatory power of their chosen input variables. Besides such a comparison of general model performance, we attempt to assess the stability of the feature selection methods, i.e., their tendency to choose the same variables when presented to different subsets of the data. Notwithstanding the flexibility of our feature selection-based bankruptcy prediction models to regularly vary the set of predictive variables, an evaluation of model stability is still potentially important for at least two reasons. First, frequent changes to the set of predictors make it difficult to interpret the relationship between input variables and the outcome, i.e., bankruptcy, in a time-consistent

[19]In unreported analyzes, we alternatively compute the values of two extensions of AIC: the corrected AIC (Hurvich and Tsai, 1989) and Schwarz's Bayesian information criterion (BIC) (Schwarz, 1978). The former introduces an extra penalty term meant to address the tendency of AIC to favor models with too many parameters, in particular when the dataset is small. The corrected AIC is defined as $AIC + \frac{2\times I^2+2\times I}{N-I-1}$. The BIC has a larger penalty term than AIC. The penalty term is $I * \log(N)$. When applied to our models, these alternative criteria yield values that are almost identical to AIC. Results are available upon request.

manner. Second, the confidence of industry representatives and domain experts into proposed methods may suffer if a method regularly selects very different variable sets, in particular if the economic environment does not change significantly. We compare the stability of the feature selection methods by considering the number of unique input variables they choose across all the eight evaluation periods covered by our sample period.

4. Results

4.1. Results from analysis using feature selection methods

We employ the wrapper, LASSO, and filter feature selection methods to generate input variable sets which we label WRA, LAS, and FIL, respectively. In addition, we generate sets of randomly selected input variables (RAN). Further, we compute and compare the in- and out-of-sample performance achieved with WRA, LAS, FIL, and RAN when combined with either LR or DNN as estimation technique. We also compare the performance of the alternative feature selection methods to that of popular models proposed in previous literature, including the model of Altman (1968) (ALT), Taffler (1984) (TAF), Zmijewski (1984) (ZMI), Altman and Sabato (2007) (ASA), and the SEBRA model (SEB). Table 1 gives an overview of the alternative bankruptcy prediction models, while Appendix C lists the input variables in the benchmark sets.

[Table 1 about here.]

Figure 4 shows the mean of AUC across the eight evaluations for the years 2010-2017 obtained when using the different input variable sets. For WRA, LAS, FIL, and RAN, we vary the predetermined number of input variables in the sets between 1 and 15. The top panels show in-sample fit while the bottom panels show out-of-sample performance when using LR (left) or the DNN (right), respectively. The panels shows that the sets selected by the feature selection methods and the benchmark sets provide an acceptable or excellent explanatory power as AUC = $[0.7, 0.9)$. However, the variable sets selected by the wrapper and LASSO methods yield the highest AUC.

Furthermore, Figure 4 shows that AUC increases in the size of the selected variable sets. Hence, by allowing feature selection methods to select a larger number of input variables the model fit can be further improved. This flexibility to extend the variable sets chosen by the model is an advantage of feature selection methods over benchmark models that include a pre-defined set of predictors. Within the group of benchmark models, the SEBRA model clearly shows the best performance. This finding is not surprising as the SEBRA model is calibrated using recent Norwegian data, while the other benchmark models are calibrated using other markets. This finding points to the importance of re-evaluating bankruptcy prediction models regularly and a market-specific selection of input variables. In contrast, models based on randomly chosen input variables perform worst in all settings. Finally, Figure 4 shows that models estimated with DNN outperform models estimated with LR. However, the difference is small. The results in Figure 4 remain qualitatively similar when evaluating the models using AIC or Brier Score instead of AUC.[20]

[Figure 4 about here.]

Figure 5 displays results on the stability of the feature selection methods based on the number of unique input variables chosen by each method across all eight evaluations, for predetermined number of predictors. The green diagonal line represents the minimum number of input variables that could be obtained in theory. The line is achieved when a feature selection method selects exactly the same input variables across all eight evaluations and thus represents the highest possible model stability. The red diagonal line represents the maximum number of input variables that could be chosen by the models in theory. Specifically, the red line represents a situation in which a feature selection method selects entirely different input variables across all eight evaluations. Remember that the input variables selected by the wrapper method depend on the estimation technique used within its algorithm. Hence, we report results on WRA in combination with both LR and DNN. The figure shows that the LASSO and filter methods are more stable than the wrapper method. This finding is consistent with results reported in Tian et al. (2015), which show that the input variables selected by LASSO are fairly stable over time.

[Figure 5 about here.]

To conclude, bankruptcy prediction models based on input variables chosen by feature selection methods outperform various benchmark bankruptcy prediction models proposed in prior literature. Within the group of models based on feature selection methods, LASSO provides a superior combination of in-sample fit, out-of-sample performance, and model stability. Moreover, both the filter and wrapper methods have several drawbacks: While the former does not consider correlations between the input variables, the latter selects variable sets heuristically and is computationally intensive. LASSO is computationally efficient and derives the optimal variable set as part of model training and thus overcomes these drawbacks. Consequently, we will henceforth focus on the input variables chosen by the LASSO method.

[20]Details and results are available upon request.

4.2. Selected variables for bankruptcy prediction

Figure 4 shows that both in- and out-of-sample model performance increases in the number of input variables. However, the optimal number of variables to be included in a bankruptcy prediction model is subject to a trade-off between model complexity and performance. Figure 6 shows the marginal percentage increase in in-sample (left panel) and out-of-sample (right panel) AUC values when expanding the number of input variables selected by the LASSO method. Based on the analysis in Figure 6, which shows that the marginal increase in AUC drops below 0.3% in all cases at nine input variables, we examine variables sets of a maximum of nine input variables in the remainder of the paper.

[Figure 6 about here.]

Table 2 shows the input variables selected by the LASSO method across the eight evaluation periods covering the years 2010-2017, when the predetermined number of variables in the model is set to nine. Panel A covers the years 2010-2013 and Panel B the years 2014-2017, respectively. The table reports coefficient estimates of the variables when the models are estimated with LR and z-scores in parentheses. The table shows that all coefficient estimates are statistically significant across all years. Further, the table reports in-sample fit measured by AUC, Brier score, AIC, and R^2 and out-of-sample prediction performance measured by AUC, Brier score, and decile rankings. The results reported in Table 2 show that the selected input variables yield good in- and out-of-sample modeling performance, which is quite stable over time.

[Table 2 about here.]

Overall, LASSO selects 10 input variables. The nine variables chosen in the years 2010-2016 are identical, again pointing to a high model stability. In 2017, *"short-term liquidity / current assets"* is replaced by *"inventory / current assets"*. Both of these variables represent measures of liquidity, suggesting that LASSO selects variables measuring the same aspects of a company's financial profile across all sample years.

We assign our 10 predictor variables to five categories, which are based on Altman (1968) and Altman and Sabato (2007) and meant to cover the main aspects of a company's financial profile. Table 2 shows that the variables selected by LASSO are well diversified across categories. Specifically, nine of the 10 predictors cover four of the five main aspects of a company's financial profile: liquidity, profitability, leverage, and solvency. One of the 10 predictors chosen, company age, does not fit into any of the five categories proposed by Altman (1968) and Altman and Sabato (2007). Moreover, none of the variables covers the fifth main aspect, activity. This latter finding is consistent with results in Altman (1968), who shows that the input variable measuring activity is insignificant. In a later study, Altman et al. (2019) provide evidence suggesting that the variable is industry-sensitive and should be excluded from the model.

Coefficient signs in Table 2 generally go in the expected direction. For instance, the sign on the coefficient of the variable *"accounts payable / total assets"* is positive, indicating that higher debt to suppliers scaled by assets translates into a higher probability of bankruptcy. Measures of financial leverage (i.e.,*"dummy: one if total liability exceeds total assets"* and *"(current liabilities - short-term liquidity) / total asset"*) show a consistently positive coefficient sign, suggesting that higher leverage is associated with a higher bankruptcy probability. The coefficient on *"net income / total assets"* is negative, indicating a lower probability of bankruptcy for more profitable firms. Also, higher outstanding tax liabilities (*"public taxes payable / total assets"*), which may indicate liquidity problems, and higher interest expenses (*"interest expenses / total assets"*) are, as expected, associated with a higher bankruptcy probability.[21] Companies with retained earnings (*"dummy: one if paid-in equity is less than total equity"*) and older, more established firms (*"log(age in years)"*) are less likely to go bankrupt. The latter finding is consistent with previous findings that newly established companies have a higher probability of bankruptcy than older companies (Eklund et al., 2001). Finally, more liquid firms are less likely to go bankrupt, as shown by the positive coefficient on *"inventory / current assets"* and the negative coefficient on *"short-term liquidity / current assets"*.

Moreover, when comparing the predictors selected by LASSO to the variables in the benchmark models, it is obvious that similarities are largest when comparing our 10 predictors to the SEBRA model (shown in Appendix C): First, four of the eight variables included in the SEBRA model are among the 10 variables selected by LASSO. Second, neither the SEBRA model nor our models based on LASSO include variables that pertain to the "activity" category. Finally, variables only included in the SEBRA model, but not selected by LASSO, and variables selected by LASSO, but not included in the SEBRA model, are highly correlated.[22] In summary, these similarities between the variables of the SEBRA model and the variables selected by LASSO point to the importance of constructing market-specific models versus following a

[21]Companies are expected to prioritize payment of tax liabilities as tax authorities have definitive procedures for treating default payments and are often the creditor filing companies for bankruptcy (Eklund et al., 2001).

[22]For instance, across all financial statements in our sample, the correlation between the variable *"total equity / total assets"* in the SEBRA model and the variables *"dummy: one if total liability exceeds total assets"* and *"(current liabilities - short-term liquidity) / total assets"*, as selected by LASSO, is -0.67 and -0.65, respectively.

“one size fits all” strategy and apply the same model across all markets and company segments (e.g., blue chip stocks versus SMEs).

Next, we analyze the relative importance of the individual predictors selected by LASSO. To this end, in Figure 7, we plot their standardized coefficients as obtained from estimating Equation (3) in the 2006-2017 sample period and applying different values of the tuning parameter λ. A sufficiently high λ, and thus a dominant penalty term, results in all coefficients converging to zero. As the penalty term is relaxed, with decreasing values of λ, more variables enter the model as their coefficient values turn non-zero. Variables entering the model at higher values of λ have stronger predictive power, indicating their higher importance. The figure shows that *”accounts payable / total assets”* and *”dummy; one if total liability exceeds total assets”* are clearly the two most important variables. In fact, these two variables are responsible for the large marginal increases in AUC associated with adding the first two variables in Figure 6. When increasing the size of the input variable set to more than five variables, four of the five main aspects of a company’s financial profile, as explained above, are covered by the selected variables.

[Figure 7 about here.]

5. Simulation study: Bankruptcy prediction model choice and bank profitability

Banks can experience costs that arise from two types of bankruptcy prediction errors. The first type of error relates to the failure to correctly predict a company’s bankruptcy. This can cause severe costs such as losses of large portions of granted loans. The second type of error results from an overestimation of the probability of bankruptcy and implicitly of credit spreads assigned to non-bankrupting potential borrowers, causing them to migrate to other banks. Thus, the costs associated with the second type of bankruptcy prediction error are opportunity costs of not lending which, however, have been shown to be lower than the costs implied by the first type of error (Altman et al., 1977; Zmijewski, 1984; Stein, 2005; Balcaen and Ooghe, 2006; Agarwal and Taffler, 2007, 2008; Bauer and Agarwal, 2014; De Bock et al., 2020).

Metrics commonly used for model evaluation, such as accuracy, AUC, Brier score, AIC, and decile rankings, assume equal costs of the different types of errors. Thus, models chosen by these metrics may prove to be unprofitable for banks. Some studies address this using ratios of the relative costs of the different types of errors (e.g., Altman et al., 1977; Frydman et al., 1985; Hopwood et al., 1989; Tam and Kiang, 1992). However, it is difficult to determine the numerator and denominator in such ratios since they depend on many time-dependent factors (De Bock et al., 2020). Hence, in this paper, we evaluate the effect of different model choices and variable sets on bank profitability by mimicking a real-world competitive credit market in a simulation following the mixed framework of Blöchlinger and Leippold (2006). This framework allows for simultaneous credit decision making and credit risk pricing by applying the setting introduced in Stein (2005). In such a simulation, even small differences in model performance according to metrics like AUC can result in huge differences in bank profitability. The reason is that the simulation takes into account the different costs associated with different types of errors, as well as that any borrower enters a loan agreement only with the one bank that offers the lowest credit spreads. Other banks' offers, no matter how close to the lowest offered credit spread, will not result in a new customer relationship.

Our simulation compares the performance of eight variable sets used as input to predict bankruptcy. These include the three sets chosen by the feature selection methods, as well as the five benchmark sets (see Table 1). Further, we employ real world data in the simulation, following Agarwal and Taffler (2007, 2008) and Bauer and Agarwal (2014), by utilizing all the financial statements in our sample from the accounting years 2010-2017. The total size of the simulated credit market amounts to the sum of interest bearing debt derived from these financial statements. Further, each of the interest bearing debt positions is assumed to represent an amount a company wants to borrow. Moreover, the simulated market consists of several banks competing for these potential borrowers. We assume that all loans are granted for a one-year period, starting at the beginning of the accounting year. The entire principal and all interest mature at year-end. Following the literature, we let the number of competing banks correspond to the number of models to be compared. Consequently, we assume that there are eight banks in the market, each using one of the eight variable sets, respectively, in a discrete hazard model with LR and training sets as defined before. The number of variables in the variable sets chosen by feature selection methods are predetermined to nine, as in Section 4.2. Each bank uses its model for predicting the probabilities of bankruptcy associated with each potential borrower. These probabilities are then used for profit maximization by optimizing credit decision making and credit risk pricing. In other words, the predictions of the eight models indirectly affect the profitability of the eight banks, which eventually serves as performance measure for assessing the quality of the eight variable sets.

Our simulation study expands the existing literature in several ways. First, while Agarwal and Taffler (2007, 2008) and Bauer and Agarwal (2014) assume a simple loan market worth USD/GBP 100 billion and that all companies want to borrow an equal amount of money, we use the effective size of the Norwegian loan market covered by our sample and derive the amount each company wants to borrow from their financial statements. Hence, our loan market setting is much more realistic as it directly takes reference to the effective Norwegian SME loan market. Second, we assume that banks acquire the money for lending at the risk-free rate and thus must write off a loss equivalent to the risk-free rate multiplied with the principal in the case of borrower bankruptcy.[23]

More formally, our competitive credit market is structured as follows: Let $\mathbf{r} = \{r_n\}_{n=1,...,N}$ be the vector of credit spreads assigned by one of the banks with n denoting the potential borrowers (financial statements). We assume that all banks set the credit spreads such that the expected revenues cover all expected losses plus profit for each loan:

$$\underbrace{(1-\hat{y}_n)\times(r_n+c)}_{\text{conditional expected revenue}} = \underbrace{\hat{y}_n\times(LGD+r_f)}_{\text{conditional expected loss}} + \underbrace{(1-\hat{y}_n)\times k}_{\text{conditional expected profit}} \tag{4}$$

where $\hat{y}_n$ is the predicted out-of-sample probability for bankruptcy derived from the bank's specific model, LGD is the loss-given default in the case of borrower bankruptcy (including recovery costs), r_f is the risk-free rate, k is the credit spread for a loan with no credit risk, i.e., the required profit margin, and c is the cost of granting loans for strategic reasons.[24] Similar to Blöchlinger and Leippold (2006), Agarwal and Taffler (2008), and Bauer and Agarwal (2014), we assume $c = 0$. Consequently, Equation (4) can be rewritten to the following equation for deriving the credit spreads assigned to each potential borrower:

$$r_n = \frac{\hat{y}_n}{1-\hat{y}_n}(LGD+r_f)+k \tag{5}$$

We follow the criteria of rejection based on the credit spread values as in Blöchlinger and Leippold (2006), and assume that banks reject potential borrowers when $r_n > 10\%$. Further, we assume $LGD = 40\%$, which is in line with the Basel III regulatory framework (Bank of International Settlements, 2017). In addition, we assume $k = 0.30\%$, following Blöchlinger and Leippold (2006), Agarwal and Taffler (2008), and Bauer and Agarwal (2014).

If the potential borrowers are granted loans from several banks, we assume they choose to borrow the whole amount from the bank offering the lowest credit spread. Further, we follow the literature and set no limits on the total amount any bank can lend. In addition, if borrowers go bankrupt, we assume that the lending bank receives no interest payment and suffers a loss

[23] For each accounting year, we use the annual effective yields of synthetic Norwegian Treasury bills for maturities of 12 months as risk free rate. The data are retrieved from: www.norges-bank.no/en/topics/Statistics/Interest-rates/Treasury-bills-annual/. The values are 2.25%, 2.12%, 1.53%, 1.52%, 1.29%, 0.73%, 0.50%, and 0.42% for 2010-2017, respectively.

[24] If the bank expects profits from a customer relationship through businesses other than lending, it may be willing to lower credit spreads in order to retain or acquire the customer. The cost for such lowering of credit spreads are captured by c.

equal to the principal weighted by LGD. In sum, the total profit of any bank for any given accounting year is given by

$$\underbrace{\sum_{n=1}^{N} p_n \times r_n \times (1 - y_n)}_{\text{total revenue}} - \underbrace{\sum_{n=1}^{N} p_n \times (LGD + r_f) \times y_n}_{\text{total loss}} \tag{6}$$

where p_n is the principal, i.e., the interest bearing debt derived from the financial statement, and y_n is the actual classification of bankrupt (1) or non-bankrupt (0). The first term of Equation (6) represents the bank's total revenue. The bank's revenue increases when the bank lends to more non-bankrupting borrowers as a result of using a variable set that yields lower credit spreads for these non-bankrupting customers compared to competing banks. Hence, avoiding the second type of error, as explained above, results in higher revenues. The second term of Equation (6) measures the bank's total loss. The bank's loss declines when a bank uses a variable set that rejects loans to bankrupting potential borrowers by providing more accurate predictions about their bankruptcies, i.e., by avoiding the first type of error explained above. In summary, Equation (6) measures a bank's total profit that results from a given variable set when taking into account realistic costs brought about by the first and second type of prediction errors.

The outcomes from the simulated credit market for all eight banks are reported in Table 3. The table shows for each bank its market share (in EUR), the share of bankrupting borrowers in its loan portfolio, as well as the share of all bankrupting borrowers which the bank has lent to.[25] Further, the table displays the revenues, losses, and profits, as derived from Equation (6) as well as profit margins for each bank, in EUR millions. Finally, following Agarwal and Taffler (2008) and Bauer and Agarwal (2014), we report return on assets (ROA), which is profit divided by total amount lent, and return on risk-weighted assets (RORWA), where risk-weighted assets are calculated in accordance with the Basel III regulatory framework.

[Table 3 about here.]

We find that using the variable set chosen by the LASSO method (LAS) yields the highest profit, profit margin, ROA, and RORWA. Further, it results in the lowest share of bankrupt borrowers in the loan portfolio and the lowest share of bankrupting borrowers which the bank has lent to. The profit is 49% and 84% higher when using LAS compared to using the variable

[25]The yearly average total market size for the accounting years 2010-2017 is EUR 2,351 billion. The number of bankrupting and non-bankrupting borrowers, i.e., financial statements are detailed in Figure 1. The total market share as well as the share of bankrupting borrowers in the market do not add up to 100% across banks, as several potential borrowers are rejected by all banks.

sets yielding the second (TAF) and third (WRA) highest profits, respectively. Note that LAS is substantially more profitable even though TAF results in a market share that is higher by 5.8 percentage points. Further, using LAS results in a decrease in the share of bankrupt borrowers in the loan portfolio by 74% and 25%, as well as a decrease in the share of loans to bankrupting borrowers among those in the market by 81% and 21%, versus using TAF and WRA, respectively. We observe that using the variable sets chosen by the filter feature selection method (FIL) results in relatively poor performance being dominated by variable sets based on prior literature.

The superiority of using LAS in terms of bank profitability is robust across all combinations of $LGD \in \{30\%, 40\%, \ldots, 70\%\}$, $k \in \{0.1\%, 0.2\%, \ldots, 0.5\%\}$, and when letting the amount each company wants to borrow be given by either interest bearing debt or total debt. Further, in all cases the superiority of using LAS is robust to the following settings: *i)* simulating a competitive credit market with eight banks, as above, where each bank uses one of the eight variable sets *ii)* simulating a competitive credit market with three banks, each using one of the input variable sets given by the three feature selection methods (LAS, WRA, and FIL), *iii)* simulating a competitive credit market with six banks, where one uses LAS and the other five use the benchmark sets (ALT, ASA, SEB, ZMI, and TAF), and *iv)* simulating seven competitive credit markets, each with two banks where one uses LAS and the other uses one of the other seven sets, respectively.[26]

In Figure 8, the borrowers are grouped into percentiles based on the amount they are willing to borrow, sorted in descending order. Panel A presents, for each percentile, the amount each bank lends to non-bankrupting borrowers up to that percentile as a proportion of the total amount lent to all non-bankrupting borrowers. Panel B shows the same for bankrupting borrowers. Results in Panel A show that banks using ASA, TAF, and LAS produce the highest proportions of total amount lent to non-bankrupting borrowers (i.e., good customers) at 30%, 27%, and 21%, respectively. This ability to select good customers results in the highest revenues for these banks, a finding that is consistent with results reported in Table 3. However, Panel B shows that banks using ASA and TAF account for 28% and 37% of the amount lent to bankrupting borrowers, respectively, which generates large losses for these banks. In contrast, the bank using LAS only makes up for 4% of the total amount lent to bankrupting borrowers. In combination, the bank using LAS generates the highest profits. The bank using WRA performs better than the banks using ASA and TAF in terms of profit margin, ROA, and RORWA.

[26]In total, our finding is thus robust across 500 different simulated markets. Results of all simulations are available upon request.

Still, it performs worse that the bank using LAS, as it lends a relatively lower amount to non-bankrupting borrowers and a relatively higher amount to bankrupting borrowers. Hence, LAS is clearly superior to all other variable sets.

[Figure 8 about here.]

In summary, even small differences in modeling performance accordingly to commonly used metrics can yield large differences in bank profitability. For example, in our simulation study, we show that using LAS compared to WRA yields a much higher bank profit (see Table 3), even though the difference in the AUC values is small (see Figure 4). The dominance of LAS over WRA stems from a different assessment of credit spreads, which ultimately affects the proportion of types of borrowers in the loan portfolio (see Figure 8). In particular, the bank employing LAS offers overall lower credit spreads to non-bankrupting borrowers compared to the bank employing WRA, at the same time keeping credit spreads to bankrupting borrowers higher. Consequently, compared to the bank using WRA, the bank using LAS achieves higher revenues because it attracts non-bankrupting borrowers, without suffering higher losses from entering into more unfavorable loan agreements with bankrupting borrowers.

6. Selecting variable sets at the industry-level

Feature selection methods do not only allow for a dynamic estimation and reconfiguration of bankruptcy prediction models, but can also be estimated within sub-samples of the company universe in a straightforward manner. This may turn out useful when bankruptcy predictors are expected to differ in the cross-section along measurable dimensions. One obvious candidate for such a cross-sectional sample split is industry affiliation. Indeed, Chava and Jarrow (2004) find that industry groupings affect coefficient values, in-sample fit, and out-of-sample predictions of discrete hazard models for bankruptcy prediction.

In this section, we evaluate the choice of bankruptcy predictors at the industry level. We do this by selecting industry-specific variable sets for all industries that include at least 500 financial statements that are categorized as bankrupt within the entire sample period. To keep the analysis traceable, we rely on our main setting, that is, we use LASSO in combination with discrete hazard models estimated with LR, and restrict the selection of bankruptcy predictors to nine variables for each of the years 2010-2017.

Table IA.1, Panels A to H, in the Internet Appendix shows the bankruptcy predictors chosen for each industry and year. Each of the eight panels summarizes results for a specific industry.[27] Results show that the variables chosen in the full sample, as reported in Table 2, are quite persistent across the 64 permutations, resulting from eight yearly analyses across eight different industries: In nine permutations, the variable set selected at the industry-level exactly matches that selected in the full sample. In 26 permutations, one of nine bankruptcy predictors is replaced. In 21 permutations, two predictors are replaced. In only eight permutations, three or four predictors are replaced.

The key question is of course, whether selecting bankruptcy predictors at the industry-level improves model performance in way that matters to bank profitability. To test this empirically, we run a simulation exercise with two banks similar to that in Section 5. The first bank uses LAS, which is the most profitable approach from the simulation described in Section 5, with results reported in Table 3. That is, the bank selects variables with the LASSO method applied to the full sample. The second bank uses the same bankruptcy prediction model, but selects variables for each industry separately and applies them to potential borrowers according to their industry affiliation (referred to as LIN). For borrowers in industries that do not qualify for a separate bankruptcy prediction model estimation due an insignificant number of bankrupt observations as explained above, we use the model configured to the full sample (i.e., LAS). Results obtained from simulating the credit market with these two banks are reported in Table 4. The table shows that LIN is associated with a slightly smaller market share of 49.1% versus 50.8%.[28] However, LIN faces lower losses, generates higher profits, including a profit margin that is higher by 16%, 33% higher ROA, and 20% higher RORWA.

[Table 4 about here.]

We conduct a number of robustness tests for these results. For instance, we find that the superiority of LIN in terms of bank profitability is robust to defining the amount to be borrowed by each company either in terms of interest-bearing debt or total debt. Specifically, in both cases, LIN is superior across all combinations of $LGD \in \{30\%, 40\%, \ldots, 70\%\}$ and $k \in \{0.1\%, 0.2\%, \ldots, 0.5\%\}$.

[27] We select industry-specific bankruptcy predictors for the industries *'Manufacturing'*, *'Construction'*, *'Wholesale and retail trade; repair of motor vehicles and motorcycles'*, *'Transportation and storage'*, *'Accommodation and food service activities'*, *'Information and communication'*, *'Professional, scientific and technical activities'*, and *'Administrative and support service activities'*. We omit the following industries from this analysis because of a small number of financial statements that are categorized as bankrupt: *'Agriculture, forestry and fishing'*, *'Mining and quarrying'*, *'Education'*, *'Human health and social work activities'*, *'Arts, entertainment and recreation'*, and *'Other service activities'*. Note that there are no firms in industries *'Activities of household as employers'* or *'Activities of extraterritorial organisations and bodies'*, which are part of the Norwegian SIC2007, in our sample.

[28] As in Table 3, neither the total market share nor the share of bankrupting borrowers in the market add up to 100% as several potential borrowers are rejected by both banks.

Similar to Figure 8 above, in Figure 9, we group borrowers into percentiles based on the amount they are willing to borrow, sorted in descending order. Panels A and B of Figure 9 show the cumulative amount lent to borrowers that did not bankrupt and borrowers that bankrupted, respectively. The panels show that the bank using industry-specific variable sets, i.e., LIN, is superior to the one that uses LAS as it extends loans to much fewer bankrupting borrowers while not giving much fewer loans to non-bankrupting borrowers.

[Figure 9 about here.]

In summary, the results in this section show that the flexibility of our feature selection-based bankruptcy prediction models allows to not only reconfigure models over time, but also to estimate models across sub-samples, such as for instance across industries, to further improve model performance. Results show that while our model is fairly stable across both time and industries, estimating our model at the industry-level further improves model performance and eventually the profitability of banks applying these models.

7. Conclusions

We show that feature selection methods improve the performance of bankruptcy prediction models in a comprehensive dataset of all privately held Norwegian SMEs. Our paper contributes to the existing literature on bankruptcy prediction by comparing three different feature selection methods, in combination with state-of-the art statistical and machine learning estimation techniques for bankruptcy prediction, also versus popular benchmark variable sets (Altman, 1968; Taffler, 1984; Zmijewski, 1984; Altman and Sabato, 2007; Bernhardsen and Larsen, 2007). We find that an embedded LASSO feature selection method performs best in terms of in-sample fit, out-of-sample performance, and stability. Our analysis is complemented by a simulation study, mimicking a real-world competitive credit market, which sheds light on bank profitability under different bankruptcy prediction model choices. This latter analysis reinforces the superiority of LASSO by showing that bankruptcy predictors selected by LASSO yield a substantially higher bank profitability compared to using other feature selection methods or benchmark sets from previous literature. Finally, we show in a second simulation that further model improvements can be achieved when adapting bankruptcy predictors to sub-sets of the company universe.

References

Adämmer, P., Schüssler, R.A., 2020. Forecasting the equity premium: Mind the news! Review of Finance 24, 1313–1355.

Agarwal, V., Taffler, R.J., 2007. Twenty-five years of the Taffler z-score model: Does it really have predictive ability? Accounting and Business Research 37, 285–300.

Agarwal, V., Taffler, R.J., 2008. Comparing the performance of market-based and accounting-based bankruptcy prediction models. Journal of Banking & Finance 32, 1541–1551.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723.

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

Altman, E.I., 1993. Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting & Avoiding Distress and Profiting from Bankruptcy. Wiley Frontiers in Finance. 2nd ed., Wiley, New York.

Altman, E.I., 2018. A fifty-year retrospective on credit risk models, the Altman Z -score family of models and their applications to financial markets and managerial strategies. The Journal of Credit Risk 14, 1–34.

Altman, E.I., Haldeman, R.G., Narayanan, P., 1977. ZETA analysis, a new model to identify bankruptcy risk of corporations. Journal of Banking & Finance 1, 29–54.

Altman, E.I., Hotchkiss, E., Wang, W., 2019. Corporate Financial Distress, Restructuring, and Bankruptcy: Analyze Leveraged Finance, Distressed Debt, and Bankruptcy. 4th ed. ed., John Wiley & Sons.

Altman, E.I., Sabato, G., 2007. Modelling credit risk for SMEs: Evidence from the U.S. market. Abacus 43, 332–357.

Azimi, M., Agrawal, A., 2021. Is Positive Sentiment in Corporate Annual Reports Informative? Evidence from Deep Learning. The Review of Asset Pricing Studies 11, 762–805.

Balcaen, S., Ooghe, H., 2006. 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. The British Accounting Review 38, 63–93.

Bank of International Settlements, 2017. Basel III: Finalising post-crisis reforms. Report.

Bauer, J., Agarwal, V., 2014. Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. Journal of Banking & Finance 40, 432–442.

Beaver, W.H., 1966. Financial ratios as predictors of failure. Journal of Accounting Research 4, 71–111.

Beck, T., Demirguc-Kunt, A., 2006. Small and medium-size enterprises: Access to finance as a growth constraint. Journal of Banking & Finance 30, 2931 – 2943.

Beck, T., Demirgüç-Kunt, A., Maksimovic, V., 2008. Financing patterns around the world: Are small firms different? Journal of Financial Economics 89, 467 – 487.

Begley, J., Ming, J., Watts, S., 1996. Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models. Review of Accounting Studies 1, 267–284.

Bernhardsen, E., Larsen, K., 2007. Modelling credit risk in the enterprise sector – further development of the SEBRA model. Economic Bulletin 2/2007.

Bharath, S.T., Shumway, T., 2008. Forecasting Default with the Merton Distance to Default Model. The Review of Financial Studies 21, 1339–1369.

Bianchi, D., Büchner, M., Tamoni, A., 2021. Bond Risk Premiums with Machine Learning. Review of Financial Studies 34, 1046–1089.

Binz, O., Schipper, K., Standridge, K., 2022. What can analysts learn from artificial intelligence about fundamental analysis? Working Paper, INSEAD and Duke University .

Blöchlinger, A., Leippold, M., 2006. Economic benefit of powerful credit scoring. Journal of Banking & Finance 30, 851–873.

Blöchlinger, A., Leippold, M., 2018. Are ratings the worst form of credit assessment except for all the others? Journal of Financial and Quantitative Analysis 53, 299–334.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthey Weather Review 78, 1–3.

Bryzgalova, S., Pelger, M., Zhu, J., 2021. Forest Through the Trees: Building Cross-Sections of Stock Returns. Working Paper, London Business School and Stanford University .

Calomiris, C.W., Mamaysky, H., 2019. How news and its context drive risk and returns around the world. Journal of Financial Economics 133, 299–336.

Campbell, J., Hilscher, J., Szilagyi, J., 2008. In search of distress risk. Journal of Finance 63, 2899–2939.

Campello, M., Gao, J., Qiu, J., Zhang, Y., 2017. Bankruptcy and the cost of organized labor: Evidence from union elections. Review of Financial Studies 31, 980–1013.

Cao, K., You, H., 2021. Fundamental analysis via machine learning. Working Paper, Hong Kong University of Science & Technology .

Carbó-Valverde, S., Rodríguez-Fernández, F., Udell, G.F., 2009. Bank market power and SME financing constraints. Review of Finance 13, 309–340.

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering 40, 16–28.

Chang, X., Chen, Y., Wang, S.Q., Zhang, K., Zhang, W., 2019. Credit default swaps and corporate innovation. Journal of Financial Economics 134, 474–500.

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

Chen, L., Pelger, M., Zhu, J., 2021. Deep learning in asset pricing. Working Paper, Stanford University .

Chen, X., Cho, Y.H.T., Dou, Y., Lev, B., 2022. Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data. Journal of Accounting Research 60, 467–515.

Chen, Z., Harford, J., Kamara, A., 2019. Operating leverage, profitability, and capital structure. Journal of Financial and Quantitative Analysis 54, 369–392.

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

Cong, L.W., Tang, K., Wang, J., Zhang, Y., 2020. AlphaPortfolio: Direct construction through reinforcement learning and Interpretable AI. Working Paper, Cornell University .

De Bock, K.W., Coussement, K., Lessmann, S., 2020. Cost-sensitive business failure prediction when misclassification costs are uncertain: A heterogeneous ensemble selection approach. European Journal of Operational Research 285, 612–630.

Deakin, E.B., 1976. Distributions of financial accounting ratios: Some empirical evidence. The Accounting Review 51, 90–96.

DeMiguel, V., Martín-Utrera, A., Nogales, F.J., Uppal, R., 2020. A transaction-cost perspective on the multitude of firm characteristics. Review of Financial Studies 33, 2180–2222.

Dietsch, M., Petey, J., 2004. Should SME exposures be treated as retail or corporate exposures? A comparative analysis of default probabilities and asset correlations in French and German SMEs. Journal of Banking & Finance 28, 773 – 788.

Eisenbeis, R.A., 1977. Pitfalls in the application of discriminant analysis in business, finance, and economics. Journal of Finance 32, 875–900.

Eklund, T., Larsen, K., Bernhardsen, E., 2001. Model for analysing credit risk in the enterprise sector. Economic Bulletin 3/2001.

Feng, G., Giglio, S., Xiu, D., 2020. Taming the factor zoo: A test of new factors. Journal of Finance 75, 1327–1370.

Fisman, R., Love, I., 2003. Trade credit, financial intermediary development, and industry growth. Journal of Finance 58, 353–374.

Fitzgerald, P., 2009. New bankruptcy-risk model looks to topple Altman's Z-score. The Wall Street Journal .

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Review of Financial Studies 33, 2326–2377.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22.

Frydman, H., Altman, E.I., Kao, D.L., 1985. Introducing recursive partitioning for financial classification: The case of financial distress. Journal of Finance 40, 269–291.

Garcia, D., Hu, X., Rohrer, M., 2021. The colour of finance words. Working paper, University of Colorado at Boulder and Norwegian School of Economics .

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Review of Financial Studies 33, 2223–2273.

Hagan, M.T., Menhaj, M.B., 1994. Training feedforward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks 5, 989–993.

Han, Y., Kong, L., 2020. The serial dependence of the commodity futures returns: A machine learning approach. Working Paper, University of North Carolina at Charlotte .

Hautsch, N., Schaumburg, J., Schienle, M., 2014. Financial network systemic risk contributions. Review of Finance 19, 685–738.

Hillegeist, S., Keating, E., Cram, D., Lundstedt, K., 2004. Assessing the probability of bankruptcy. Review of Accounting Studies 9, 5–34.

Hopwood, W., McKeown, J., Mutchler, J., 1989. A test of the incremental explanatory power of opinions qualified for consistency and uncertainty. The Accounting Review 64, 28–48.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression. John Wiley & Sons, Incorporated, New York, United States.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Härdle, W., Lee, Y.J., Schäfer, D., Yeh, Y.R., 2009. Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies. Journal of Forecasting 28, 512–534.

John, G.H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. Machine Learning Proceedings 1994 , 121–129.

Jones, S., Johnstone, D., Wilson, R., 2015. An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. Journal of Banking & Finance 56, 72–85.

Jones, S., Johnstone, D., Wilson, R., 2017. Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. Journal of Business Finance & Accounting 44, 3–34.

Joy, O.M., Tollefson, J.O., 1975. On the financial applications of discriminant analysis. Journal of Financial and Quantitative Analysis 10, 723–739.

Kaastra, I., Boyd, M., 1996. Designing a neural network for forecasting financial and economic time series. Neurocomputing 10, 215–236.

Keles, D., Scelle, J., Paraschiv, F., Fichtner, W., 2016. Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks. Applied Energy 162, 218–230.

Kelly, B.T., Manela, A., Moreira, A., 2021. Text selection. Journal of Business & Economic Statistics 39, 859–879.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324.

Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. Journal of Financial Economics 135, 271–292.

Kumar, P.R., Ravi, V., 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. European Journal of Operational Research 180, 1–28.

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

Light, N., Maslov, D., Rytchkov, O., 2017. Aggregation of information about the cross section of stock returns: A latent variable approach. Review of Financial Studies 30, 1339–1381.

Mansi, S.A., Maxwell, W.F., Zhang, A., 2012. Bankruptcy prediction models and the cost of debt. Journal of Fixed Income 21, 25.

Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial Applied Mathematics 11, 431–441.

Martin, D., 1977. Early warning of bank failure: A logit regression approach. Journal of Banking & Finance 1, 249–276.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behaviour. P. Zarembka, Ed., Frontiers in Econometrics 1, 105–142.

Meier, J.M., Servaes, H., 2019. The bright side of fire sales. Review of Financial Studies 32, 4228–4270.

Merton, R.C., 1974. On the pricing of corporate debt: The risk structure of interest rates. Journal of Finance 29, 449–470.

Nagel, S., Purnanandam, A., 2019. Banks' risk dynamics and distance to default. Review of Financial Studies 33, 2421–2467.

Nakariyakul, S., Casasent, D.P., 2009. An improvement on floating search algorithms for feature subset selection. Pattern Recognition 42, 1932–1940.

Ogneva, M., Piotroski, J.D., Zakolyukina, A.A., 2019. Accounting fundamentals and systematic risk: Corporate failure over the business cycle. The Accounting Review 95, 321–350.

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

Reisz, A.S., Perlich, C., 2007. A market-based framework for bankruptcy prediction. Journal of Financial Stability 3, 85–131.

Robb, A.M., Robinson, D.T., 2012. Capital structure decisions of new firms. Review of Financial Studies 27, 153–179.

Rossi, A.G., 2018. Predicting stock market returns with machine learning. Working Paper, Georgetown University .

Sadhwani, A., Giesecke, K., Sirignano, J., 2021. Deep Learning for Mortgage Risk. Journal of Financial Econometrics 19, 313–368.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

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

Stein, R.M., 2005. The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. Journal of Banking & Finance 29, 1213–1236.

Taffler, R.J., 1984. Empirical models for the monitoring of UK corporations. Journal of Banking & Finance 8, 199–227.

Tam, K.Y., Kiang, M.Y., 1992. Managerial applications of neural networks: The case of bank failure predictions. Management Science 38, 926–947.

Tian, S., Yu, Y., 2017. Financial ratios and bankruptcy predictions: An international evidence. International Review of Economics & Finance 51, 510–526.

Tian, S., Yu, Y., Guo, H., 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.

Toplensky, R., 2020. Banks are only as sound as their models. The Wall Street Journal .

Weigand, A., 2019. Machine learning in empirical asset pricing. Financial Markets and Portfolio Management 33, 93–104.

Zhang, A.L., 2021. Uncovering mutual fund private information with machine learning. Working Paper, Georgia State University .

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

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

Table 1: Overview of the input variable sets

Method	Input variable sets
WRA	Chosen by the wrapper feature selection method
LAS	Chosen by the embedded LASSO feature selection method
FIL	Chosen by the filter feature selection method
RAN	Selected randomly
ALT	Employed in Altman (1968)
TAF	Employed in Taffler (1984)
ZMI	Employed in Zmijewski (1984)
ASA	Employed in Altman and Sabato (2007)
SEB	Employed in the SEBRA model (Bernhardsen and Larsen, 2007)

The feature selection methods are detailed in Section 3.3. Appendix C lists the input variables of the benchmark sets (ALT, TAF, ZMI, ASA, and SEB).

Table 2: Estimation results of the set of nine input variables selected by the LASSO method

Panel A: For the years 2010-2013

	Category	**2010**	**2011**	**2012**	**2013**
accounts payable / total assets	Liquidity	1.64(26.47)	1.60(26.20)	1.62(25.82)	1.61(25.41)
dummy; one if total liability exceeds total assets	Leverage	0.45(11.43)	0.38(9.58)	0.36(8.84)	0.40(9.35)
(current liabilities - short-term liquidity) / total assets	Leverage	0.16(3.96)	0.22(5.86)	0.23(6.01)	0.26(6.72)
net income / total assets	Profitability	-0.72(-15.78)	-0.65(-14.76)	-0.63(-13.95)	-0.70(-14.51)
public taxes payable / total assets	Liquidity	3.42(27.83)	3.37(28.23)	3.45(28.38)	3.51(29.06)
interest expenses / total assets	Solvency	8.49(18.20)	9.28(20.54)	9.36(20.10)	10.15(20.41)
dummy; one if paid-in equity is less than total equity	Solvency	-0.86(-21.35)	-0.84(-20.89)	-0.82(-19.47)	-0.72(-16.66)
log(age in years)	Age	-0.50(-28.84)	-0.52(-30.19)	-0.51(-28.50)	-0.53(-29.19)
inventory / current assets	Liquidity	0.77(17.42)	0.81(18.71)	0.80(17.65)	0.85(18.71)
short-term liquidity / current assets	Liquidity				
Intercept		-3.83(-89.97)	-3.81(-88.69)	-3.94(-86.60)	-4.06(-86.28)
In-sample AUC		0.85	0.86	0.86	0.86
In-sample Brier score		0.02	0.02	0.02	0.02
AIC		48360	48937	46080	45218
R^2		0.20	0.20	0.20	0.21
Out-of-sample AUC		0.86	0.86	0.87	0.86
Out-of-sample Brier score		0.02	0.02	0.02	0.01
Out-of-sample decile rankings:					
Decile 1		0.58	0.60	0.59	0.60
Decile 2		0.18	0.17	0.16	0.16
Decile 3		0.10	0.10	0.09	0.10
Decile 4		0.05	0.03	0.06	0.05
Decile 5		0.02	0.03	0.04	0.04
Decile 6-10		0.06	0.07	0.06	0.06

Estimation results of the set of nine input variables selected by the LASSO method for the years 2010-2013. We use LR as estimation technique, as well as a rolling window and a forward validation procedure as shown in Figure 1. Estimation results for 2014-2017 are shown in Panel B. We show coefficient estimates and z-scores in parentheses. We report in-sample fit by AUC, Brier score, AIC, and R^2. We report out-of-sample prediction performance by AUC, Brier score, and decile rankings.

Table 2: Estimation results of the set of nine input variables selected by the LASSO method

Panel B: For the years 2014-2017

	Category	2014	2015	2016	2017
accounts payable / total assets	Liquidity	1.64(25.76)	1.62(25.76)	1.51(23.92)	1.75(27.20)
dummy; one if total liability exceeds total assets	Leverage	0.39(8.94)	0.37(8.20)	0.37(8.02)	0.42(9.10)
(current liabilities - short-term liquidity) / total assets	Leverage	0.30(7.80)	0.38(9.76)	0.45(11.69)	0.10(2.40)
net income / total assets	Profitability	-0.67(-13.71)	-0.64(-12.78)	-0.58(-11.60)	-0.62(-12.53)
public taxes payable / total assets	Liquidity	3.63(30.86)	3.92(34.55)	3.95(35.79)	4.22(38.68)
interest expenses / total assets	Solvency	10.27(19.50)	9.26(16.99)	9.08(16.02)	8.81(14.93)
dummy; one if paid-in equity is less than total equity	Solvency	-0.66(-15.01)	-0.59(-13.46)	-0.50(-11.14)	-0.45(-10.10)
log(age in years)	Age	-0.54(-30.24)	-0.56(-31.52)	-0.56(-31.79)	-0.60(-33.41)
inventory / current assets	Liquidity	0.79(16.98)	0.84(18.03)	0.87(18.31)	
short-term liquidity / current assets	Liquidity				-1.44(-22.35)
Intercept		-4.11(-86.06)	-4.16(-86.49)	-4.22(-85.73)	-3.48(-66.77)
In-sample AUC		0.87	0.87	0.87	0.87
In-sample Brier score		0.02	0.01	0.01	0.01
AIC		44676	45104	44773	44988
R^2		0.21	0.21	0.21	0.25
Out-of-sample AUC		0.86	0.87	0.87	0.87
Out-of-sample Brier score		0.01	0.01	0.01	0.01
Out-of-sample decile rankings:					
Decile 1		0.59	0.60	0.59	0.60
Decile 2		0.17	0.18	0.17	0.17
Decile 3		0.09	0.08	0.10	0.08
Decile 4		0.06	0.05	0.05	0.06
Decile 5		0.03	0.03	0.03	0.03
Decile 6-10		0.06	0.06	0.06	0.06

Estimation results of the set of nine input variables selected by the LASSO method for the years 2014-2017. We use LR as estimation technique, as well as a rolling window and a forward validation procedure as shown in Figure 1. Estimation results for 2010-2013 are shown in Panel A. We show coefficient estimates and z-scores in parentheses. We report in-sample fit by AUC, Brier score, AIC, and R^2. We report out-of-sample prediction performance by AUC, Brier score, and decile rankings.

Table 3: Simulation of a competitive credit market with eight banks

	WRA	LAS	FIL	ALT	TAF	ZMI	ASA	SEB
Market share of each bank in EUR (%)	15.1	21.1	1.6	1.3	26.9	0.3	30.2	3.4
Share of bankrupting borrowers in portfolio (%)	1.6	1.2	6.0	9.9	4.7	22.3	1.7	11.2
Share of bankrupting borrowers in the market (%)	4.2	3.3	3.9	5.8	17.5	10.9	38.8	14.7
Revenue	11,804.2	16,958.7	1,693.5	1,605.7	26,502.0	496.6	18,467.1	3,857.5
Loss	3,632.1	1,950.1	838.1	1,955.5	16,414.3	670.4	12,267.0	6,181.2
Profit	8,172.2	15,008.6	855.4	-349.7	10,087.7	-173.8	6,200.1	-2,323.7
Profit margin (%)	69.2	88.5	50.5	-21.8	38.1	-35.0	33.6	-60.2
Return on assets (ROA) (%)	0.3	0.4	0.3	-0.1	0.2	-0.3	0.1	-0.4
Return on risk-weighted assets (RORWA) (%)	0.6	0.8	0.4	-0.2	0.3	-0.3	0.4	-0.7

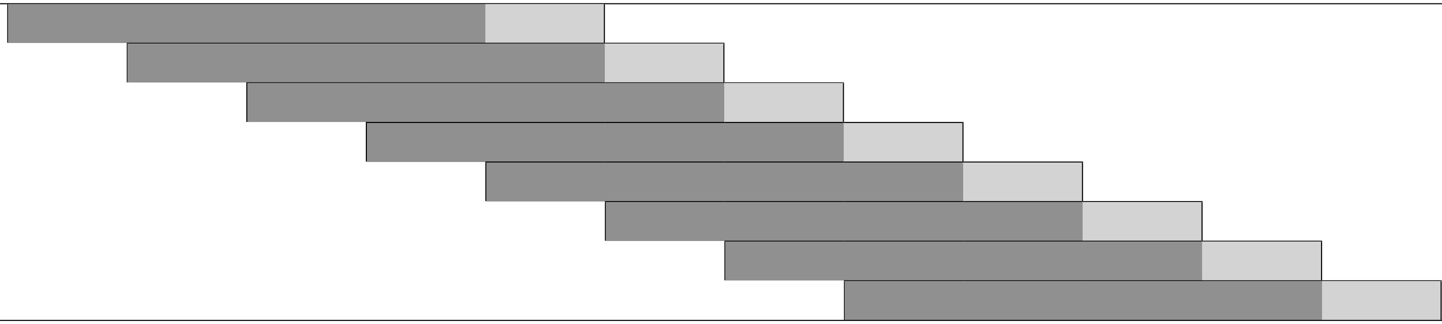
Each bank in the simulation uses one of the variable sets WRA, LAS, FIL, ALT, TAF, ZMI, ASA and SEB, as shown in Table 1. The first three are the sets of nine input variables selected by the wrapper, LASSO, and filter methods, respectively. The latter five are the sets employed in Altman (1968), in Taffler (1984), in Zmijewski (1984), in Altman and Sabato (2007), and in the SEBRA model (Bernhardsen and Larsen, 2007), respectively. The table shows the market share for each bank, as well as the share of borrowers that went bankrupt in each bank's loan portfolio and the share of all borrowers in the market that went bankrupt who borrow from each bank. Further, it reports revenues, losses, and profits of each bank in EUR millions, as given in Equation (6). Finally, the panel presents profit margins, return on assets (ROA), which is profit divided by total amount lent, and return on risk-weighted assets (RORWA), calculated by following the Basel III regulatory framework. We use in our simulation all financial statements in our sample from the accounting years 2010-2017 (see Figure 1).

Table 4: Simulation of a competitive credit market with two banks – general versus industry-specific datasets

	LAS	LIN
Market share of each bank in EUR (%)	50.8	49.1
Share of bankrupting borrowers in portfolio (%)	2.7	1.9
Share of bankrupting borrowers in the market (%)	57.1	34.6
Revenue	55,854.8	50,954.0
Loss	24,169.3	17,302.1
Profit	31,685.5	33,652.0
Profit margin (%)	56.7	66
Return on assets (ROA) (%)	0.3	0.4
Return on risk-weighted assets (RORWA) (%)	0.5	0.6

The first bank in the simulation uses the variable set chosen by the LASSO method when using the full sample (LAS). The second bank, LIN, uses for each potential borrower the variable set chosen by LASSO that corresponds to the potential borrower's industry (see Table IA.1, Panels A to H, in the Internet Appendix) or the variable set chosen by the overall model, as in LAS, for industries that are too small to select a separate variable set. The table shows the market share for each bank, as well as the share of borrowers that went bankrupt in each bank's loan portfolio and the share of all borrowers in the market that went bankrupt who borrow from each bank. Further, it reports revenues, losses, and profits of each bank in EUR millions, as given in Equation (6). Finally, the table presents profit margins, return on assets (ROA), which is profit divided by total amount lent, and return on risk-weighted assets (RORWA), calculated by following the Basel III regulatory framework. We use in our simulation all financial statements in our sample from the accounting years 2010-2017 (see Figure 1).

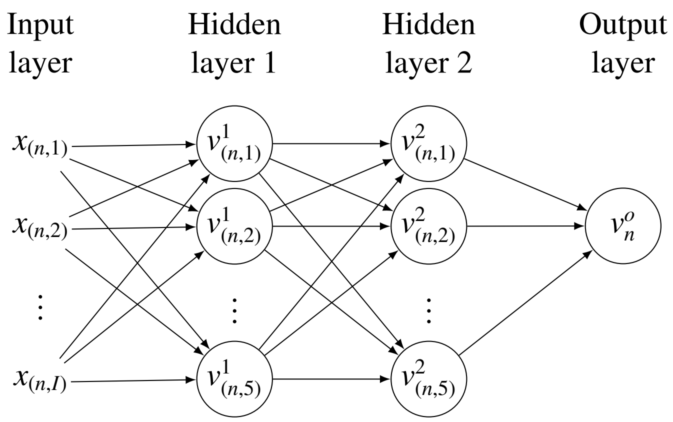
Figure 1: The procedure for selecting input variables, training models, and evaluating



	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
Obs. (Bankrupt)	1,230	1,847	1,580	1,503	1,313	1,380	1,448	1,392	1,349	1,276	1,487	1,485	17,290
Obs. (Non-bankrupt)	71,771	74,999	72,027	76,825	79,160	80,370	85,102	85,627	86,216	87,864	92,066	93,008	985,035
Bankruptcy frequency	1.68%	2.4%	2.15%	1.92%	1.63%	1.69%	1.67%	1.6%	1.54%	1.43%	1.59%	1.57%	1.72%

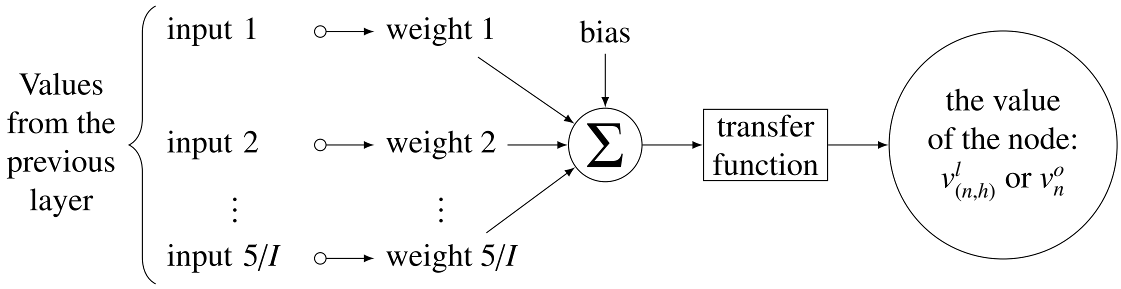
We predict one-year-ahead bankruptcy and use a rolling window and a forward validation procedure. We select the input variables, train the models, and evaluate their in-sample fit using all financial statements over four consecutive accounting years (dark gray). Out-of-sample performance is evaluated by using all financial statements in the subsequent accounting year (light gray). We use a dataset of financial statements of Norwegian SMEs in the time period 2006-2017, categorized as bankrupt or non-bankrupt. The number of bankrupt and non-bankrupt financial statements, as well as bankruptcy frequency, per accounting year and total across all years is shown at the bottom of the figure.

Figure 2: Illustration of the DNN for any financial statement n



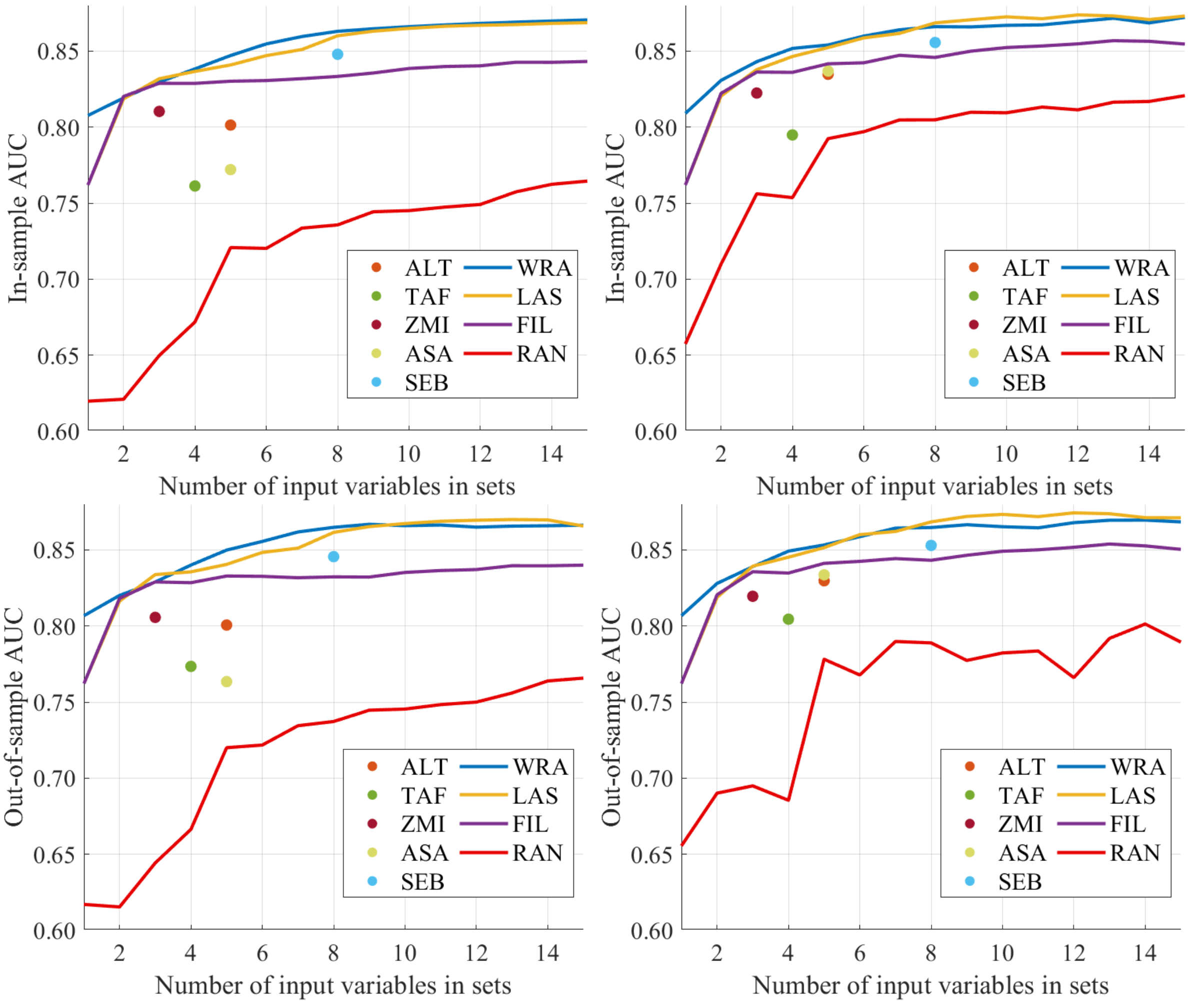
The circles represent nodes, where $v^l_{(n,h)}$ is the value of the nodes $h \in \{1, \ldots, 5\}$ for hidden layer $l \in \{1, 2\}$, and v^o_n is the value of the output node. The values are computed as illustrated in Figure 3. All the values of the input variables $\{x_{(n,i)}\}_{i=1,\ldots,I}$ are inputs for the computation of values for each of the nodes in the first hidden layer. Further, all the values of the nodes in the first hidden layer are inputs for the computation of values for each of the nodes in the second layer. Finally, all the values of the nodes in the second hidden layer are inputs for the computation of the value of the node in the output layer. Appendix B further details the DNN.

Figure 3: The computation of the values of the nodes in the hidden and output layers of the DNN



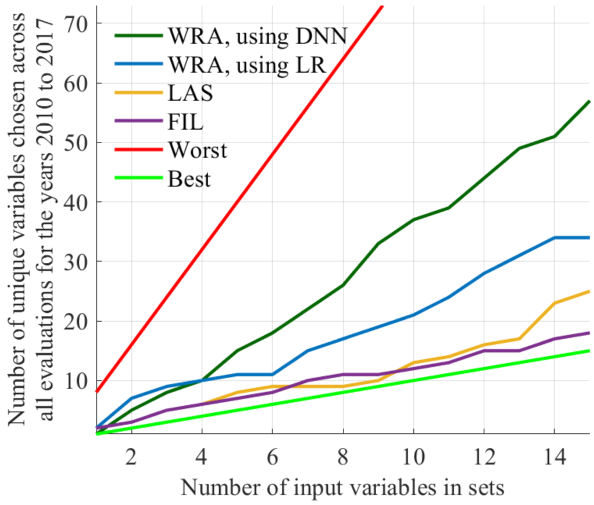
The values of all the nodes (or input variables) in the previous layer are inputs for the computation of each node. First, each of these inputs are multiplied by an associated weight. Second, a sum of all these products and a bias value are the input of a transfer function. Finally, the output of this transfer function is the value of the node.

Figure 4: AUC obtained when using the different input variable sets given in Table 1



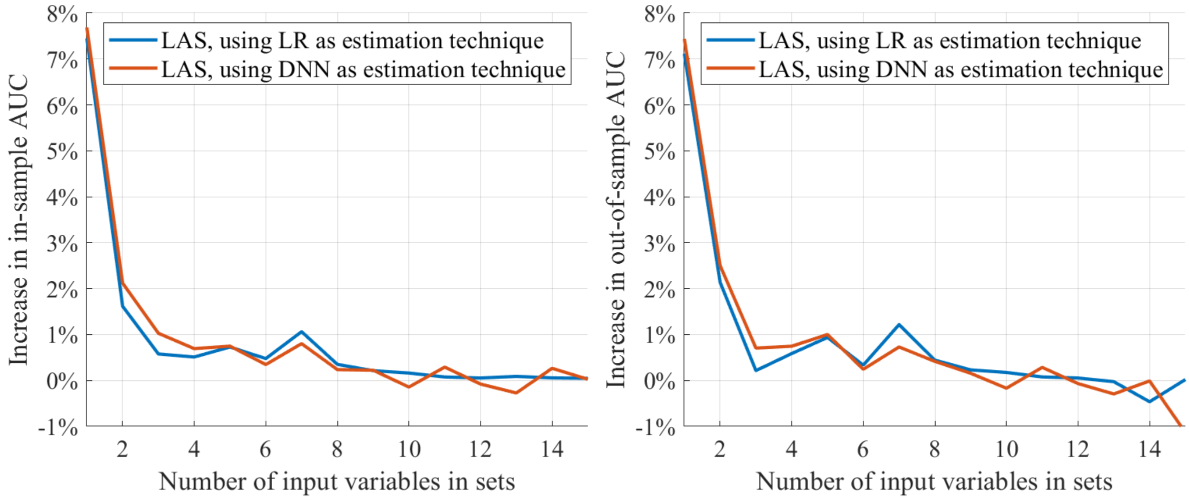
The AUC values are the mean of AUC across the eight evaluations for the years 2010-2017, as shown in Figure 1. For WRA, LAS, FIL, and RAN, we let the predetermined number of input variables in the sets vary between 1 and 15. Top and bottom panels show in-sample AUC and out-of-sample AUC, respectively. Left and right panels display results when using LR and DNN, respectively, as estimation technique.

Figure 5: Stability of variable selection



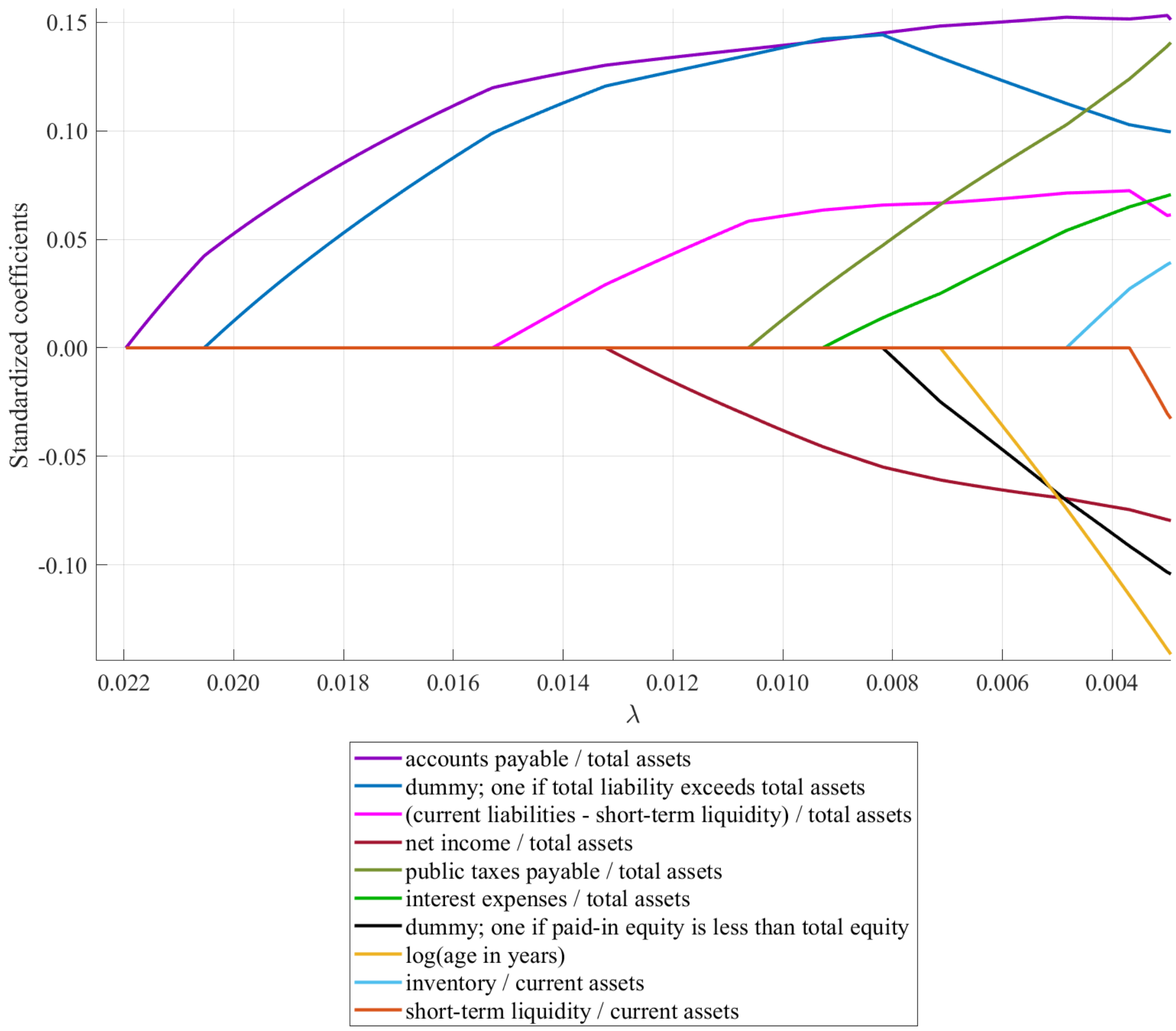
Total number of unique input variables chosen by each feature selection method across the eight evaluations over the years 2010-2017 for different predetermined number of variables in sets. The green and red diagonal lines represent the best and worst possible cases, respectively.

Figure 6: The percentage marginal increase in AUC



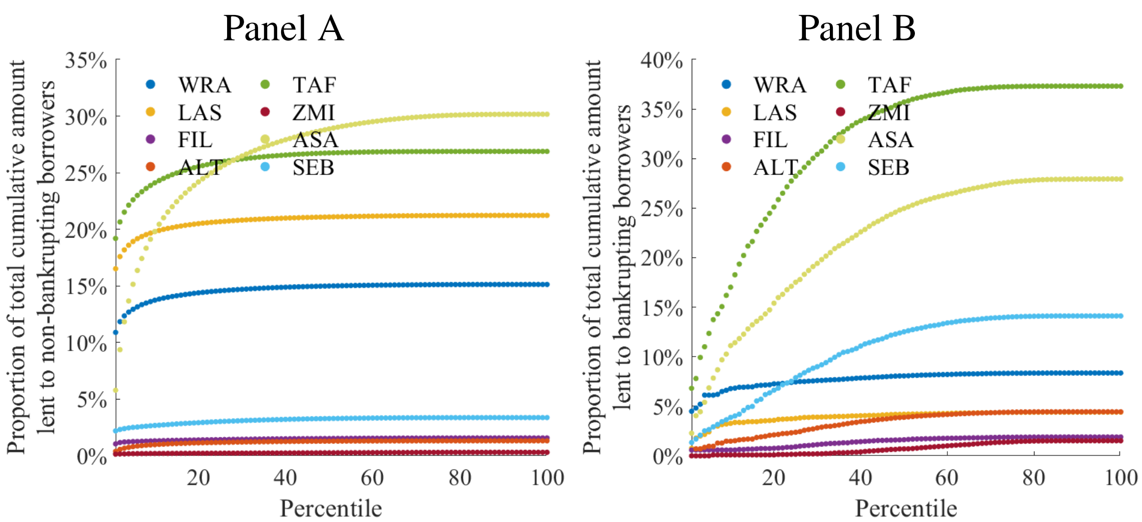
The percentage marginal increase in in-sample (left panel) and out-of-sample (right panel) AUC when expanding the size of the input variable set selected by the LASSO method. The AUC values represent the mean of AUC across the eight evaluations for the years 2010-2017, as defined in Figure 1.

Figure 7: The importance of the variables selected by the LASSO method



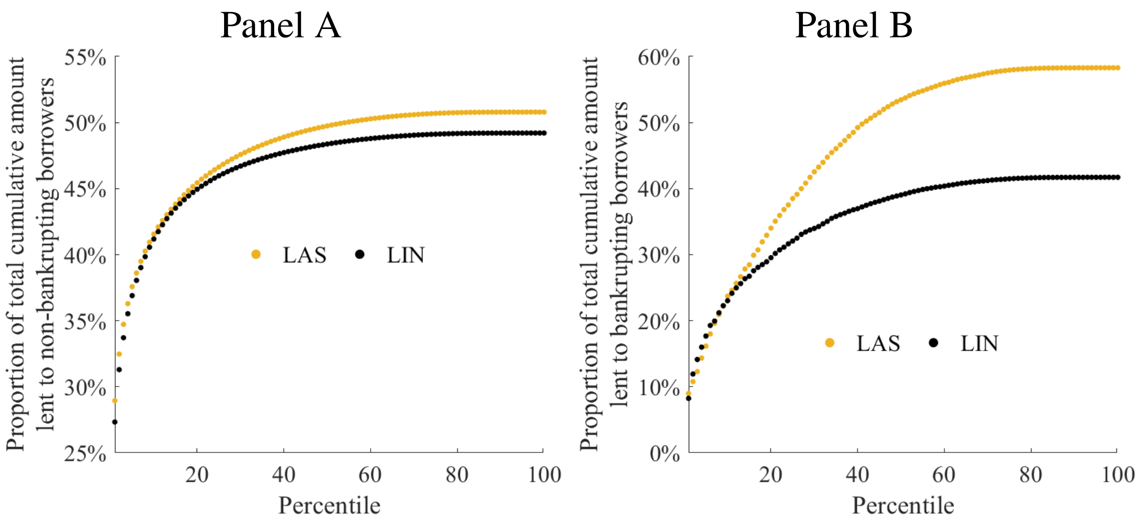
The standardized coefficients in Equation (3) for the variables selected by the LASSO method (see Table 2) for different values of the positive tuning parameter λ when using all data in the sample period 2006-2017.

Figure 8: Total cumulative amount lent within percentiles of amount borrowers are willing to borrow – General versus industry-specific variable sets



All potential borrowers, i.e., financial statements of accounting years 2010-2017 with interest bearing debt, are grouped into percentiles based on the amount they are willing to borrow ordered descendingly. Panel A shows, for each percentile, the amount each bank lends to borrowers who actually not bankrupted up to that percentile as a proportion of the total amount lent to all borrowers who actually not bankrupted. Panel B displays the same for borrowers who actually bankrupted. There are eight banks in the simulation, each using one of the variable sets WRA, LAS, FIL, ALT, TAF, ZMI, ASA and SEB, as shown in Table 1. Each potential borrower chooses to borrow exclusively from the bank offering the lowest credit spread.

Figure 9: Total cumulative amount lent within percentiles of amount borrowers are willing to borrow when assessing bank profitability of industry-specific variable sets



All potential borrowers, i.e., financial statements of accounting years 2010-2017 with interest bearing debt, are grouped into percentiles based on the amount they are willing to borrow ordered descendingly. Panel A shows, for each percentile, the amount each bank lends to borrowers who actually not bankrupted up to that percentile as a proportion of the total amount lent to all borrowers who actually not bankrupted. Panel B displays the same for borrowers who actually bankrupted. There are two banks in the simulation. The first bank uses the variable sets selected by the LASSO method when applied to financial statements of companies from all industries in the sample (LAS). The second bank uses LIN. That is, for each potential borrower, the bank uses the variable set chosen by LASSO that corresponds to the potential borrower's industry (see Table IA.1, Panels A to H, in the Internet Appendix) or the variable set chosen when applying the full sample, as in LAS, for industries that are too small for selecting a separate variable set. Each potential borrower chooses to borrow exclusively from the bank offering the lowest credit spread.

Appendix A. All input variables considered in this study

Table A.1: Input variables for bankruptcy prediction

The table lists all 155 input variables considered in our study. It also reports the Pearson correlation coefficient (PCC) between the values of the input variables and the actual classifications of non-bankrupt (0) or bankrupt (1) for each financial statement. The sample consists of 1,002,325 financial statements whereas 1.72% are categorized as bankrupt. Input variables are sorted according to the absolute value of PCC (in descending order).

Description	PCC
accounts payable / total assets	0.17
dummy; one if total liability exceeds total assets	0.16
total equity / total assets	- 0.15
(current liabilities - short-term liquidity) / total assets	0.15
total liabilities / total assets	0.15
current liabilities / total assets	0.15
net income / total assets	- 0.13
retained earnings / total assets	- 0.13
dummy; one if paid-in equity is less than total equity	- 0.13
pre-tax profit / total assets	- 0.13
EBIT / total assets	- 0.12
working capital / total assets	- 0.12
EBIT / total tangible assets	- 0.12
operating profit / total assets	- 0.12
EBITDA / total assets	- 0.11
(short-term assets - total liabilities) / total assets	- 0.11
interest expenses / total assets	0.11
public taxes payable / total assets	0.10
total expenses / assets	0.10
operating expenses / total assets	0.10
(non-interest expenses - salary) / total assets	0.09
accounts payable / current liabilities	0.09
sales / total tangible assets	0.08
sales / current assets	0.08
sales / total assets	0.08
(shareholder’s equity + total revenues) / total assets	0.08
total revenues / total assets	0.08
investment turnover (sales / (total equity + total liabilities))	0.08
net income / (total liabilities + paid-in capital)	- 0.07
inventory / current assets	0.07
retained earnings / tangible assets	- 0.07
operating profit / paid-in capital	- 0.07
log(age in years)	- 0.07
short-term liquidity / current assets	- 0.07
salary / total assets	0.06

Table A.1: Input variables in our study, descending by the absolute value of PCC (continued)

Description	PCC
(current assets - short-term liquidity) / total assets	0.06
pre-tax profit / paid-in capital	- 0.06
net income / paid-in capital	- 0.06
short-term liquidity / total assets	- 0.05
log(total assets)	- 0.05
effective tax rate	- 0.05
retained earnings / inventory	- 0.05
working capital / long-term liabilities	- 0.05
total equity / long-term liabilities	- 0.04
interest income / total assets	- 0.04
dividends / net income	- 0.03
quick assets / sales	- 0.03
sales / tangible assets	0.03
net quick assets / inventory	- 0.03
operating expenses / sales	- 0.03
(total revenues - sales) / total revenues	- 0.03
short-term liquidity / sales	- 0.03
total equity / (total equity + long-term liabilities)	- 0.03
operating profit / total revenues	0.03
total equity / fixed assets	- 0.03
short-term liquidity as a percentage of the capital employed	- 0.03
quick assets /total assets	- 0.03
interest income / interest expenses	- 0.03
long-term liabilities / total assets	0.03
sales / short-term liquidity	0.02
EBIT / interest expense	- 0.02
current liabilities / total equity	- 0.02
sales / total equity	- 0.02
intangibles / total assets	0.02
current assets / total equity	- 0.02
(total revenues + interest income) / total expenses	- 0.02
sales / working capital	- 0.02
total liabilities / total equity	- 0.02
total revenues / net working capital	- 0.02
working capital / operational expenditure	- 0.02
total revenues / sales	- 0.02
current assets / total assets (net liquid assets / total assets)	0.02
fixed assets / total assets	- 0.02
fixed assets / total equity	- 0.02
(long-term liabilities + total equity) / fixed assets	- 0.02

Table A.1: Input variables in our study, descending by the absolute value of PCC (continued)

Description	PCC
total assets / total revenues	- 0.02
sales / stock holders equity	0.02
working capital / sales	- 0.02
return on net fixed assets	- 0.02
interest-bearing debt / total equity	- 0.02
total revenues / total expenses	- 0.02
long-term liabilities / total equity	- 0.02
personnel costs / added value	- 0.02
current assets / sales	- 0.02
net income / gross profit	0.02
retained earnings / sales	- 0.02
total equity / sales	- 0.02
current liabilities / sales	- 0.02
accounts receivable / accounts payable	- 0.02
non-interest expenses / operating profit	- 0.01
(inventory + accounts receivables) / total equity	- 0.01
accounts payable / sales	- 0.01
operating profit / sales	0.01
net income / total equity	0.01
EBIT / sales	- 0.01
working capital / current liabilities	- 0.01
pre-tax profit / sales	- 0.01
net income / sales	- 0.01
interest expenses / total revenues	- 0.01
operating profit / (operating profit - interest expense)	- 0.01
financial expenses / sales	- 0.01
financial expenses / sales	- 0.01
income gearing	- 0.01
pre-tax profit / ordinary expenses	- 0.01
interest expenses / total expenses	- 0.01
accounts receivable / sales	- 0.01
intangible assets / sales	- 0.01
inventory / working capital	- 0.01
accounts payable / inventories	- 0.01
cost of goods sold / inventory	0.01
quick assets / current liabilities	- 0.01
current liabilities / total liabilities	0.01
sales / inventories	- 0.01
owners equity / total assets	0.01
short-term liquidity / current liabilities	- 0.01

Table A.1: Input variables in our study, descending by the absolute value of PCC (continued)

Description	**PCC**
tales / assets employed	- 0.01
long-term liabilities / current assets	- 0.01
pre-tax profit / total equity	0.01
(total equity - intangible assets) / (total assets - intangible assets - short-term liquidity)	- 0.01
total equity / total liabilities	- 0.01
retained earnings / current liabilities	- 0.01
current assets/total liabilities	- 0.01
pre-tax profit as a percentage of the capital employed	- 0.01
profit before tax/current liabilities	0.01
inventory / cost of goods sold	- 0.01
(quick assets / current liabilities) × interest earned ratio	- 0.01
working capital / total equity	0.00
solvency ratio	- 0.00
return on debt (earnings / total liabilities)	- 0.00
EBIT / total liabilities	- 0.00
short-term liquidity / total liabilities	- 0.00
EBITDA / interest expense	- 0.00
EBITDA / total liabilities	0.00
interest bearing debt / total liabilities	- 0.00
sales / fixed assets	0.00
share of labour costs	0.00
working capital / total revenues	- 0.00
current liabilities / earnings before tax and interest charge	- 0.00
no-credit interval	0.00
inventory / current liability	0.00
(sales - cost of goods sold) / sales	0.00
cost of goods sold / sales	- 0.00
sales / accounts receivable	0.00
current assets / common shareholder's equity	0.00
total revenues / fixed assets	0.00
return on capital employed	- 0.00
operation asset / total asset	0.00
current liabilities / current assets	- 0.00
earnings after tax and interest charge / net capital employed	0.00
profits / net working capital	0.00
inventory / sales	- 0.00
net income / total revenues	- 0.00
fixed assets / (paid-in capital + long-term liabilities)	0.00
accounts receivable / total liabilities	0.00
interest expenses / total liabilities	0.00

Appendix B. The Deep Artificial Neural Network

In this appendix, we provide details about the DNN and how we train it. Let $\mathbf{V^1} = \{v^l_{(n,h)}\}_{n=1,...,N,h=1,...,5}$ be the values of the nodes h for the financial statements n for hidden layer $l \in \{1, 2\}$. The values of the nodes in the first and second hidden layers, respectively, are given by

$$\mathbf{V^1} = \phi\left(\mathbf{X}\mathbf{W^1} + \iota\left(\mathbf{b^1}\right)^\top\right) \tag{B.1}$$

$$\mathbf{V^2} = \phi\left(\mathbf{V^1}\mathbf{W^2} + \iota\left(\mathbf{b^2}\right)^\top\right) \tag{B.2}$$

where $\mathbf{X} = \{x_{(n,i)}\}_{n=1,...,N,i=1,...,I}$ is a matrix of values for input variables i, $\mathbf{W^1} = \{w^1_{i,h}\}$ and $\mathbf{W^2} = \{w^2_{h,h}\}$ are matrices of weights, $\mathbf{b^1} = \{b^1_h\}$ and $\mathbf{b^2} = \{b^2_h\}$ are vectors of biases, $^\top$ denotes the transpose of any matrix or vector, ι is an $N \times 1$ vector of ones, and $\phi(a) = \iota \oslash (\iota + \exp(-a))$ is the logistic transfer function. Further, the vector of the nodes in the output layer $\mathbf{v^o} = \{v^o_n\}$ is given by

$$\mathbf{v^o} = \varphi\left(\mathbf{V^2}\mathbf{w^o} + \iota b^o\right) \tag{B.3}$$

where $\mathbf{w^o} = \{w^o_h\}$ is a vector of weights, b^o is a bias value, and $\varphi(a) = a$ is the linear transfer function. We stack all the Q weights and biases of the DNN in the vector $\pi = \{\pi_q\}_{q=1,...,Q}$. This means that the values of the output nodes can be written as $\mathbf{v^o}(\pi) = \{v^o_n(\pi)\}$.

We train the DNN (i.e., estimate all weights and biases) by using the Levenberg-Marquardt backpropagation algorithm (Marquardt, 1963; Hagan and Menhaj, 1994). It follows an iterative procedure which updates π. Each iteration is called an epoch. When training, we use all financial statements from the four previous accounting years, as described in Figure 1. The training procedure starts with randomly grouping 70% of these financial statements into a training set of size M. Further, it groups the remaining 30% into a validation set of size K. Further, before any epoch, we set the values of π to randomly chosen initial values.

Let $\mathbf{y}^\tau = \{y^\tau_m\}_{m=1,...,M} \in \{0, 1\}^M$ and $\mathbf{y}^\upsilon = \{y^\upsilon_k\}_{k=1,...,K} \in \{0, 1\}^K$ be the vectors of actual classifications of non-bankrupt (0) or bankrupt (1) of the financial statements in the training set and validation set, respectively. Further, let $\mathbf{v}^{\mathbf{o},\tau}(\pi) = \{v^{o,\tau}_m(\pi)\}_{m=1,...,M}$ and $\mathbf{v}^{\mathbf{o},\upsilon}(\pi) = \{v^{o,\upsilon}_k(\pi)\}_{k=1,...,K}$ be the values of the output nodes of the DNN for the financial statements in the training set and validation set, respectively. For each epoch, the training procedure first computes the sum of squared validation errors by

$$\Psi(\pi) = \sum_{k=1}^{K} \left(e^\upsilon_k(\pi)\right)^2 \tag{B.4}$$

where $\mathbf{e}^\upsilon(\pi) = \{e^\upsilon_k(\pi)\}_{k=1,...,K} = \mathbf{y}^\upsilon - \mathbf{v}^{\mathbf{o},\upsilon}(\pi)$ is the vector of validation errors. Second, the fitting procedure computes the numerical approximation of the Jacobian matrix of the training errors $\mathbf{e}^{\tau}(\pi) = \{e_m^{\tau}(\pi)\}_{m=1,\dots,M} = \mathbf{y}^{\tau} - \mathbf{v}^{\mathbf{o},\tau}(\pi)$ at the current point π arranged by:

$$\mathbf{J} = \begin{bmatrix} \frac{de_1^{\tau}}{d\pi_1} & \cdots & \frac{de_1^{\tau}}{d\pi_Q} \\ \vdots & \ddots & \vdots \\ \frac{de_M^{\tau}}{d\pi_1} & \cdots & \frac{de_M^{\tau}}{d\pi_Q} \end{bmatrix}$$

Third, the fitting procedure updates π by

$$\pi = \pi + \left[\mathbf{J}^{\intercal}\mathbf{J} + \mu\mathbf{I}\right]^{-1} \mathbf{J}^{\intercal}\mathbf{e}^{\tau}(\pi) \tag{B.5}$$

where $\mathbf{I}$ is the identity matrix and μ is a scalar for controlling the step size. We use 0.001 as the initial value of μ in the first epoch. Fourth, the fitting procedure uses this updated π to recompute the sum of squared validation errors by $\Gamma(\pi) = \sum_{k=1}^{K}\left(e_k^{\upsilon}(\pi)\right)^2$. Fifth, the fitting procedure updates the value of μ by $\mu = \mu\delta$, where $\delta = 0.1$ if $\Gamma(\pi) < \Psi(\pi)$ and $\delta = 10$ if $\Gamma(\pi) \geq \Psi(\pi)$. This results in bigger steps in Equation (B.5) for each epoch when the training procedure is taking steps in the direction of lower sum of squared validation errors, and smaller steps when the steps are in the direction of higher sum of squared validation errors. Finally, the fitting procedure checks if any stopping criteria is reached. If so, the training procedure stops, and we obtain the values of the weights and biases π. If not, the training procedure continues with the next epoch. As stopping criteria, we use:

1. $\mu > 10^{10}$
2. $\Gamma(\pi) \geq \Psi(\pi)$ for six epochs in a row
3. The training procedure has done 1000 epochs
4. $\Gamma(\pi) = 0$
5. The performance gradient, given as $(\mathbf{J}^{\upsilon})^{\intercal}\,\mathbf{e}^{\upsilon}(\pi)$ where $\mathbf{J}^{\upsilon}$ is the numerical approximation of the Jacobian matrix of $\mathbf{e}^{\upsilon}(\pi)$ at π, falls below 10^{-7}

For the first epochs, typically $\Gamma(\pi) < \Psi(\pi)$, meaning that the predicting accuracy of the DNN on the validation set keeps improving. However, after several epochs, typically the predicting accuracy on the validation set deteriorates for each epoch. This is an indication of overfitting. The training procedure therefore stops, to prevent overfitting, when $\Gamma(\pi) \geq \Psi(\pi)$ for several epochs. This is expressed by the first two stopping criteria above.

Appendix C. The input variables in the benchmark sets

Table A.2: Input variables in the benchmark sets

The model of Altman (1968) (ALT)
EBIT / total assets
retained earnings / total assets
sales / total assets
total equity / total liabilities
working capital / total assets

The model of Taffler (1984) (TAF)
current assets/total liabilities
current liabilities / total assets
no-credit interval
profit before tax/current liabilities

The model of Zmijewski (1984) (ZMI)
current liabilities / current assets
net income / total assets
total liabilities / total assets

The model of Altman and Sabato (2007) (ASA)
current liabilities / total equity
EBITDA / interest expense
EBITDA / total assets
retained earnings / total assets
short-term liquidity / total assets

The SEBRA model (Bernhardsen and Larsen, 2007) (SEB)
accounts payable / total assets
dummy: one if paid-in equity is less than total equity
EBITDA / total liabilities
log(age in years)
log(total assets)
public taxes payable / total assets
total equity / total assets
working capital / total revenues

Internet Appendix for:

Bankruptcy Prediction of Privately Held SMEs Using Feature Selection Methods

Florentina Paraschiv[a,b], Markus Schmid[c], Ranik Raaen Wahlstrøm[b]

[a]*Zeppelin University, Department of Finance, Am Seemooser Horn 20, Friedrichshafen 88045, Germany*
[b]*NTNU Business School, Norwegian University of Science and Technology, 7491 Trondheim, Norway*
[c]*Swiss Finance Institute (SFI) and Swiss Institute of Banking and Finance, University of St. Gallen, Unterer Graben 21, 9000 St. Gallen, Switzerland*

Table IA.1: Industry-specific variable sets

Panel A: Manufacturing

		2010	2011	2012	2013	2014	2015	2016	2017
Variables selected when using all data	**accounts payable / total assets**	X	X	X	X	X	X	X	X
	dummy; one if total liability exceeds total assets	X	X	X	X	X	X	X	X
	(current liabilities - short-term liquidity) / total assets			X	X	X	X	X	X
	net income / total assets	X	X	X	X	X			
	public taxes payable / total assets	X	X	X	X	X	X	X	X
	interest expenses / total assets	X	X	X	X	X	X	X	X
	dummy; one if paid-in equity is less than total equity	X	X	X	X	X	X	X	X
	log(age in years)	X	X				X		X
	inventory / current assets	X	X	X	X		X		
	short-term liquidity / current assets							X	
Variables not selected when using all data	**pre-tax profit / total assets**					X	X	X	X
	EBIT / total tangible assets								X
	earnings after tax and interest charge / net capital employed	X							
	operating profit / total assets		X	X					
	return on capital employed				X				
	pre-tax profit as a percentage of the capital employed					X			
	intangibles / total assets							X	

The nine input variables selected for the years 2010-2017 by the LASSO method when applied to financial statements of companies from the industry *'Manufacturing'*. The upper part of the table includes variables also chosen when using the full sample (see Table **??**). The lower part of the table lists other variables that are chosen in this specific industry for the respective year. The industry includes 65,071 observations, 1,161 of which are classified as bankrupt (1.75%).

Panel B: Construction

		2010	2011	2012	2013	2014	2015	2016	2017
Variables selected when using all data	**accounts payable / total assets**	X	X	X	X	X	X	X	X
	dummy; one if total liability exceeds total assets	X	X	X	X	X	X	X	X
	(current liabilities - short-term liquidity) / total assets	X	X	X	X	X	X	X	X
	net income / total assets	X	X	X	X	X	X	X	X
	public taxes payable / total assets	X	X	X	X	X	X	X	X
	interest expenses / total assets		X	X	X	X			
	dummy; one if paid-in equity is less than total equity	X	X	X	X	X	X		
	log(age in years)	X	X	X	X	X	X	X	X
	short-term liquidity / current assets						X	X	X
Variables not selected when using all data	**total equity / total assets**	X	X						
	sales / stock holders equity							X	X
	pre-tax profit / total assets	X							
	operating profit / paid-in capital			X	X				
	retained earnings / tangible assets					X			
	cost of goods sold / inventory						X		
	tales / assets employed							X	
	net quick assets / inventory								X

The nine input variables selected for the years 2010-2017 by the LASSO method when applied to financial statements of companies from the industry *'Construction'*. The upper part of the table includes variables also chosen when using the full sample (see Table **??**). The lower part of the table lists other variables that are chosen in this specific industry for the respective year. The industry includes 191,528 observations, 3,939 of which are classified as bankrupt (2.02%).

Panel C: Wholesale and retail trade; repair of motor vehicles and motorcycles

		2010	2011	2012	2013	2014	2015	2016	2017
Variables selected when using all data	**accounts payable / total assets**	X	X	X	X	X	X	X	X
	dummy; one if total liability exceeds total assets	X	X	X	X	X	X	X	X
	(current liabilities - short-term liquidity) / total assets	X	X	X	X	X	X	X	X
	net income / total assets	X	X	X	X	X	X	X	X
	public taxes payable / total assets				X	X	X	X	X
	interest expenses / total assets	X	X	X	X	X	X	X	X
	dummy; one if paid-in equity is less than total equity	X	X	X	X	X	X	X	X
	log(age in years)	X	X	X	X	X	X	X	X
	inventory / current assets		X	X	X	X	X	X	X
Variables not selected when using all data	**EBITDA / total assets**	X	X	X					
	pre-tax profit / total assets	X							

The nine input variables selected for the years 2010-2017 by the LASSO method when applied to financial statements of companies from the industry *'Wholesale and retail trade; repair of motor vehicles and motorcycles'*. The upper part of the table includes variables also chosen when using the full sample (see Table **??**). The lower part of the table lists other variables that are chosen in this specific industry for the respective year. The industry includes 238,178 observations, 6,243 of which are classified as bankrupt (2.55%).

Panel D: Transportation and storage

		2010	2011	2012	2013	2014	2015	2016	2017
Variables selected when using all data	**accounts payable / total assets**	X	X	X	X	X	X	X	X
	dummy; one if total liability exceeds total assets	X	X	X	X	X	X	X	X
	(current liabilities - short-term liquidity) / total assets				X	X	X	X	X
	net income / total assets	X			X	X	X	X	X
	public taxes payable / total assets	X	X	X	X	X	X	X	X
	interest expenses / total assets	X	X	X	X	X	X	X	X
	dummy; one if paid-in equity is less than total equity	X	X	X	X				
	log(age in years)	X	X	X	X	X	X	X	X
	short-term liquidity / current assets		X	X	X				
Variables not selected when using all data	**operating profit / total assets**	X	X	X					
	net quick assets / inventory					X	X	X	X
	return on capital employed					X			
	total equity / (total equity + long-term liabilities)	X					X	X	
	sales / current assets		X						
	accounts payable / current liabilities			X					
	short-term liquidity as a percentage of the capital employed								X

The nine input variables selected for the years 2010-2017 by the LASSO method when applied to financial statements of companies from the industry *'Transportation and storage'*. The upper part of the table includes variables also chosen when using the full sample (see Table **??**). The lower part of the table lists other variables that are chosen in this specific industry for the respective year. The industry includes 48,885 observations, 868 of which are classified as bankrupt (1.74%).

Panel E: Accommodation and food service activities

		2010	2011	2012	2013	2014	2015	2016	2017
Variables selected when using all data	**accounts payable / total assets**	X	X	X	X	X	X	X	X
	dummy; one if total liability exceeds total assets	X	X	X	X	X	X	X	X
	(current liabilities - short-term liquidity) / total assets						X		X
	net income / total assets	X	X	X	X	X	X	X	X
	public taxes payable / total assets	X	X	X	X	X	X	X	X
	interest expenses / total assets	X	X	X	X	X	X	X	
	dummy; one if paid-in equity is less than total equity	X	X	X	X	X		X	X
	log(age in years)	X	X	X	X	X	X	X	X
	inventory / current assets	X	X	X	X	X			
Variables not selected when using all data	**working capital / total assets**			X			X		
	sales / current assets						X	X	X
	log(total assets)	X	X			X		X	X
	retained earnings / tangible assets				X				

The nine input variables selected for the years 2010-2017 by the LASSO method when applied to financial statements of companies from the industry *'Accommodation and food service activities'*. The upper part of the table includes variables also chosen when using the full sample (see Table **??**). The lower part of the table lists other variables that are chosen in this specific industry for the respective year. The industry includes 47,644 observations, 1,631 of which are classified as bankrupt (3.31%).

Panel F: Information and communication

		2010	2011	2012	2013	2014	2015	2016	2017
Variables selected when using all data	**accounts payable / total assets**	X	X	X	X	X	X	X	X
	dummy; one if total liability exceeds total assets	X	X	X	X	X	X	X	X
	(current liabilities - short-term liquidity) / total assets		X	X	X	X	X	X	X
	net income / total assets	X	X	X	X	X			
	public taxes payable / total assets	X	X	X	X	X	X	X	X
	interest expenses / total assets	X	X	X	X	X	X	X	X
	dummy; one if paid-in equity is less than total equity	X	X	X	X				
	log(age in years)	X	X						
Variables not selected when using all data	**total equity / total assets**								X
	working capital / total assets						X		
	retained earnings / tangible assets					X	X	X	X
	EBIT / total tangible assets						X	X	X
	pre-tax profit / total assets	X							
	operating expenses / total assets					X	X		
	working capital / long-term liabilities				X	X			
	effective tax rate		X						
	working capital / total equity				X				
	net quick assets / inventory							X	X
	operating profit / paid-in capital	X							
	net income / total equity			X					
	(non-interest expenses - salary) / total assets			X					
	sales / current assets							X	

The nine input variables selected for the years 2010-2017 by the LASSO method when applied to financial statements of companies from the industry *'Information and communication'*. The upper part of the table includes variables also chosen when using the full sample (see Table **??**). The lower part of the table lists other variables that are chosen in this specific industry for the respective year. The industry includes 51,950 observations, 553 of which are classified as bankrupt (1.05%).

Panel G: Professional, scientific and technical activities

		2010	2011	2012	2013	2014	2015	2016	2017
Variables selected when using all data	accounts payable / total assets	X	X	X	X	X	X	X	X
	dummy; one if total liability exceeds total assets	X	X	X	X	X			
	(current liabilities - short-term liquidity) / total assets	X	X	X	X	X	X	X	X
	net income / total assets	X	X	X	X	X	X	X	X
	public taxes payable / total assets	X	X	X	X		X	X	X
	interest expenses / total assets	X	X	X	X	X	X	X	
	dummy; one if paid-in equity is less than total equity	X	X	X	X	X	X	X	X
	log(age in years)		X						
	inventory / current assets			X	X				
Variables not selected when using all data	retained earnings / total assets	X				X			
	(non-interest expenses - salary) / total assets					X	X	X	X
	total liabilities / total assets					X			
	pre-tax profit / total assets	X							
	EBIT / total tangible assets							X	
	sales / current assets								X
	accounts payable / current liabilities		X	X	X		X		
	sales / stock holders equity								X
	intangibles / total assets						X		
	net quick assets / inventory							X	X

The nine input variables selected for the years 2010-2017 by the LASSO method when applied to financial statements of companies from the industry *'Professional, scientific and technical activities'*. The upper part of the table includes variables also chosen when using the full sample (see Table **??**). The lower part of the table lists other variables that are chosen in this specific industry for the respective year. The industry includes 169,498 observations, 1,090 of which are classified as bankrupt (0.64%).

Panel H: Administrative and support service activities

		2010	2011	2012	2013	2014	2015	2016	2017
	accounts payable / total assets	X	X	X	X	X	X	X	X
	dummy; one if total liability exceeds total assets	X	X	X	X	X	X	X	X
	(current liabilities - short-term liquidity) / total assets	X	X	X	X	X	X	X	X
Variables	**net income / total assets**	X	X	X	X	X	X		
selected when	**public taxes payable / total assets**	X	X	X	X	X	X	X	X
using all data	**interest expenses / total assets**	X	X	X	X	X	X	X	X
	dummy; one if paid-in equity is less than total equity	X				X			X
	log(age in years)	X	X	X	X	X	X	X	X
	short-term liquidity / current assets	X	X	X	X	X	X	X	X
	sales / stock holders equity							X	X
Variables not	**working capital / total equity**			X					
selected when	**sales / current assets**						X	X	
using all data	**current liabilities / total equity**		X						
	(inventory + accounts receivables) / total equity				X				

The nine input variables selected for the years 2010-2017 by the LASSO method when applied to financial statements of companies from the industry *'Administrative and support service activities'*. The upper part of the table includes variables also chosen when using the full sample (see Table **??**). The lower part of the table lists other variables that are chosen in this specific industry for the respective year. The industry includes 52,198 observations, 862 of which are classified as bankrupt (1.62%).