Hjelseth, Ida Nervik; Raknerud, Arvid; Vatne, Bjørn Helge

Working Paper

A bankruptcy probability model for assessing credit risk on corporate loans with automated variable selection

Working Paper, No. 7/2022

Provided in Cooperation with:
Norges Bank, Oslo

Suggested Citation: Hjelseth, Ida Nervik; Raknerud, Arvid; Vatne, Bjørn Helge (2022) : A bankruptcy probability model for assessing credit risk on corporate loans with automated variable selection, Working Paper, No. 7/2022, ISBN 978-82-8379-237-9, Norges Bank, Oslo, https://hdl.handle.net/11250/3011180

This Version is available at:
https://hdl.handle.net/10419/298407

WORKING PAPER

A bankruptcy probability model for assessing credit risk on corporate loans with automated variable selection

NORGES BANK RESEARCH

7 | 2022

IDA NERVIK HJELSETH,
ARVID RAKNERUD,
BJØRN H. VATNE

NORGES BANK

Working papers fra Norges Bank, fra 1992/1 til 2009/2 kan bestilles over e-post:
111facility@norges-bank.no

Fra 1999 og senere er publikasjonene tilgjengelige på www.norges-bank.no

Working papers inneholder forskningsarbeider og utredninger som vanligvis ikke har fått sin endelige form. Hensikten er blant annet at forfatteren kan motta kommentarer fra kolleger og andre interesserte. Synspunkter og konklusjoner i arbeidene står for forfatternes regning.

Working papers from Norges Bank, from 1992/1 to 2009/2 can be ordered by e-mail
FacilityServices@norges-bank.no

Working papers from 1999 onwards are available on www.norges-bank.no

Norges Bank's working papers present research projects and reports (not usually in their final form) and are intended inter alia to enable the author to benefit from the comments of colleagues and other interested parties. Views and conclusions expressed in working papers are the responsibility of the authors alone.

ISSN 1502-8190 (online)
ISBN 978-82-8379-237-9 (online)

A bankruptcy probability model for assessing credit risk on corporate loans with automated variable selection *

Ida Nervik Hjelseth[†] Arvid Raknerud[‡] Bjørn H. Vatne[§]

Norges Bank, Financial Stability

June 20, 2022

Abstract

We propose an econometric model for predicting the share of bank debt held by bankrupt firms by combining a novel set of firm-level financial variables and macroeconomic indicators. Our firm-level data include payment remarks in the form of debt collections from private agencies and attachments from private and public agencies and cover all Norwegian limited liability companies for the period 2010–2021. We use logistic Lasso regressions to select bankruptcy predictors from a large set of potential predictors, comparing a highly sparse variable selection criterion ("the one standard error rule") with the minimum cross validation error (CVE) criterion. Moreover, we examine the implications of using debt shares as weights in the estimation and find that weighting has a large impact on variable selection and predictions and, generally, leads to lower out-of-sample prediction errors than alternative approaches. Debt weighting combined with sparse variable selection gives the best predictions of the risk of bankruptcy in firms holding high shares of the bank debt.

JEL: C25, C33, C53, G33, D22
Keywords: Bankruptcy prediction, credit risk, corporate bank debt, Lasso, weighted logistic regression

*This paper should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. We would like to thank Kasper Roszbach, Paolo Giordani, Christian Bjørland, Henrik Borchgrevink and Haakon Solheim for valuable comments and suggestions.
[†]Norges Bank: Division of Financial Stability (ida-nervik.hjelseth@norges-bank.no)
[‡]Statistics Norway: Research Department and Norges Bank: Division of Financial Stability. Corresponding author: arvid.raknerud@ssb.no
[§]Norges Bank: Division of Financial Stability (bjorn-helge.vatne@norges-bank.no)

1 Introduction

Assessing corporate credit risk is an important part of monitoring the soundness of the banking system, as a large proportion of bank lending is to the corporate sector and losses on corporate loans have historically exceeded losses on household loans both during banking crises and in normal times (see Kragh-Sørensen and Solheim (2014)).

Before the pioneering work of Beaver (1966) and Altman (1968), financial institutions' analyses of credit risk on corporate loans were largely subjective judgments ("expert" opinions) based on a few key variables such as leverage, collateral and earnings. Beaver (1966) was the first to use financial ratios for predicting bankruptcy. While his study look at one ratio at a time, Altman (1968) used multiple discriminant analysis to combine information from several financial ratios in a single prediction. Altman's so-called Z-model was popular for several decades. The risk-of-ruin model of Wilcox (1973) and the option pricing model of Merton (1974) are early examples of more theoretically founded credit risk models.[1]

Modern empirical credit score analysis combine information from several data sources to obtain a credit risk score or default probability, typically using a probability model (linear, logit or probit). To improve default predictions and understand which variables are the most important drivers of defaults, the more recent literature has focused on the importance of dynamic (panel data) as opposed to static (cross-section) models (see Shumway (2001)); the importance of flexible functional forms, e.g. hazard functions with splines (see Giordani et al. (2014)); and machine learning methods, such as classification trees, neural networks, gradient boosting and the Lasso (see Matin et al. (2019), Christoffersen et al. (2018), Li and Sun (2013), Kim and Kang (2010), Min and Lee (2005), and Jones et al. (2017)).

A bankrupt firm will almost surely default on the claims held by its creditors – a large share of which are banks – and thereby invoke credit losses. Bank debt held by bankrupt firms, henceforth referred to as *bankruptcy debt*, is therefore a potentially important mechanism through which corporate failure is propagated in the economy. Although not all loan defaults are related to bankruptcies, aggregate bankruptcy debts (e.g. for an industry) follow very similar patterns over time as banks' defaults on loans to the corporate sector. This has been shown by Hjelseth and

[1] In the risk-of-ruin model, a firm will go bankrupt if the value of its assets falls below that of its debt obligations, whereas in the Merton model a firm's probability of bankruptcy depends on asset value relative to the value of outside debt.

Raknerud (2016), Bernhardsen (2001) and Bernhardsen and Larsen (2007).[2] When assessing the credit risk associated with a bankruptcy, the amount of debt held by the firm is of key importance. For the solvency of a bank, or the soundness of the banking sector as a whole, the potential bankruptcy of firms holding high shares of total bank loans poses higher risk than that of firms with only small loans. Because the true bankruptcy probability function is unknown – and maybe of infinite complexity – the empirical method of fitting prediction models should be tuned towards fitting the observations best that are most important to the target of prediction. In our case, the target of prediction is bankruptcy debt rates at the industry level, defined as:

$$s_t(I) = \sum_{i \in \mathcal{B}(I,t)} w_{i,t-1}$$

where $\mathcal{B}(I,t)$ is the *unknown* set of firms that go bankrupt in the given year (t) and industry (I) and $w_{i,t-1}$ is the pre-determined share of firm i's total bank debt in the industry (observed from the outgoing balance of year $t-1$).

When fitting models used to make probability predictions about $\mathcal{B}(I,t)$, it may be useful to put more weight on observations corresponding to large $w_{i,t-1}$, as these will have a larger impact on the target variable, $s_t(I)$. However, in the vast literature on bankruptcy modelling, equal weight is given to "small" and "large" firms when selecting and estimating models, with heterogeneity in size characteristics controlled for through control variables, such as e.g. number of employees (see Jacobson and von Schedvin (2015)), total sales (Carling et al. (2007)), total assets (Christoffersen et al. (2018)), or numerous other measures of firm size. The first – of two – main contributions of this paper, is that we explicitly consider the weighting of observations in the estimation of bankruptcy probability models as an *alternative* to including size-related control variables.

The second main contribution is that we assemble a unique panel data set that includes firm-level payment remarks related to financial claims from both private and public agencies, covering all Norwegian limited liability companies over a relatively long time period, 2011–2020. To predict bankruptcy debt, we combine payment remarks with accounting variables and industry-specific economic indicators (as in Carling et al. (2007)). Payment remarks are based on notifications of overdue payments from private debt collection companies, and attachments from local governments, tax authorities, the Norwegian Labour and Welfare Administration, and

[2]Since bank-firm-customer data on loan defaults are available in Norway for a very short time period, we focus in this paper on bankruptcy debt.

private companies. In the existing literature on bankruptcy risk, payment remarks are seldom available and, if available, not from public registers, but for selected credit agencies and with limited coverage of firms.[3] The broad coverage, makes our data uniquely valuable for predicting bankruptcy debt rates at the aggregate level, such as for a whole industry.

When using machine learning methods to choose among candidate models with similar prediction errors but different number of predictors, sparsity is a highly desirable feature. Loosely speaking, a sparse statistical model is one in which a small number of predictors play an important role. Sparsity facilitates interpretability and makes the models easier to estimate and apply in practice, e.g. for scenario analyses or forecasting purposes. However, many of the most popular and successful machine learning methods, such as neural networks and gradient boosting, are "black box" prediction models that do not provide interpretable relations between variables (see e.g. the discussion in Hastie et al. (2015)). To facilitate interpretability, we propose using logistic Lasso regression for variable selection, with regularization parameter chosen by means of cross validation (CV). Our most sparse model selection criterion combines CV with the "one standard error (SE) rule": the method selects the most sparse model whose prediction error is one standard error worse than the minimum CV error (CVE), see Chen and Yang (2021).

We first show that using debt shares, $w_{i,t-1}$, as weights, has a large impact both on variable selection by means of (weighted) Lasso and on the subsequent fitting of bankruptcy models by means of (weighted) logistic regressions ("post selection estimation"). Weighted Lasso in combination with the one SE rule generally leads to selection of a small number of variables. In particular, variables related to firm size are never among the chosen predictors. Second, we compare the variable selection and predictions using weights, with a model fitted without weighting, henceforth referred to as the *unweighted benchmark* model. In this model, firm size measures are included as control variables *ex ante*, which is necessary in order not to vastly over-predict bankruptcy debt rates. The reason is that large firms *cet. par.* are associated with significantly lower bankruptcy risk and higher debt shares than small firms. In the benchmark model, variables are selected by means of unweighted Lasso augmented with a polynomial in log assets to capture the effect of size (since the polynomial terms are not automatically selected in the first place). The model is then fitted to bankruptcy data by means of ordinary

[3]For example Carling et al. (2007) analyze default probabilities on the business loan portfolio of a large Swedish bank in 1994–2000.

logistic regression. Compared to the unweighted benchmark model, debt-weighted Lasso and debt-weighted logistic regression are associated with both fewer predictors being selected and lower out-of-sample root mean squared error (RMSE) when used to predict $s_t(I)$. On the other hand, the latter method is more vulnerable to outliers in the data and considerably reduces the efficient sample size (see Kish (1965)). Third, we consider weighted and unweighted models where the predictors are chosen by minimizing CVE. These models include many more predictors, and perform similarly to the unweighted benchmark model – as they generally include a polynomial in log assets as predictors – but generally worse than the weighted model based on the one SE rule.

The rest of the paper is organized as follows. In Section 2, we present the data and our sample. In Section 3, we introduce our econometric model of bankruptcy prediction and in Section 4 we present estimates of parameters, marginal effects and (out-of-sample) predictions of aggregate bankruptcy debt at the industry level. Finally, Section 5 concludes.

2 Data, sample and operationlizations

Our data consist of income statements, balance sheets, bankruptcies and other firm-specific information for all Norwegian-registered firms that submit their financial statements to the Brønnøysund Register Centre, which is the national registry data manager in Norway.[4] This data are merged with firm-level data on payment remarks from a register using organization numbers as firm identifiers.[5]

We restrict our sample to non-consolidated financial statements for all non-financial limited liability firms with a registered industry code. Limited liability firms stand for nearly 90 percent of the total debt to credit institutions held by all non-financial firms that submit financial statements. Since we are interested in banks' credit risk associated with loans to the corporate sector, we exclude observations of firms without bank debt.[6]

We have grouped the firms into six different industries: fishing and fish farming, manufacturing, construction, retail trade, commercial real estate (CRE) and services. That leaves us with one

[4]The financial statements data are delivered by a credit rating agency (Dun & Bradstreet), while the other firm-specific information is delivered directly from the Brønnøysund Register Centre.
[5]The data are collected and stored by Dun & Bradstreet.
[6]The financial statements data include information about debt to credit institutions, here referred to as "bank debt", at the end of the accounting year for each firm.

residual group: "other industries", which includes firms in international shipping, oil and gas exploration, support activities for oil and gas exploration, electricity and water supply, renovation activities, agriculture and forestry. These are industries where there is a considerable mismatch between banks' loan portfolios and the financial statements population. In the shipping industry, for example, a large share of Norwegian banks' loans is to foreign firms. These are not included in our data, and the relationship between banks' loan losses in shipping and the corresponding bankruptcy debt rates is weak. The residual group also include some very large publicly owned companies, especially in electricity and water supply (e.g., Statkraft) and in oil exploration (e.g., Equinor). These companies represent a large share of total debt in the industry, but have a negligible probability of bankruptcy. For the above mentioned reasons we exclude "other industries" from our analyses.

As in Carling et al. (2007), our model includes macroeconomic indicators relevant for predicting bankruptcy in the industry. Annual mainland real GDP growth is the macroeconomic indicator used for manufacturing, construction, retail trade and services. Furthermore, a real prime yield rate for office space in Oslo is used for CRE, and the log real salmon price is used for for fishing and fish farming.

We have financial statements data for the accounting years 1999–2020, bankruptcy registrations for 1999–2021 and payments remarks (debt collections and attachments) for 2010–2021. Our reported estimation results are restricted to bankruptcies registered in 2011–2021, because this is the (longest) period where payment remark variables can be used as *predictors* of bankruptcy, i.e., observed at least one year before the bankruptcy event.

2.1 Definition of bankruptcy and descriptive statistics

The timing of a bankruptcy registration can vary because bankruptcy proceedings may be uncertain and time consuming. There is typically a lag of one or two years between the date of the last registered activity and the date of bankruptcy in the registers (see Hjelseth and Raknerud (2016) for details). To address the timeliness issue, we identify t as the year of the bankruptcy event if the firm was active at the *end* of the previous year $(t-1)$ and is declared bankrupt in year t *or* $t+1$.[7] Registered activity in a given year either means that the firm

[7]The firm is also defined as bankrupt if the liquidation of the firm was registered as compulsory, as these firms are shown to have some of the same properties as bankrupt firms.

Table 2.1: Number of firms and share of total bank debt in the accounting year 2020. Average bankruptcy frequencies and bankruptcy debt rates, $s_t(I)$, 2011–2020. By industry. Percent.

Industry	No. of firms	Share of total bank debt	Bankruptcy frequency	Bankruptcy debt rate
Fishing and fish farming	1,435	4.6	0.7	0.2
Manufacturing[1)]	4,651	5.8	1.9	0.8
Construction	10,333	2.3	2.7	1.3
Retail trade	11,679	3.7	3.1	1.3
Commercial real estate[2)]	28,816	42.3	0.4	0.2
Services[3)]	20,953	27.1	1.8	0.3
Other industries[4)]	1,619	14.1	1.0	0.6

1) Includes mining and quarrying.
2) Includes property development.
3) Information and communication, commercial services, public services, transportation and storage services, accommodation and food service activities, arts, entertainment, recreation and other personal service activities.
4) Oil and gas exploration, support activities for oil and gas exploration, international shipping, electricity and water supply and renovation activities, agriculture and forestry.

filed financial statements, or there was a new credit rating (by Dun & Bradstreet) of the firm. In the case of missing financial statements for a firm that had a new credit rating, these were imputed using the financial statements from the previous year. Our bankruptcy definition picks up about 85 percent of the registered bankruptcies in the sample. In the remaining cases, there is a gap of more than two years between the last registered economic activity and the bankruptcy registration. Thus, in these cases the corresponding firm exits are not treated as bankruptcy events by our definition.

Table 2.1 shows the number of firms in our final sample for the accounting year 2020, together with the percentage share of bank debt in each industry, and each industry's average annual bankruptcy frequency and bankruptcy debt rate in 2011–2020 (the latter is the average of $s_t(I)$ over t, for each industry, I). As seen from the table, the average bankruptcy debt rate is much lower than the (arithmetic) average bankruptcy frequency in all industries, showing that there is generally a negative relationship between the amount of bank debt in firms and their bankruptcy probability.

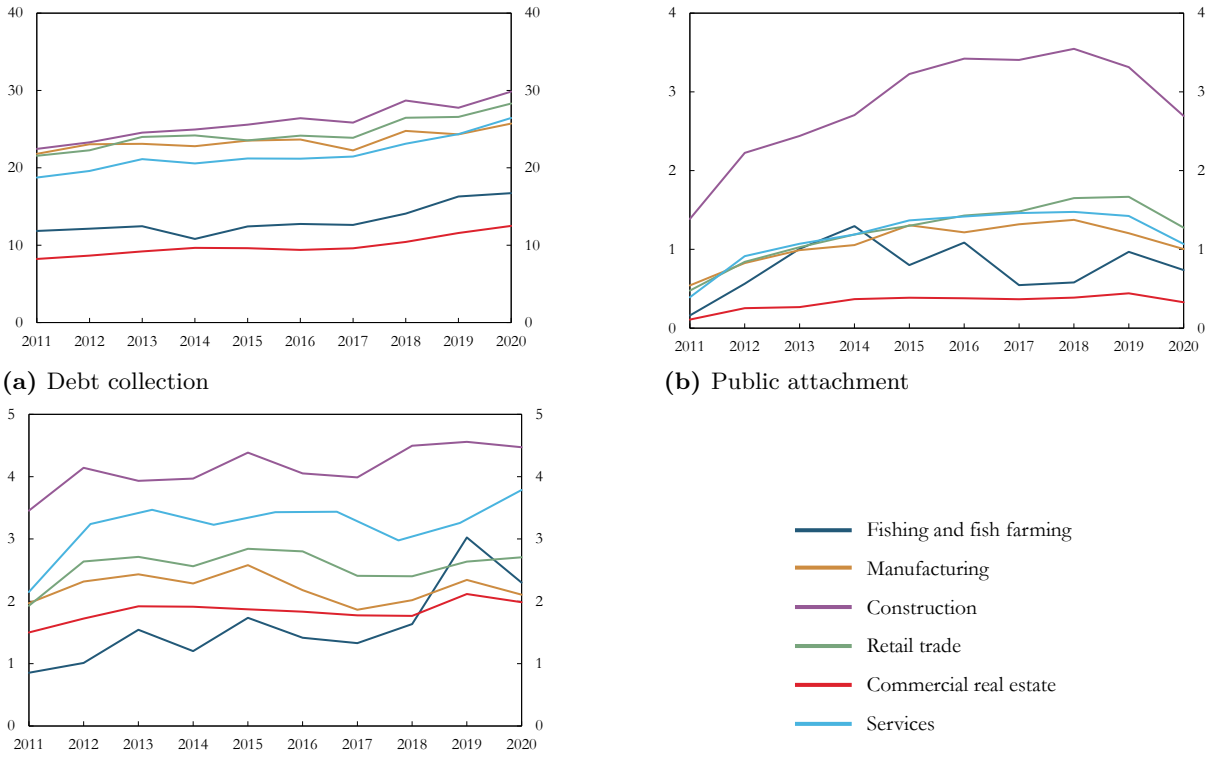
2.2 Firm-specific predictors of bankruptcy

From the financial statements, we construct standard financial indicators related to profitability, liquidity, financial structure and equity: return on assets, equity ratio, current liability ratio, interest coverage ratio, log of real total assets, and many more. In the variable selection stage of our bankruptcy prediction modelling (see Section 3), we also allow for lagged values and interaction terms – in total more than 1000 variables (see Table A.1 in Appendix A.2 for a full

listing and definitions).

Hjelseth and Raknerud (2016) document that a firm's credit rating is a strong predictor of bankruptcy. However, a drawback of using the credit rating as an explanatory variable is that the model becomes nontransparent. Since information on payment remarks is an important determinant of credit ratings (see Appendix A.1), we use, in addition to financial statements data, information on payment remarks related to financial claims from private and public agencies. Payment remarks are published in real time and based on notifications of overdue payments from private debt collection companies, and attachments from local governments, tax authorities, the Norwegian Labour and Welfare Administration (which collects payroll taxes), and private companies.

Chart 2.1: Three types of payment remarks by industry. Share of firms in each industry with at least one remark of the given type. 2011–2020. Percent.



(a) Debt collection

(b) Public attachment

(c) Private attachment

Source: Norges Bank

We divide payment remarks into three groups: 1) debt collections from private agencies, 2) attachments by public authorities and 3) attachments by private agencies. Charts 2.1a–2.1c show the pattern of debt collections and attachments over time and across industries. The share of firms with debt collections varies widely across industries, but displays an increasing trend:

from 8–22 percent in 2011 to 12–30 percent in 2020. The shares are highest in construction and lowest in CRE. Attachments are more rare than debt collections: less than 4 percent of firms in any year or industry have attachments by public authorities and less than 5 percent by private creditors. For both types of attachments, the shares are highest in construction. There was a drop in public attachments from 2019 to 2020 in all industries. The drop in public attachments (Chart 2.1b) reflects government policies to extend deadlines for payments of taxes during the covid-19 pandemic. Since 75 percent of all bankruptcy petitions in Norway are filed by tax authorities (see Oslo Tingrett (2021)), this policy measure is likely to have substantially reduced the number of bankruptcy registrations in 2020 and 2021.

There are many possibilities for constructing bankruptcy predictors from payment remarks data. As guidance in constructing relevant indicators, we utilize the relationship between the credit ratings from Dun & Bradstreet and payment remarks, since credit ratings can be seen as a ranking of default probabilities. In Appendix A.1 we show, by means of a classification tree, that dummy variables for the three types of payment remarks can be mapped into an ordered categorical variable, Claims (C), and combined with a dummy for negative equity, to accurately predict the two lowest credit rating categories. The variable C is constructed as follows:

$$C = \begin{cases} 0: & \text{no payment remarks} \\ 1: & \text{debt collection without attachment} \\ 2: & \text{collection with public and/or private attachment} \end{cases}$$

As there are very few observations with all three types of remarks, we do not separate between firms that have both types of attachments and firms with only one type: In both cases $C = 2$. There are a few anomalous cases with attachment but no collection in the data. These are also classified as $C = 2$.

3 An econometric model of bankruptcy debt prediction

It is useful to reformulate the targets of prediction defined in Section 1, i.e., the industry-level bankruptcy debt rates, $s_t(I)$, as:

$$s_t(I) = \sum_{i \in \mathcal{F}(I)} w_{i,t-1} B_{it}$$

where B_{it} is the bankruptcy indicator which is 1 if t is the year of bankruptcy (see Section 2.1 for definition) and $\mathcal{F}(I)$ is the set of firms in industry I. The debt shares $w_{i,t-1}$ sum to one when summing over all $i \in \mathcal{F}(I)$ for given I and t. In real time, $s_t(I)$ must be predicted from an information set, $\mathcal{I}_t$, which is assumed to include $w_{i,t-1}$ (observed from the *outgoing* balance of year $t-1$). Replacing B_{it} in Equation (3) with firm-level bankruptcy probability predictors, $\widehat{B}_{it}$, yields the corresponding aggregate predictor, $\widehat{s}_t(I)$.

3.1 Bankruptcy probability

We assume that the probability of bankruptcy in t of a firm that is active at the end of $t-1$ is given by $Pr(B_{it} = 1|\mathcal{I}_t) = p_{it}(\theta)$, where $p_{it}(\theta)$ is a logit function:

$$\ln\left(\frac{p_{it}(\theta)}{1 - p_{it}(\theta)}\right) = \beta X_{i,t-1} + \gamma z_t \tag{1}$$

The *predicted* bankruptcy debt rate based on the logit model is then:

$$\widehat{s}_t(I) = \sum_{i \in F(I)} w_{i,t-1} p_{it}(\widehat{\theta}). \tag{2}$$

In Equation (1), $X_{i,t-1}$ is a (high-dimensional) column vector of firm-specific variables, z_t is a macro economic indicator relevant to the given industry and θ is an unknown parameter vector to be estimated.

The vector of firm-specific predictors, $X_{i,t-1}$, is dated $t-1$ to indicate a time lag between the dating of the predictors and the bankruptcy indicator B_{it}. $X_{i,t-1}$ includes a large number of standard *financial* indicators related to profitability, liquidity and financial structure; size-related variables; and indicators for payment remarks, C (see Section 2.2). We also include a one year lag of all the mentioned variables and interactions between categorical variables.

The macroeconomic indicator, z_t, is assumed to be observed in "real time" or, at least, published with a much shorter lag than X_{it} (for example, quarterly GDP comes with a two month publication lag and financial statements come with up to a nine month lag).[8] The estimation of the coefficient γ of z_t, is hampered by the short time series of the payment remarks variables (2010–2021) (see a related discussion in Jacobson et al. (2013)). In Appendix A.4 we propose a "measurement error model" as a practical remedy, where credit ratings from Dun & Bradstreet are included as auxiliary predictors instead of payment remarks *prior* to 2010. The proposed procedure enables us to obtain an estimate, $\widehat{\gamma}$, using *all* bankruptcy data, and then, by imposing the constraint $\gamma = \widehat{\gamma}$, estimate β using payment remarks (and other variables) available for 2010–2021.

3.2 Variable selection by Lasso and two-stage estimation

We now consider applying a two-stage estimator of θ. In the first stage, we estimate the model (1) using logistic Lasso regression on the combined data from all industries to select the most important predictors, that is, the components of $X_{i,t-1}$ with a non-zero coefficient estimate. To allow for industry heterogeneity, we include in $X_{i,t-1}$ industry dummies and a full set of interactions between industry dummies and explanatory variables. Thus, in principle, some economic predictors selected by Lasso could be specific to certain industries, whereas others could be common to all industries. In the second stage, we estimate a logit model separately for each industry, using the selected variables from the first stage as predictors ("post selection estimation").

Technically, Lasso minimizes a weighted average of the negative logit log-likelihood and a penalty term related to the sum of the absolute value of the coefficients (see Appendix A.3 for formulas). Because of the absolute value penalty, coefficient estimates may be exactly zero. This is the reason Lasso is referred to as a method of "feature" selection. The number of included predictors depends on the regularization parameter, λ, i.e., on how much weight is given to the Lasso penalty term. The higher λ, the fewer variables (features) are selected. The regularization parameter is chosen by first minimizing CVE. To obtain a sparse model, we apply the one SE rule. That is, we choose the simplest model (the highest λ) with a CVE which is no more than one standard error worse than the model with the lowest CVE (see Hastie et al. (2009), pp. 61

[8] We avoid discussing details about publication frequencies and publication lags here, as the main purpose of this paper is neither nowcasting nor real time forecasting.

and 244). We also consider the corresponding (weighted and unweighted) model *without* applying the one SE rule in the first stage. These models include more predictors and is henceforth refered to as "minimum CVE-based models".[9]

If we include all variables selected by Lasso in the second stage logit estimation, there is a risk of over-fitting. For example, interactions between categorical variables may identify narrow groups of observations with little or no variation in B_{it} in a given industry, in which case the estimation may even fail to converge. Therefore, as a refinement of the variable selection, we retain in the second stage only the subset of predictors selected by Lasso that minimizes the Bayesian information criterion (BIC) for the given industry. BIC is a standard criterion for model selection when you have a relatively small set of potential models. If the set of variables chosen by Lasso includes the true model (with probability one), BIC will asymptotically select the correct set of predictors.[10] We will henceforth use the notation x_{it} to refer to the (final) set of industry-specific features selected *after* applying BIC.

3.3 Debt weighting of observations in the estimation

Because the aim of this analysis is to predict the credit risk of bank loans to non-financial firms as defined in Equation (3), it is more important to accurately predict the bankruptcy probability of a firm with a high share of the bank debt than a firm with a low share. We will therefore consider modifying the two-stage estimator described above, by giving more weight to firms with high shares of the bank debt both when selecting predictors by Lasso (stage 1) and estimating θ (stage 2).

To address the issue of weighting formally, we start by noting that the variable of main interest for the prediction of $s_t(I)$ is: $w_{i,t-1}B_{it}$, the bankruptcy debt, with corresponding predictor $w_{i,t-1}p_{it}(\theta)$ and prediction error: $w_{i,t-1}(B_{it} - p_{it}(\theta))$. Since this prediction error is proportional to $w_{i,t-1}$, $w_{i,t-1}$ is a natural weight both in the first (variable selection) stage and in the second (post selection estimation) stage.

A more formal justification of this weighting scheme is that the asymptotic first order condition

[9] The CV method allows clustering of the observations by the same firm to take into account dependencies between intra-firm observations.

[10] Our two-stage approach mimics that of Tutz et al. (2015), who shows that feature selection by Lasso in a first stage, followed by maximum likelihood estimation ("re-fitting") in a second stage, improve the accuracy of estimators compared to one-stage logit-Lasso estimation.

for estimating θ in the logit model with $w_{i,t-1}$ as weights is:

$$E[(Y_{it} - w_{i,t-1}p_{it}(\theta))(x'_{i,t-1}, z_t)] = 0 \tag{3}$$

where $Y_{it} = w_{i,t-1}B_{it}$ is the bankruptcy debt (recall that $x_{i,t-1}$ is the subset of $X_{i,t-1}$ selected in the first stage).[11] Equation (3) says that the weighted logit-estimator of θ with B_{it} as dependent variable, can be equivalently seen as a generalized method of moments estimator of θ with Y_{it} as dependent variable, where the moment condition is that the predictors are orthogonal to the *bankruptcy debt* prediction error. In order for $\widehat{s}_t(I)$ to be unbiased, Equation (3) needs to be satisfied.

On the other hand, in the case without weighting, the asymptotic first order condition for estimating θ is:

$$E[(B_{it} - p_{it}(\theta))(x'_{i,t-1}, z_t)] = 0 \tag{4}$$

In Equation (4), the identifying condition is that the predictors are orthogonal to the *bankruptcy* prediction error, $B_{it} - p_{it}(\theta)$. However, since our purpose is to predict $s_t(I)$, Equation (3) still needs to be satisfied. That is, we must ensure that $w_{i,t-1}$ is uncorrelated with $B_{it} - p_{it}(\theta)$.[12]

The minimization of the Lasso penalty function is, in practice, done by iteratively replacing the negative log-likelihood expression with a weighted sum of squares (see Appendix A.3). With and without debt weighting, the individual sum of squares terms at the final Lasso estimate equal

$$\frac{w_{i,t-1}(B_{it} - p_{it}(\widehat{\theta}))^2}{p_{it}(\widehat{\theta})(1 - p_{it}(\widehat{\theta}))}$$

and

$$\frac{(B_{it} - p_{it}(\widehat{\theta}))^2}{p_{it}(\widehat{\theta})(1 - p_{it}(\widehat{\theta}))}$$

respectively. This shows that the squared residuals $(B_{it} - p_{it}(\widehat{\theta}))^2$ are penalized proportionally to the amount of debt, $D_{i,t-1}$, in the case of debt weighting. A drawback of debt weighting is that the estimation could be highly influenced by a few large outlier firms, depending on the

[11] Equation (3) follows from the minimization of the Lasso penalty function in Appendix A.3 (with general weight $\omega_{it} = w_{i,t-1}$ and regularization parameter $\lambda = 0$; see Equation (5)) in Appendix A.3.

[12] We explore the difference between the weighted and unweighted estimator in Appendix A.3, by giving explicit formulas in the special case with one predictor. The intuition from this exercise is that, in the case of debt weighting, each firm-year observation is "replicated" in the data file in proportion to the amount of debt it represents.

distribution of the weights. Another drawback is that debt-weighting will generally increase the standard error of $\widehat{\theta}$ because it reduces the effective sample size.[13] Thus, if the model is correctly specified, weighting may reduce the efficiency of the estimator considerably. On the other hand, if the model is *not* correctly specified, weighting could reduce the bias of the fitted model by ensuring that Equation (3) is satisfied. This is analogous to the way local linear regressions may reduce the bias of the fitted model if the correct model is not (globally) linear. In the next section, we compare results from both weighted and unweighted estimators, with a separate variable selection in each case.

4 Results

4.1 Lasso results

The choice of predictors to be included in the model is based on cross validation (CV), as explained in Section 3. The debt-weighted CV error (CVE) function is shown in Chart 4.1. The CVE minimizer is: $\lambda_{CV} = 0.0014$, which corresponds to 67 variables with non-zero coefficients. However, the optimal λ is barely identified, as the function is flat in a large interval around the minimum. The implication is that the selection of predictors is unstable. In contrast, the one SE rule only depends on the minimal value of the CVE function and its standard error, yielding an almost ten times higher λ ($\lambda_{SE} = 0.012$). At the point λ_{SE} there are only 10 non-zero coefficients and the CVE function is steep, indicating that the selection of variables will not be changed by small perturbations in the data.

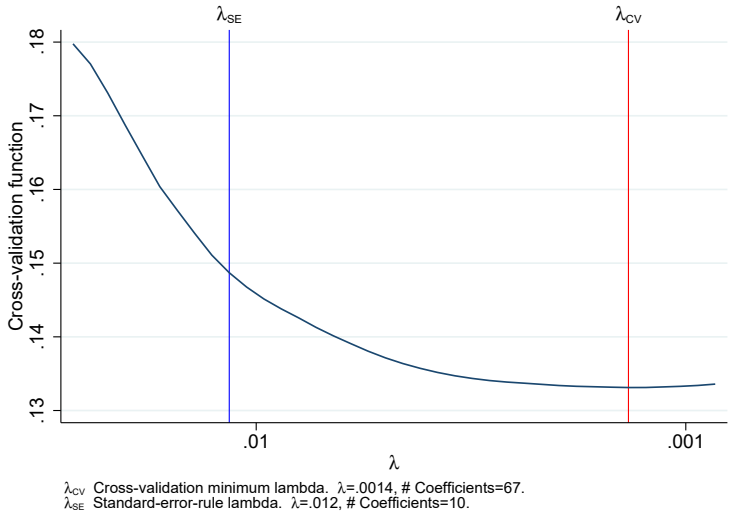
Unweighted Lasso gives quite similar results with regard to variable selection and exhibits the same problem with finding a minimizer as does weighted Lasso. The CVE minimizer is: $\lambda_{CV} = 0.0005$, with 159 non-zero coefficients – most of which are interactions between a (much smaller) set of economic variables and various dummy variables. The one SE rule yields $\lambda_{SE} = 0.009$ and 14 non-zero coefficients. Notably, regardless of weighting, size-related variables are not included among the selected predictors at λ_{SE}.

[13] The concept of effective sample size is derived by considering linear (weighted) estimators of a population mean, see Kish (1965). By applying this concept to $E(B_{it})$, the effective sample size associated with debt-weighting relative to the optimal (i.e., variance-minimizing) weights (see Appendix A.3), is

$$\frac{\left(\sum p_{it}(1-p_{it})\right)^2 \sum \frac{w_{i,t-1}^2}{p_{it}(1-p_{it})}}{\left(\sum w_{i,t-1} p_{it}(1-p_{it})\right)^2 \sum \frac{1}{p_{it}(1-p_{it})}}$$

Since large firms are associated with significantly lower bankruptcy risk and higher shares of the debt than small firms (see Hjelseth and Raknerud (2016)), the unweighted model is likely to seriously over-predict bankruptcy rates if we do not include any measure of firm size among the predictors. We therefore augment the set of predictors with a polynomial in Log real assets to obtain our *unweighted benchmark* model, as explained in Section 1. On the other hand, in the case of the weighted model the impact of firm size is captured by giving higher weight in the estimation to firms with high shares of the debt – not by including size-related control variables. This means that the weighted model might capture the bankruptcy risk of large firms better than of small firms.

Chart 4.1: Debt-weighted CVE function with λ chosen either as the CVE minimizer (λ_{CV}) or by applying the one SE rule (λ_{SE}).



The final list of predictors based on the one SE rule are shown in the first two columns of Table A.2. The included predictors are also shown in the Table 4.1, where a missing value means that the given variable is not selected in the given industry after applying Lasso in the first stage and then BIC in the second stage. In general, the automated feature selection with and without weighting are similar, although there are more variables included in the latter case. The small list of automatically selected predictors in Table 4.1 includes: Return on assets (RoA), Equity ratio (ER), Current liabilities ratio, a dummy for negative equity ($E < 0$), and the categorical variable Claims (C); see Table A.1 in Appendix A.2 for definitions. The last five rows of Table 4.1 refer to dummy variables (e.g., $C = 0$) or interactions between (products of) dummy variables. Predictors are dated either t or $t-1$ and the dependent variable, B, is dated

$t+1$ ($B_{i,t+1}$).

4.2 Marginal effects

Table 4.1 reports estimated average marginal effects (AME) with and without weighting for each industry. The method of weighting does not only apply to the estimation method, as explained in Section 3, but also to the averaging of the (firm-level) marginal effects. AME in the case of weighting expresses the expected change in percentage points (p.p.) in the bankruptcy debt rate for the given industry. Similarly, AME without weighting expresses the expected p.p. change in the industry's average bankruptcy frequency.

The marginal effect of a *categorical* variable is the estimated (weighted or unweighted) average change in bankruptcy probability in p.p. of the given category relative to the reference category: firms with debt collection ($C_t = 1$) and positive equity in at least one of the two last years ($E_t > 0$ or $E_{t-1} > 0$). When calculating AME, all *other* variables are evaluated at their actual value in the data. For variables that refer to, respectively, a rate and a log value, AME can be interpreted as the p.p. change in probability of, respectively, a 1 *p.p.* and 1 *percent* partial change in the variable. Reported z-scores are conditional on the first stage variable selection.

In Table 4.1 we first note that the macroeconomic indicators are highly significant in all industries, with p-values below 0.01, except for fishing and fish farming in the case of weighting, where the estimate is significant at the 10 percent level. The weighted estimates reported in Table 4.1 show that a 1 p.p. increase in GDP growth reduces the bankruptcy debt rate at the industry level by roughly 0.1–0.2 p.p. Moreover, a 1 p.p. increase in the prime yield rate increases the bankruptcy rate with 0.07 p.p. in CRE, whereas a 1 percent increase in the real salmon price decreases the bankruptcy probability by 0.004 p.p. in fishing and fish farming. Without weighting, the estimated AME of the macro indicator is substantially larger in absolute value than with weighting in all industries – sometimes more than two or three times larger.

Focusing next on the firm-specific variables in the case of weighting, we see from Table 4.1 that, in all industries, increased RoA_t significantly reduces bankruptcy debt rates. We find the strongest partial effect related to RoA_t in manufacturing, construction and retail trade. Across industries, lagged equity ratio (ER_{t-1}) and Current liability ratio are either much less significant predictors than RoA_t, or not selected at all. Regarding the categorical variables, we observe that

the dummy variables $E_t < 0$ and $E_{t-1} < 0$, the categorical variable Claims (C_t), and interactions between these variables are associated with highly significant AME in all industries. Firms with negative equity in two consecutive periods have significantly increased bankruptcy probability, and even more so when combined with payment remarks. For example, in CRE the estimated AME of the dummy variable interaction $E_t < 0 \cdot \ E_{t-1} < 0$ is 0.27 p.p. and the *additional* effect of having attachments ($C_t = 2$) is 0.18 p.p. (0.44 p.p. in total relative to the reference category). Compared to firms with zero payment remarks ($C_t = 0$) and positive equity ($E_t > 0$) , the estimated difference in AME is $0.44 - (-0.31) = 0.75$ p.p.

The most notable difference between the estimated weighted and unweighted AME is that the former are generally much larger in magnitude and more significant. The first finding is as expected, since average bankruptcy frequencies are generally higher than average bankruptcy debt rates (see Table 2.1). The second finding was also expected, since the effective sample size is much smaller when debt-weights are used in the estimation (see Footnote 13).

In the unweighted benchmark model, the *ex ante* included variable Log real assets is an important predictor. It enters the logit function both through a linear and quadratic term (we have not included higher order polynomial terms as these are insignificant in all the industries). AME for Log real assets is the sum (not displayed) of the AME for the linear and quadratic term reported in Table 4.1. The two terms therefore cannot be interpreted independently. We see that the effect of Log real assets is non-monotone: the typical pattern is that moderately large firms have a higher probability of bankruptcy than very small ones (positive linear AME term), but when the firm's asset size cross a certain threshold, the bankruptcy probability starts to decrease (negative quadratic AME-term). The first (positive) relation could reflect that creditors have more to gain from bankruptcy proceeding in the case of an asset-rich firm compared to a firm with little assets. The second (negative) relation is the dominant one according to numerous empirical studies about bankruptcy and firm liquidation; some examples are Mata et al. (1995), Olley and Pakes (1996) and Foster et al. (2008). It could reflect that larger firms have more financial muscle to withstand temporary economic setbacks, or to renegotiate debt conditions in times of crisis.

Table 4.1: Estimated average marginal effects (AME) and z-scores by industry.

	Fishing and fish farming				Manufacturing			
	With weighting		Without weighting		With weighting		Without weighting	
Dependent variable: $B(t+1)$	AME	z	AME	z	AME	z	AME	z
Macro indicator	-0.004^{*}	-1.73	-0.017^{**}	-2.35	-0.078^{***}	-3.21	-0.251^{***}	-9.32
Continuous firm variables:								
RoA_t	-0.004^{***}	-4.22			-0.028^{***}	-5.78	-0.045^{***}	-16.46
RoA_{t-1}	-0.002^{*}	-1.88	-0.014^{***}	-4.15				
ER_{t-1}					-0.025^{***}	-5.31	-0.017^{***}	-6.32
Current liabilities ratio_t							0.015^{***}	5.72
Current liabilities ratio_{t-1}								
Log real assets_t			0.007	1.15			0.015^{***}	4.23
Log real assets_t^2			-0.008	-1.40			-0.014^{***}	-4.35
Categorical firm variables:								
$\text{C}_t = 0$	-0.248^{***}	-2.51	-1.650^{***}	-6.22	-0.671^{**}	-2.41	-1.538^{***}	-7.38
$\text{C}_t = 2$							1.593^{***}	5.53
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0$	0.195^{**}	2.39	0.723^{***}	3.83			1.269^{***}	4.68
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0 \cdot \text{C}_t = 1$							0.816^{***}	2.83
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0 \cdot \text{C}_t = 2$	0.124^{**}	1.42						
AUROC	0.901		0.922		0.834		0.841	
Number of observations		10,034				46,106		

	Construction				Retail trade			
	With weighting		Without weighting		With weighting		Without weighting	
Dependent variable: $B(t+1)$	AME	z	AME	z	AME	z	AME	z
Macro indicator	-0.155^{***}	-3.60	-0.243^{***}	-7.90	-0.083^{***}	-3.83	-0.194^{***}	-8.68
Continuous firm variables:								
RoA_t	-0.036^{***}	-6.06	-0.051^{***}	-22.29	-0.031^{***}	-8.06	-0.061^{***}	-30.40
RoA_{t-1}			-0.012^{***}	-5.67	-0.014^{***}	-3.00	-0.016^{***}	-8.59
ER_{t-1}			-0.019^{***}	-6.65			-0.016^{***}	-7.57
Current liabilities ratio_t			0.013^{***}	4.39			0.016^{***}	8.03
Current liabilities ratio_{t-1}			0.009^{***}	3.10				
Log real assets_t			0.016^{***}	4.85			0.028^{***}	6.76
Log real assets_t^2			-0.016^{***}	-5.12			-0.032^{***}	-7.81
Categorical firm variables:								
$\text{C}_t = 0$	-1.644^{***}	-4.36	-3.303^{***}	-16.53	-0.863^{***}	-2.90	-2.909^{***}	-24.33
$\text{C}_t = 2$			2.669^{***}	12.07			1.883^{***}	10.51
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0$	2.392^{***}	8.61	2.618^{***}	9.50	2.163^{***}	10.48	2.676^{***}	15.69
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0 \cdot \text{C}_t = 1$			-0.889^{***}	-3.20				
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0 \cdot \text{C}_t = 2$	1.472^{***}	3.68	-1.880^{***}	-5.83	0.803^{***}	3.53		
AUROC	0.850		0.858		0.841		0.850	
Number of observations		79,935				122,221		

	Commercial real estate				Services			
	With weighting		Without weighting		With weighting		Without weighting	
Dependent variable: $B(t+1)$	AME	z	AME	z	AME	z	AME	z
Macro indicator	0.070^{***}	8.20	0.093^{***}	11.99	-0.024^{***}	-3.28	-0.116^{***}	-7.86
Continuous firm variables:								
RoA_t	-0.007^{***}	-8.47	-0.009^{***}	-11.70	-0.006^{***}	-3.22	-0.029^{***}	-26.55
RoA_{t-1}							-0.006^{***}	-5.50
ER_{t-1}	-0.003^{***}	-5.33	-0.005^{***}	-8.87				
Current liabilities ratio_t			0.004^{***}	9.04	0.005^{***}	3.39	0.002	1.13
Current liabilities ratio_{t-1}							0.007^{***}	4.83
Log real assets_t			0.002^{***}	2.82			0.005^{***}	3.09
Log real assets_t^2			-0.003^{***}	-3.38			-0.007^{***}	-4.49
Categorical firm variables:								
$\text{C}_t = 0$	-0.314^{***}	-6.51	-0.898^{***}	-16.60	-0.353^{***}	-4.28	-2.554^{***}	-19.93
$\text{C}_t = 2$			0.533^{***}	8.80			2.186^{***}	14.16
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0$	0.266^{***}	5.68	0.528^{***}	10.46	0.583^{***}	5.58	2.739^{***}	21.86
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0 \cdot \text{C}_t = 1$			-0.213^{***}	-3.42			-0.517^{***}	-3.33
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0 \cdot \text{C}_t = 2$	0.178^{***}	2.92	-0.441^{***}	-5.95	0.238^{***}	3.56	-1.484^{***}	-7.71
AUROC	0.873		0.885		0.849		0.860	
Number of observations		261,189				164,867		

Notes: Estimation period: 2010–2019 (accounting years). The asterisks indicate significance levels at: * p<0.1, ** p<0.05 and *** p<0.01. Macro indicator for manufacturing, construction, retail trade and services: Real GDP growth rate. Fishing and fish farming: log of real salmon prices. Commercial real estate: real prime yield for office spaces in Oslo. The Average Marginal Effect (AME) is the estimated probability of bankruptcy for each observation in p.p. of a unit change in the explanatory variables. Categorical variables: change in probability in p.p. relative to the reference category. Rate variables: change in probability in p.p of a 1 p.p. partial change in the variable. Logarithmic scale variables: p.p. change in probability of a 1 percent partial change in the variable. AME for *log real assets* is the sum (not displayed) of the AME for the linear and quadratic term. AUROC is the area under the ROC curve.

4.3 The importance of categorical variables

All possible combinations for the values of the five categorical variables listed in Table 4.1 constitute six non-overlapping categories of observations, as listed in the first column of Table 4.2. The next two columns in Table 4.2 show share of total bank debt and actual bankruptcy debt rate for the corresponding category of observations aggregated over all years and industries. The highest average actual bankruptcy debt rate, 15.7 percent, are found in the category with $E_t < 0$, $E_{t-1} < 0$ and $C_t = 2$. Next comes the category with $E_t < 0$, $E_{t-1} < 0$ and $C_t = 1$, with a bankruptcy debt rate of 5.8 percent. Third comes the category with $C_t = 2$ and positive equity in at least one of the last two years, with a bankruptcy debt rate of 1.6 percent.

The share of bank debt held by the different categories are ranked in the opposite order of the corresponding bankruptcy debt rates. In particular, the debt shares in the three categories of firms with highest bankruptcy debt rates are very small: 0.2, 1.0 and 1.1 percent, respectively. In contrast, the category with no payment remarks ($C_t = 0$) and positive equity in t or $t-1$ (or both), have average bankruptcy debt rate of 0.1 percent and the highest debt share: 73.1 percent. Firms with debt collection but positive equity in t or $t-1$ hold 19.5 percent of the debt and have average bankruptcy debt rate of 0.6 percent. The differences in bankruptcy debt rates across some of the categories in Table 4.2 are much larger than explained by the AME of the categorical variables reported in Table 4.1. These discrepancies are due to the contribution of the other continuous variables, such as RoA, which, of course, have different distributions in different categories.

Table 4.2: Debt shares, and actual and predicted bankrupcty debt rates by combinations of categorical variables. Average over 2011–2020. Percent.

			Predicted bankruptcy debt rates			
			Out-of-sample		In-sample	
Categorical variables	Share of total bank debt	Actual bankruptcy debt rate	With weighting	Without weighting	With weighting	Without weighting
$C_t = 0$ and ($E_t > 0$ and/or $E_{t-1} > 0$)	73.1	0.10	0.10	0.11	0.10	0.11
$C_t = 1$ and ($E_t > 0$ and/or $E_{t-1} > 0$	19.5	0.58	0.61	0.49	0.64	0.46
$C_t = 0$ and $E_t < 0$ and $E_{t-1} < 0$	5.2	1.01	1.14	0.89	1.15	0.85
$C_t = 2$ and ($E_t > 0$ and/or $E_{t-1} > 0$)	1.1	1.64	0.83	3.30	0.66	2.16
$C_t = 1$ and $E_t < 0$ and $E_{t-1} < 0$	1.0	5.78	6.71	5.88	6.60	5.14
$C_t = 2$ and $E_t < 0$ and $E_{t-1} < 0$	0.2	15.73	15.33	12.13	15.60	10.63

4.4 Out-of-sample predictions

So far, our results refer to estimates and predictions where we have used all available data. However, good in-sample performance is no guarantee of good predictive properties. If, for

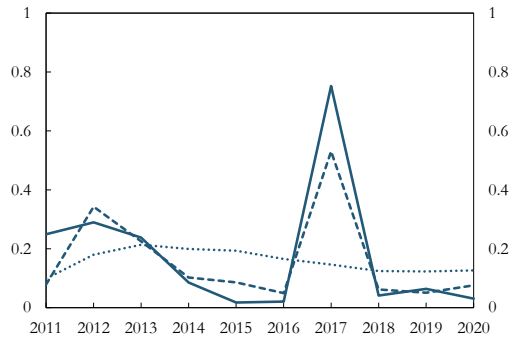
example, the estimated relation between the macroeconomic indicator and the dependent variable is unstable, the model may quickly break down when additional years of data are included in the analysis.

In order to examine genuine predictive properties, we show in Chart 4.2, for each industry, actual bankruptcy debt rates versus out-of-sample predictions obtained using the estimated logit models, with and without weighting, for 2011–2020. The predictions for any year t are constructed by excluding all observations dated t when estimating the model from which the predictions are derived. Moreover, we split the firms randomly in two sub-samples of equal size, say A_{-t} and B_{-t} *excluding* year t (i.e., excluding bankruptcies dated t from both sub-samples). Then, to predict bankruptcy probabilities in year t for the firms in sub-sample A_{-t}, the model is estimated on the sub-sample B_{-t}, and vice versa. In contrast, in-sample predictions use the same data twice: first for estimation and then for prediction, which potentially may lead to over-fitting. Our sample splitting procedure mirrors the sample splitting used in CV. It is repeated for every year in 2011–2020 to generate out-of-sample predictions.

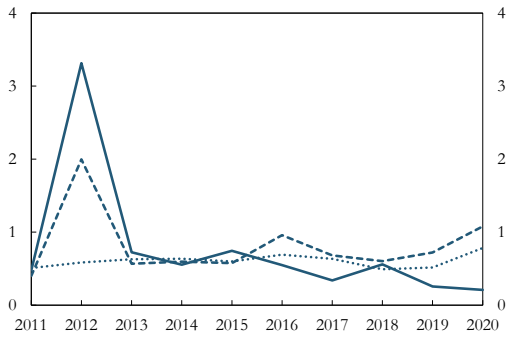
From the graphs in Chart 4.2, we see that in some industries the weighted model yields predictions of bankruptcy debt rates that are visibly better than the predictions based on the unweighted estimates (with generally different predictors), in other industries neither method appear to yield particularly good predictions, whereas in services the weighted model is visibly much worse than the unweighted benchmark model in 2020. Improvements achieved by weighting are clearly seen in fishing and fish farming, manufacturing and CRE. It is notable that the predictions based on weighting use fewer predictors, which is *cet. par.* beneficial when pure forecasting is not the only purpose of the model.

Chart 4.3 shows the corresponding graphs for all industries aggregated, i.e., obtained as the debt-weighted average across industries of the (actual and predicted) series in Chart 4.2. The aggregate actual bankruptcy debt rate exhibits a slight downward trend during 2011–2018, which is closely tracked by the predicted series using weights in the estimation. In contrast, the predictions of the unweighted benchmark model are almost flat. The year 2020 is exceptional because of the extraordinary policy measures that took effect due to the pandemic (see Section 2.2). The predicted bankruptcy debt rate obtained using weights is 0.15 p.p. higher than the actual one in

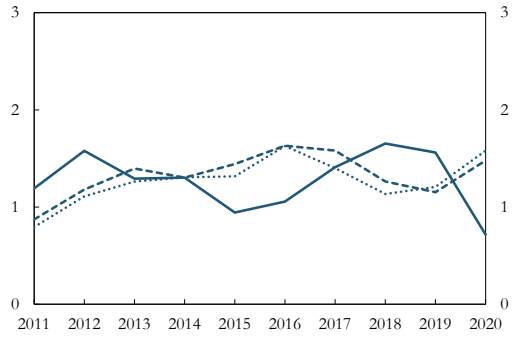
Chart 4.2: Bankruptcy debt rates by industry. Actual rates and out-of-sample predictions with and without weighting. Predictors selected by applying the one SE rule in the first stage. 2011–2020. Percent.



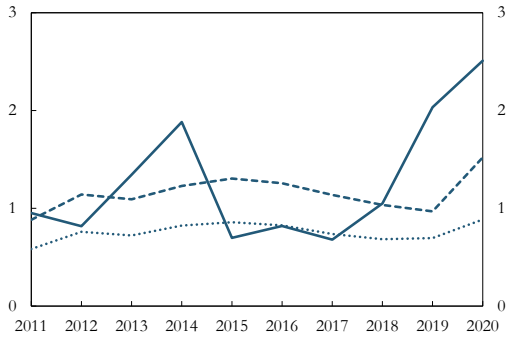
(a) Fishing and fish farming



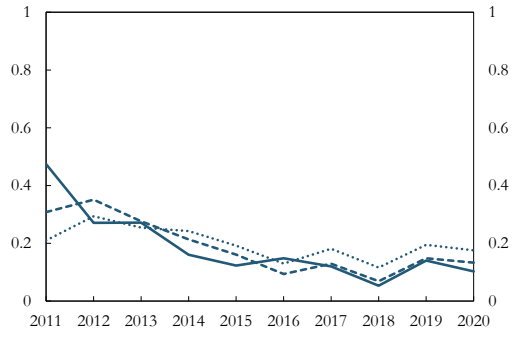
(b) Manufacturing



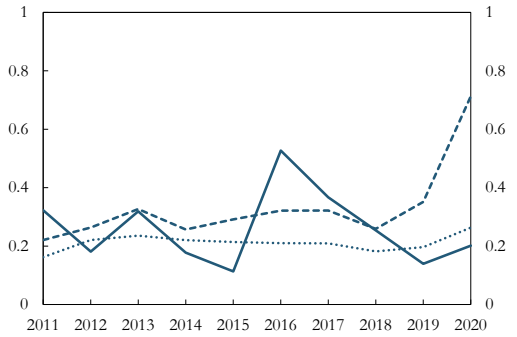
(c) Construction



(d) Retail trade



(e) Commercial Real estate



(f) Services

Actual share of bankruptcy debt

Predicted share of bakruptcy debt: with weights

Predicted share of bankruptcy debt: without weights

2020, whereas the prediction from the unweighted benchmark model is 0.07 p.p. higher.[14]

The main reason that the predictions from the unweighted benchmark model is so stable over time, is that these are dominated by Log real assets for large firms, whose bankruptcy probabilities are very small regardless of large changes in economic fundamentals. In Chart 4.3 we have included a graph showing aggregate predictions after removing the asset size variables from the unweighted benchmark model (i.e., only including the macro economic indicator and the automated selected variables based on the one SE rule). This graph exhibits much larger fluctuations over time than the unweighted benchmark model. Even more strikingly, the graph uniformly over-predicts bankruptcy debt rates by a wide margin. As anticipated in Section 4.1, large firms have lower bankruptcy probabilities and more debt than small firms and this relationship is not picked up without size-related control variables, leading to a serious prediction bias. The graph corresponding to the weighted model indicates that the bias can be remedied by debt-weighting as an alternative to including size-related control variables.

Chart 4.3: Aggregate bankruptcy debt rate for the six industries. Actual rate and out-of-sample predictions with and without weighting. Predictors selected by applying the one SE rule in the first stage. 2011–2020. Percent.

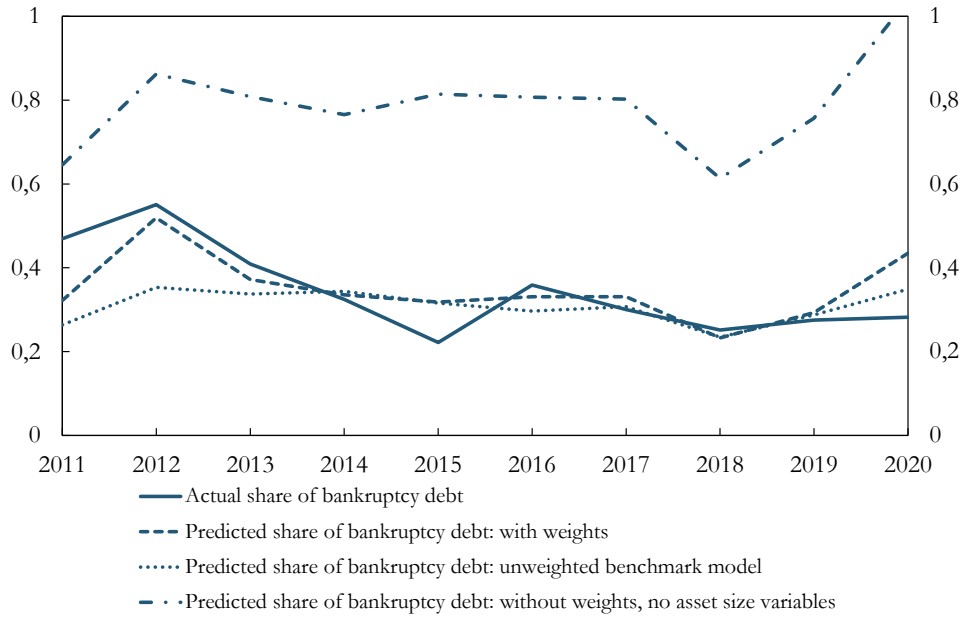


Table 4.3 summarizes the predictive properties of the weighted model and the unweighted benchmark model by displaying root mean squared prediction error (RMSE) in-sample and

[14]Remember that the predictions of the bankruptcy debt rates for 2020 are based on financial statements and payment remarks from 2019, i.e., before the pandemic, and macroeconomic indicators from 2020.

out-of-sample for 2011–2020. Because 2020 is an extreme outlier, we also show in the last row aggregate RMSE for all industries for 2011–2019, i.e. excluding 2020. The results in Table 4.3 confirm the conclusions we draw from Charts 4.2 and 4.3. For all industries (i.e., debt-weighted average statistics across industries), RMSE with and without weighting are: 0.08 vs. 0.10 p.p. out-of-sample, and 0.10 vs. 0.09 p.p. in-sample. If we exclude 2020, out-of-sample RMSE with and without weighting are 0.06 and 0.11 p.p., respectively. The 1–2 p.p. increase in average out-of-sample RMSE when 2020 is included, can be attributed to one industry, services, where the weighted model over-predicts by 0.33 p.p. in 2020, whereas the unweighted model over-predicts by 0.15 p.p. Also if we compare the arithmetic means across industries (2011–2020), the model with weighting performs better than the model without weighting, with out-of-sample RMSE of 0.31 p.p. and 0.43 p.p., respectively. Furthermore, when we look at the individual industries, the out-of-sample RMSE is always lower in the model with weighting than in the unweighted benchmark model, except in services where it is slightly higher – again this is caused by the 2020-outlier.

Table 4.3 also shows correlation coefficients between actual and predicted bankruptcy debt rates. The arithmetic average and aggregated correlation coefficients, with and without weights, vary wildely from -0.02 to 0.80, but out-of-sample they are much higher for the weighted than the unweighted model. These results indicate that the model with weighting is better at capturing changes in the bankruptcy debt rate over time than the unweighted benchmark model, whose predictions tend to be dominated by the impact of the asset size variables.

Table 4.3: Root mean squared prediction error (RMSE) at the industry level and correlations (Corr.) between the actual and predicted bankruptcy debt rate. Predictors selected by applying the one SE rule in the first stage. Average over 2011–2020.

	Out of sample				In sample			
	With weighting		Without weighting		With weighting		Without weighting	
Industry	RMSE	Corr.	RMSE	Corr.	RMSE	Corr.	RMSE	Corr.
Fishing and fish farming	0.09	0.93	0.22	-0.03	0.22	0.26	0.21	0.34
Manufacturing	0.55	0.84	0.89	-0.13	0.86	0.23	0.89	0.00
Construction	0.42	-0.38	0.45	-0.50	0.36	-0.24	0.40	-0.45
Retail trade	0.59	0.33	0.79	0.27	0.57	0.55	0.78	0.42
CRE	0.07	0.83	0.10	0.54	0.05	0.91	0.08	0.77
Services	0.16	-0.11	0.14	-0.28	0.16	0.01	0.12	0.09
Arithmetic mean	0.31	0.41	0.43	-0.02	0.37	0.29	0.41	0.20
All industries aggregated	0.08	0.64	0.10	0.26	0.10	0.26	0.09	0.52
All industries aggregated excl. 2020	0.06	0.80	0.11	0.37	0.11	0.37	0.10	0.63

Finally, we comment on the last four columns of Table 4.2, which show the in-sample and out-of-sample average predicted bankruptcy debt rates across six categories of firms when θ is

estimated with or without weighting. The predicted bankruptcy debt rates are generally similar to the actual ones, confirming Table 4.3. In general, predictions are better with than without weighting, especially out-of-sample. The improvement obtained by weighting is most notable in the category with the highest bankruptcy risk (last row of Table 4.2) .

4.5 Minimum CVE-based variable selection

The results presented so far refer to variable selection by means of the one SE rule in the first stage and application of the BIC at the industry level in the second stage. The one SE rule is motivated by the desire to achieve a parsimonious representation – possibly at a cost in terms of forecasting performance. To shed light on this issue, Table 4.4 presents results for weighted and unweighted models when the minimum CVE criterion is used instead of the one SE rule in the first stage. The complete list of variables included in each industry for the two different variable selection criteria is shown in Table A.2 in Appendix A.5. The final variable selection based on the minimum CVE criterion includes a polynomial in Log real assets in most industries, both for models with or without weighting, in addition to many more interactions between dummy variables compared to the models based on the one SE rule. The weighted model based on the one SE rule contains at most three continuous firm-specific predictors and three categorical predictors (across industries). The unweighted model based on the minimum CVE criterion contain at most seven continuous firm-specific predictors and 14 categorical predictors.

For all industries aggregated, out-of-sample RMSE in Table 4.4 is 0.08 p.p. both with and without weighting (0.09 p.p. and 0.08 p.p. respectively, when excluding 2020). Weighting seems in general to have little impact on performance. However, comparing the results in Table 4.3 and 4.4 by industry, we see that the out-of-sample RMSE is usually larger when variable selection is based on the minimum CVE criterion compared to the one SE rule with weighting (which never includes Log real assets as a predictor). The relatively good performance of the highly sparse weighted model is particularly striking in fishing and fish farming and manufacturing. The fit of the out-of-sample predictions is further displayed in Chart A.2 and A.3 in Appendix A.5 (which correspond to Chart 4.2 and 4.3). The main conclusion from these charts is that with minimum CVE-based variable selection, the time series pattern of the bankruptcy debt series are dominated by the asset size variables for large firms, which contributes to a considerable smoothing of predicted bankruptcy debt rates over time, making them relatively immune to

Table 4.4: Root mean squared prediction error (RMSE) at the industry level and correlations (Corr.) between the actual and predicted bankruptcy debt rate. Predictors selected by applying minimum CVE in the first stage. Average over 2011–2020.

	Out of sample				In sample			
	With weighting		Without weighting		With weighting		Without weighting	
Industry	RMSE	Corr.	RMSE	Corr.	RMSE	Corr.	RMSE	Corr.
Fishing and fish farming	0.23	0.11	0.28	-0.35	0.22	0.25	0.25	0.00
Manufacturing	0.90	-0.09	0.87	0.14	0.88	0.08	0.87	0.18
Construction	0.45	-0.44	0.46	-0.56	0.44	-0.53	0.47	-0.70
Retail trade	0.65	0.26	0.70	0.00	0.64	0.31	0.69	0.06
CRE	0.06	0.90	0.10	0.61	0.03	0.96	0.08	0.80
Services	0.13	-0.13	0.13	-0.22	0.12	0.27	0.12	0.08
Arithmetic mean	0.40	0.10	0.42	-0.06	0.39	0.22	0.41	0.07
All industries aggregated	0.08	0.68	0.08	0.56	0.08	0.76	0.07	0.75
All industries aggregated excl. 2020	0.09	0.85	0.08	0.58	0.08	0.87	0.07	0.77

even large changes in economic fundamentals.

5 Conclusion

Predicting the credit risk of banks' corporate lending is important from a financial stability perspective. In this paper, we have proposed a model framework for predicting the share of bank debt held by bankrupt firms at the industry level. In addition to macroeconomic indicators and a huge set of potential standard firm-level variables from the financial statements, a key feature of our model is the inclusion of records of payment remarks, including debt collections from private agencies and attachments from private and public agencies.

Our model framework consists of a two-stage procedure, where we first use logistic Lasso for variable selection, followed by logistic regressions for (post-selection) estimation and prediction. We have compared models with and without weighting of observations, where the weights are equal to each firm's share of its industry bank debt.

We have shown that weighting has a large impact on both variable selection and estimation, and hence on the predictions of bankruptcy debt rates. Weighted Lasso combined with a highly sparse variable selection criterion ("the one SE rule") leads to a smaller number of variables being selected compared to alternative methods, such as the minimum CV error criterion, and – more importantly – to generally lower out-of-sample prediction errors at the industry level.

References

Altman, E. I. (1968): "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *The Journal of Finance*, 23, 589–609.

Barndorff-Nielsen, O. and D. Cox (1994): *Inference and Asymptotics*, Monographs on Statistics and Applied Probability, Chapman and Hall.

Beaver, W. H. (1966): "Financial Ratios As Predictors of Failure," *Journal of Accounting Research*, 4, 71–111.

Bernhardsen, E. (2001): "A model of bankruptcy prediction," Working Paper 10, Norges Bank.

Bernhardsen, E. and K. Larsen (2007): "Modelling credit risk in the enterprise sector - further development of the SEBRA model." Economic Bulletin 2, Norges Bank.

Carling, K., T. Jacobson, J. Lindé, and K. Roszbach (2007): "Corporate credit risk modeling and the macroeconomy," *Journal of Banking & Finance*, 31, 845–868.

Chen, Y. and Y. Yang (2021): "The One Standard Error Rule for Model Selection: Does It Work?" *Stats*, 4, 868–892.

Christoffersen, B., R. Matin, and P. Mølgaard (2018): "Can Machine Learning Models Capture Correlations in Corporate Distresses?" Working Paper 128, Danmarks Nationalbank.

Foster, L., J. Haltiwanger, and C. Syverson (2008): "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *The American Economic Review*, 98, 394–425.

Giordani, P., T. Jacobson, E. von Schedvin, and M. Villani (2014): "Taking the Twists into Account: Predicting Firm Bankruptcy Risk with Splines of Financial Ratios," *The Journal of Financial and Quantitative Analysis*, 49, 1071–1099.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer series in statistics, Springer.

Hastie, T., R. Tibshirani, and M. Wainwright (2015): *Statistical Learning with Sparsity: The Lasso and Generalizations*.

Hjelseth, I. N. and A. Raknerud (2016): "A Model of Credit Risk in the Corporate Sector Based on Bankruptcy Prediction," Staff Memo 20, Norges Bank.

Jacobson, T., K. Roszbach, and J. Lindé (2013): "Firm default and aggregate fluctuations," *Journal of the European Economic Association*, 11, 945–972.

Jacobson, T. and E. von Schedvin (2015): "Trade Credit And The Propagation Of Corporate Failure: An Empirical Analysis," *Econometrica*, 83, 1315–1371.

Jones, S., D. J. Johnstone, and R. Wilson (2017): "Predicting Corporate Bankruptcy: An Evaluation of Alternative Statistical Frameworks," *S&P Global Market Intelligence Research Paper Series*.

Kim, M.-J. and D.-K. Kang (2010): "Ensemble with neural networks for bankruptcy prediction," *Expert Systems with Applications*, 37, 3373–3379.

Kish, L. (1965): *Survey sampling*, New York: John Wiley & Sons, Inc.

Kragh-Sørensen, K. and H. Solheim (2014): "What do banks lose money on during crises?" Staff Memo 3, Norges Bank.

Li, H. and J. Sun (2013): "Predicting Business Failure Using an RSF-based Case-Based Reasoning Ensemble Forecasting Method," *Journal of Forecasting*, 32, 180–192.

Mata, J., P. Portugal, and P. Guimarães (1995): "The survival of new plants: Start-up conditions and post-entry evolution," *International Journal of Industrial Organization*, 13, 459–481, the Post-Entry Performance of Firms.

Matin, R., C. Hansen, C. Hansen, and P. Mølgaard (2019): "Predicting distresses using deep learning of text segments in annual reports," *Expert Systems with Applications*, 132, 199–208.

Merton, R. C. (1974): "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," *The Journal of Finance*, 29, 449–470.

Min, J. H. and Y.-C. Lee (2005): "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," *Expert Systems with Applications*, 28, 603–614.

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

Oslo Tingrett (2021): "Årsmelding 2020 konkurs," https://www.domstol.no/enkelt-domstol/oslo-tingrett/om-domstolen/publikasjoner/arsmeldinger2/arsmelding-2020/konkurs/ Accessed: 2022-03-15.

Shumway, T. (2001): "Forecasting Bankruptcy More Accurately: A Simple Hazard Model," *The Journal of Business*, 74, 101–124.

Tutz, G., W. Pössnecker, and L. Uhlmann (2015): "Variable selection in general multinomial logit models," *Computational Statistics & Data Analysis*, 82, 207–222.

Wilcox, J. W. (1973): "A Prediction of Business Failure Using Accounting Data," *Journal of Accounting Research*, 11, 163–179.

A Appendix

A.1 Classification tree

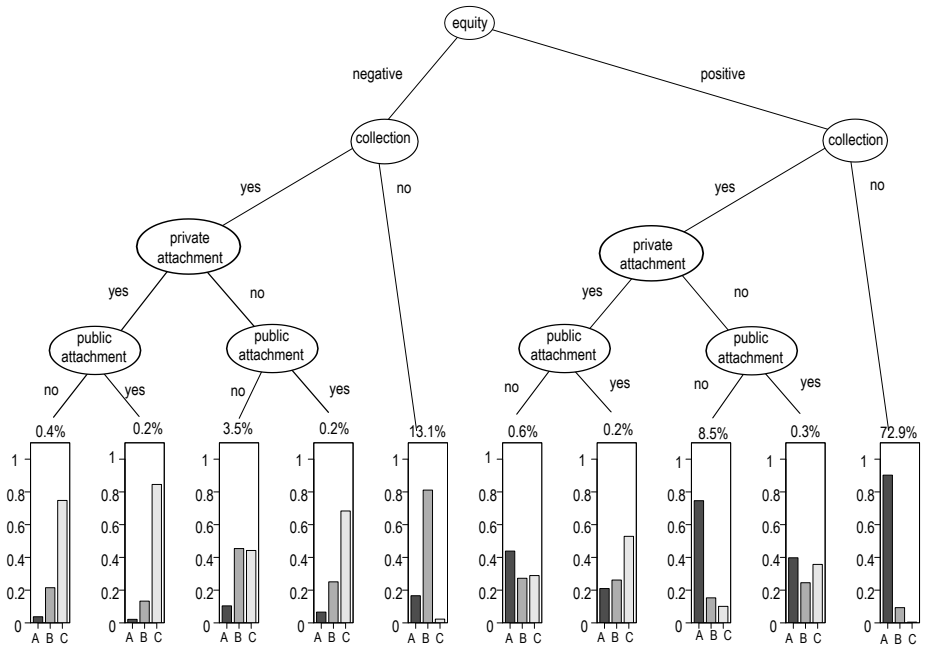
Chart A.1 depicts a classification tree for the relationship between the probability of the rating categories C, B and A (A is the residual category[15]) and dummy variables for the three types of payment remarks and negative equity.[16]

The depicted classification tree shows that the first split is between firms with positive or negative equity. The second is between firms with or without debt collections. The third split is between firms with or without private attachments, and the fourth is between firms with or without public attachments. Firms with positive equity and no debt collection (73 percent of all observations) are rated A with 90 percent probability. Firms with positive equity is rated A with 75 percent probability if they have debt collections but no attachments. This reflects that debt collections are not unusual and, in isolation, is a weak signal of financial distress. We see from the figure that about 17 percent of the firms have negative equity. These firms are most likely to be rated B if they have no debt collections (80 percent probability). If they have debt collections, but no attachments they are about equally likely to be rated B or C (45 percent probability). They are most likely rated C if they have debt collections and only one type of attachment (75 percent if private and 65 percent if public). If they have all three types of payment remarks, they are rated C with 85 percent probability. We conclude that by using four dichotomous covariates, the three rating categories can be predicted correctly with a high probability.

[15] Category A contains everything else than rating B or C, i.e., rating A, AA, AAA, AN ("new firm") and non-rated firms. C is the lowest rating and AAA is the highest.

[16] Having negative equity or not is also an important factor in determining rating in Dun & Bradstreet's credit rating model.

Chart A.1: Classification tree of credit rating



A.2 Variables included in LASSO

Table A.1: List of variables X_{it} included in the Lasso. All variables in the list are (in addition) interacted with industry dummies.

Variables	Description
Continuous variables from financial statements data, dated both year t and $t-1$	
Return on assets (RoA)	(Results bf. xo items and taxes + int. exp.)/total assets
Return on equity	Results bf. xo items and taxes/equity
Equity ratio (ER)	Equity/total assets (opening balance)
Current ratio	Current assets/current liabilities
Quick ratio	(Current assets - inventories)/current liabilities
Cash-to-revenue ratio	Cash and cash equivalents/revenue
Working capital-to-revenue ratio	(Current assets - current liabilities)/revenue
Current liabilities ratio	Current liabilities/total assets
Interest coverage ratio (ICR)	EBIT/net interest expenses
Debt-servicing ratio 1	EBITDA/bank debt
Debt-servicing ratio 2	Results bf. xo items and taxes/bank debt
Debt-servicing ratio 3	(Results bf. xo items and taxes+depr. and amort.)/bank debt
Debt-servicing ratio 4	EBITDA/interest-bearing debt
Debt-servicing ratio 5	Results bf. xo items and taxes/interest-bearing debt
Debt-servicing ratio 6	(Results bf. xo items and taxes+depr. and amort.)/interest-bearing debt
EBITDA margin	EBITDA/revenue
Revenue ratio 1	Revenue/total assets
Revenue ratio 2	Revenue/total liabilities
Revenue ratio 3	Revenue/bank debt
Revenue ratio 4	Revenue/interest-bearing debt
Log real assets	ln(total assets/GDP deflator)
Log real assets^2	
Categorical variables from financial statements data, dated both year t and $t-1$	
Dummy for pos. equity ($\delta(E)$)	$\delta(E)=1$ if $E \geq 0$: positive equity $\delta(E)=0$ if $E<0$: negative equity
Categorical variables from payment remarks data, dated both year t and $t-1$	
Claims (C)	$C=0$: no payment remarks, $C=1$: debt collection, no attachments, $C=2$: debt collection and public and/or private attachment.
Other categorical variables, dated year t	
Firm age group	1: firm is 3 years or younger, 2: firm is between 4 to 9 years old, 3: firm is 10 years or older.
Dummy variable interaction terms. For $n, m \in \{0,1\}$ and $j, k \in \{0,1,2\}$:	
$\delta(E_t) = n \cdot \delta(E_{t-1}) = m$	
$C_t = j \cdot C_{t-1} = k$	
$\delta(E_t) = n \cdot C_t = j$	
$\delta(E_t) = n \cdot C_{t-1} = k$	
$\delta(E_{t-1}) = m \cdot C_t = j$	
$\delta(E_{t-1}) = m \cdot C_{t-1} = k$	
$\delta(E_t) = n \cdot \delta(E_{t-1}) = m \cdot C_t = j$	
$\delta(E_t) = n \cdot \delta(E_{t-1}) = m \cdot C_{t-1} = k$	
$\delta(E_t) = n \cdot C_t = j \cdot C_{t-1} = k$	
$\delta(E_{t-1}) = m \cdot C_t = j \cdot C_{t-1} = k$	
$\delta(E_t) = n \cdot \delta(E_{t-1}) = m \cdot C_t = j \cdot C_{t-1} = k$	

A.3 Lasso with weights

In the presence of a huge number of potential predictors, we select the most relevant features using logit-Lasso, with general weights ω_{it}; where either $\omega_{it} \equiv 1$ (unweighted Lasso) or $\omega_{it} = w_{i,t-1}$ (debt-weighted Lasso):

$$(\widetilde{\beta}, \widetilde{\gamma}) = \arg\min_{(\beta,\gamma)} \sum_{(i,t)} \omega_{it}[\ln(1 + \exp(\beta X_{i,t-1} + \gamma z_t)) - B_{it}(\beta X_{i,t-1} + \gamma z_t)] + \lambda||\beta||_1 \tag{5}$$

The summation represents the negative weighted logit log-likelihood and the last term is the penalty related to the absolute value of the β-coefficients. Moreover, λ is the regularization parameter and $||\cdot||_1$ refer to the L_1-norm (see Hastie et al. (2009) for details). The parameter λ is chosen by first minimizing the cross validation (CV) prediction error with respect to λ and then apply the one SE rule (see Hastie et al. (2009), pp. 61 and 244). In practice, the minimization is solved by iteratively replacing the terms inside the summation with the the quadratic approximation of Equation (6) – trivially generalized to include more than one predictor.

To understand the difference between the weighted and unweighted estimator, we examine the logit model in the case of a single predictor, x_{it}:

$$\ln(\frac{p_{it}(\theta)}{1 - p_{it}(\theta)}) = \beta_0 + \beta_1 x_{it}$$

As shown by Hastie et al. (2009) (Section 9, Algorithm 9.2), the logit estimator can be found by iteratively minimizing a weighted sum of squares:

$$\min_{\beta_0,\beta_1} \sum_{it} (\widehat{u}_{it} - \beta_0 - \beta_1 x_{it})^2 \widehat{\omega}_{it} \tag{6}$$

where

$$\widehat{u}_{it} = \frac{B_{it} - p_{it}(\widehat{\theta})}{p_{it}(\widehat{\theta})(1 - p_{it}(\widehat{\theta}))} + \widehat{\beta}_0 + \widehat{\beta}_1 x_{it}$$

are auxiliary dependent variables that depend on the current best parameter estimate $\widehat{\theta} = (\widehat{\beta}_0, \widehat{\beta}_1)$ and $\widehat{\omega}_{it}$ are parameter dependent weights which are proportional to the conditional variance of B_{it}:

$$\widehat{\omega}_{it} = c \cdot p_{it}(\widehat{\theta})(1 - p_{it}(\widehat{\theta})) \tag{7}$$

where the constant c is chosen such that the $\widehat{\omega}_{it}$ sum to one. The logit estimator is the solution for which the minimizer in (6) equals the current best estimate, $\widehat{\theta}$. By explicitly solving (6) we obtain:

$$\widehat{\beta}_1 = \frac{\sum \widehat{\omega}_{it}(\widehat{u}_{it} - \sum \widehat{u}_{it}\widehat{\omega}_{it})(x_{it} - \sum x_{it}\widehat{\omega}_{it})}{\sum \widehat{\omega}_{it}(x_{it} - \sum x_{it}\widehat{\omega}_{it})^2}$$

and

$$\widehat{\beta}_0 = \sum \widehat{\omega}_{it}\widehat{u}_{it} - \widehat{\beta}_1 \sum \widehat{\omega}_{it}x_{it}$$

The effect of debt-weighting is to replace the original weight with:

$$\widehat{\omega}_{it} = \tilde{c} \cdot \; w_{i,t-1}p_{it}(\widehat{\theta})(1 - p_{it}(\widehat{\theta})) \tag{8}$$

for a constant, $\tilde{c}$, so that the sum of the weights is one.

A.4 Method for obtaining a pooled estimator of γ

The main steps are as follows. Using the data for $t = 2011, ..., 2020$ (henceforth main period), we fit two equations. First, we estimate b and γ in:

$$\ln(\frac{p_{it}(\theta)}{1 - p_{it}(\theta)}) = bx_{i,t-1} + \gamma z_{it} \tag{9}$$

Then, given estimated coefficients, b, we estimate c in:

$$bx_{it} = c\widetilde{x}_{it} + u_i + e_{it} \tag{10}$$

where $\widetilde{x}_{it}$ contains the same variables as x_{it} except that $(E < 0)\cdot(C = j)$ is replaced by indicators of rating categories. The combined error term, $u_i + e_{it}$, represents the measurement error from approximating bx_{it} by $c\widetilde{x}_{it}$, where u_i is the firm-specific and e_{it} the idiosyncratic measurement error component.

Given estimates of the error component variances σ_u^2, σ_e^2 in the random effects equation (10) with bx_{it} as the dependent variable and data $(d_1, X_1, \widetilde{X}_1)$, where $d_1 = \{D_{it}\}_{i,t\geq 2011}$, $X_1 = \{x_{i,t-1}\}_{i,t\geq 2011}$ and $\widetilde{X}_1 = \{\widetilde{x}_{i,t-1}\}_{i,t\geq 2011}$, we can represent the log-likelihood function of d_1 as either:

1. $l\left(d_1|X_1;\gamma,b\right)$: the weighted log-likelihood corresponding to the logit equation (9), or

2. $l^{me}(d_1|\widetilde{X}_1;\gamma,c)$: the weighted *mixed effect* log-likelihood obtained by inserting (10) into (9) and integrating out the measurement error components $u_i \sim N(0,\sigma_u^2)$ and $e_{it} \sim N(0,\sigma_e^2)$.

Our results show that maximizing $l^{me}(d_1|\widetilde{X}_1;\gamma,c)$ with respect to (γ,c) and maximizing $l(d_1|X_1;\gamma,b)$ with respect to (γ,b) give two almost identical maximum likelihood estimates of γ. This validates the relation (10) for the main period. However, for $t < 2011$, the only available data are $(d_0,\widetilde{X}_0)$, where the subscript 0 refers to data for $t = 2001,...,2010$ (auxiliary period).

We will now detail how to combine $l(d_0|\widetilde{X}_0;\gamma,c)$ (auxiliary period) and $l(d_1|X_1;\gamma,b)$ (main period) to obtain a pooled estimator of γ that uses rating data in the auxiliary period and payment remark data in the main period. First, we construct the profile log-likelihood of the data d_0 by maximizing out the nuisance parameters, c, from $l(d_0|\widetilde{X}_0;\gamma,c)$ (see Barndorff-Nielsen and Cox (1994) for the general theory of profile likelihood):

$$l^P(d_0|\gamma) = \max_c l(d_0|\widetilde{X}_0;\gamma,c) \simeq -\frac{1}{2}\frac{(\gamma-\widehat{\gamma}^0)^2}{Var(\widehat{\gamma}^0)} \tag{11}$$

The approximation in equation (11) is based on the quadratic Taylor expansion of $l(d_0|\widetilde{X}_0;\gamma,c)$ about the maximizer $(\widehat{\gamma}^0,\widehat{c}^0)$ (ignoring the uninteresting constant term), with corresponding variance estimate $Var(\widehat{\gamma}^0)$.

Second, we construct the profile log-likelihood of d_1 by maximizing out b from $l(d_1|X_1;\gamma,b)$:

$$l^P(d_1|\gamma) = \max_b l(d_1|X_1;\gamma,b) \simeq -\frac{1}{2}\frac{(\gamma-\widehat{\gamma}^1)^2}{Var(\widehat{\gamma}^1)} \tag{12}$$

where $(\widehat{\gamma}^1,\widehat{b}^1)$ is the maximizer of $l(d_1|X_1;\gamma,b)$ with respect to (γ,b). Since d_0 and d_1 are conditionally independent, the profile log-likelihood of the complete data for 2001–2020 is:

$$l^P(d|\gamma) = l^P(d_1|\gamma) + l^P(d_0|\gamma)$$

From (11)-(12), we obtain an approximate maximum profile likelihood estimator of γ as follows:

$$\widehat{\gamma} = Var(\widehat{\gamma})\left(\frac{\widehat{\gamma}^0}{Var(\widehat{\gamma}^0)} + \frac{\widehat{\gamma}^1}{Var(\widehat{\gamma}^1)}\right)$$

where

$$Var(\widehat{\gamma}) = (\frac{1}{Var(\widehat{\gamma}^0)} + \frac{1}{Var(\widehat{\gamma}^1)})^{-1}$$

Finally, we estimate b by replacing γ with $\widehat{\gamma}$ in the log likelihood function of the main period data, y_1:

$$\widehat{b} = \arg\max_b l(d_1|X_1; \widehat{\gamma}, b)$$

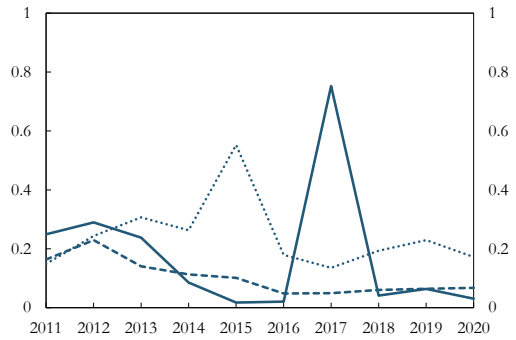
A.5 Table and charts refering to the variable selection based on the minimum CVE criterion

Table A.2: Variables included in final model selection by applying the one SE rule or the minimum CVE criterion in the first stage. The integers (1-6) indicates the industries[1)] where the given variable is included

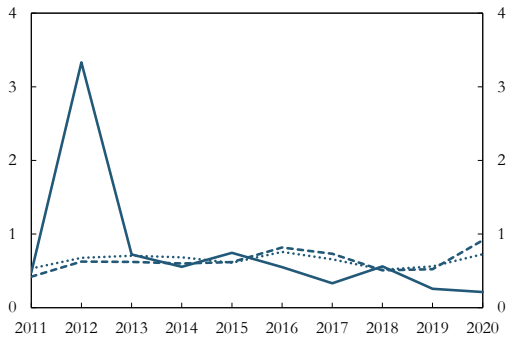
Variables	One SE rule		Minimum CVE	
	With weighting	Without weighting[2)]	With weighting	Without weighting
Continuous firm variables:				
RoA_t	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,4,5,6
RoA_{t-1}	1,4	1,2,4,6	2,4	1,3,4
E_{t-1}	1,2,5	2,3,4,5	2,3,4,5	1,2,3,4,5,6
Current liabilities ratio_t	6	2,3,4,5,6	4,6	2,3,4,5,6
Current liabilities ratio_{t-1}	2,3,5	3,6	2,3,4,5	4,5,6
Log real assets_t		1,2,3,4,5,6	3,5,6	2,3,4,6
Log real assets_t^2		1,2,3,4,5,6	2,3,4,5,6	2,3,4,5,6
Quick ratio_t			5	
Categorical firm variables:[3)]				
$\text{E}_t < 0$			1,5,6	1,2,3,4,5,6
$\text{C}_t = 0$	1,2,3,4,5,6	1,2,3,4,5,6	1,2,3,6	
$\text{C}_t = 2$		2,3,4,5,6	1,5	
Firm $\text{age group}_t = 1$			2,3,4,5	1,2,3,4,5,6
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0$	1,3,4,5,6	1,2,3,4,5,6	1,5	6
$\text{C}_t = 1 \cdot \text{C}_{t-1} = 0$			5	3,4,5
$\text{C}_t = 2 \cdot \text{C}_{t-1} = 0$			1	
$\text{E}_t < 0 \cdot \text{C}_{t-1} = 0$			1	1,2,3,4,5
$\text{E}_t < 0 \cdot \text{C}_t = 2$			1,5	3,5,6
$\text{E}_t \geq 0 \cdot \text{C}_t = 0$			1,2,5	1,2,3,4,5,6
$\text{E}_{t-1} < 0 \cdot \text{C}_t = 2$			2,3,5,6	5
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0 \cdot \text{C}_{t-1} = 0$			1,6	3,4,6
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0 \cdot \text{C}_t = 1$		2,3,5,6	1,6	1,2,3,4,5,6
$\text{E}_t < 0 \cdot \text{E}_{t-1} < 0 \cdot \text{C}_t = 2$	1,3,4,5,6	1,3,5,6	1,4	2,3,4,6
$\text{E}_t \geq 0 \cdot \text{E}_{t-1} \geq 0 \cdot \text{C}_t = 2$			1,5	3,4,5,6
$\text{E}_t < 0 \cdot \text{C}_t = 1 \cdot \text{C}_{t-1} = 0$			1,4,5	2,5,6
$\text{E}_t < 0 \cdot \text{C}_t = 2 \cdot \text{C}_{t-1} = 0$			1,3,4	3,6
$\text{E}_t \geq 0 \cdot \text{C}_t = 2 \cdot \text{C}_{t-1} = 0$			1,3,4	3,6
$\text{E}_{t-1} < 0 \cdot \text{C}_t = 1 \cdot \text{C}_{t-1} = 0$			1,2,4,5	2,3,4,5,6
$\text{E}_{t-1} < 0 \cdot \text{C}_t = 2 \cdot \text{C}_{t-1} = 2$			1	1,3,6

1)1=fishing and fish farming, 2=manufacturing, 3=construction, 4=retail trade, 5=commercial real estate, 6=services.
2) This is the model referred to as the unweighted benchmark model. It includes Log real assets and Log real assets^2 by default, i.e., without being selected by the one SE rule.
3) For the categorical variables, a statement A (e.g. $E < 0$) is short-hand notation for a dummy variable which is 1 if A is true (e.g. if $E < 0$).

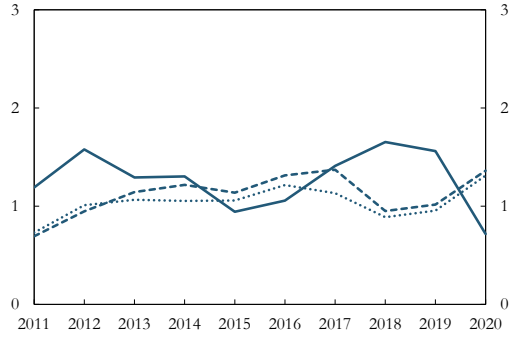
Chart A.2: Bankruptcy debt rates by industry. Actual rates and out-of-sample predictions with and without weighting. Predictors selected by applying minimum CVE in the first stage. 2011–2020. Percent.



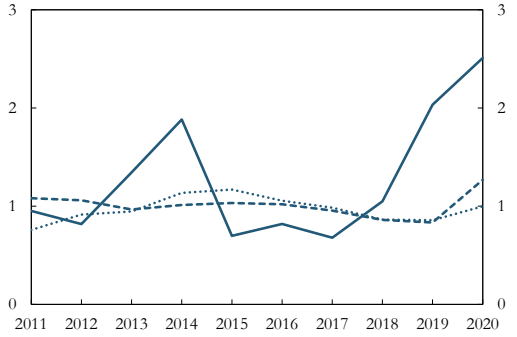
(a) Fishing and fish farming



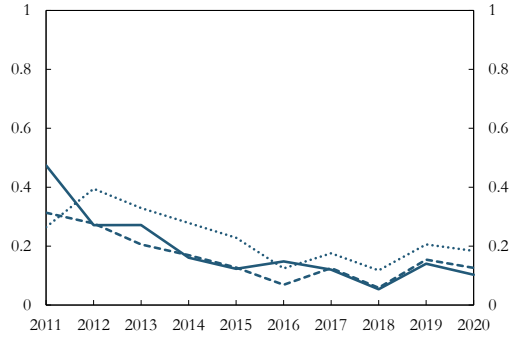
(b) Manufacturing



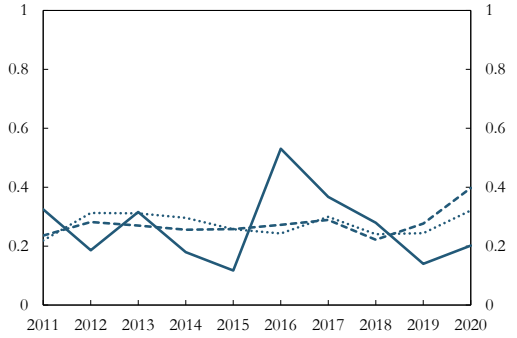
(c) Construction



(d) Retail trade



(e) Commercial Real estate



(f) Services

Actual share of bankruptcy debt
Predicted share of bakruptcy debt: with weights
Predicted share of bankruptcy debt: without weights

Chart A.3: Aggregate bankruptcy debt rates for the six industries. Actual rates and out-of-sample predictions with and without weighting. Predictors selected by applying minimum CVE in the first stage. 2011–2020. Percent.