Christoffersen, Benjamin; Matin, Rastin; Mølgaard, Pia

Working Paper

Can machine learning models capture correlations in corporate distresses?

Danmarks Nationalbank Working Papers, No. 128

Provided in Cooperation with:
Danmarks Nationalbank, Copenhagen

Suggested Citation: Christoffersen, Benjamin; Matin, Rastin; Mølgaard, Pia (2018) : Can machine learning models capture correlations in corporate distresses?, Danmarks Nationalbank Working Papers, No. 128, Danmarks Nationalbank, Copenhagen

This Version is available at:
https://hdl.handle.net/10419/202868

WORKING PAPER

DANMARKS NATIONALBANK

26 OCTOBER 2018 — NO. 128

Can Machine Learning Models Capture Correlations in Corporate Distresses?

Benjamin Christoffersen
bch.fi@cbs.dk
COPENHAGEN BUSINESS SCHOOL

Pia Mølgaard
pim@nationalbanken.dk
DANMARKS NATIONALBANK

Rastin Matin
rma@nationalbanken.dk
DANMARKS NATIONALBANK

Can Machine Learning Models Capture Correlations in Corporate Distresses?

Abstract
Accurate probability-of-distress models are central to regulators, firms, and individuals who need to evaluate the default risk of a loan portfolio. A number of papers document that recent machine learning models outperform traditional corporate distress models in terms of accurately ranking firms by their riskiness. However, it remains unanswered whether advanced machine learning models can capture correlation in distresses, which traditional distress models struggle to do. We implement a regularly top-performing machine learning model and find that prediction accuracy of individual distress probabilities improves while there is almost no difference in the predicted aggregate distress rate relative to traditional distress models. Thus, our findings suggest that complex machine learning models do not eliminate the need for a latent variable that captures correlations in distresses. Instead, we propose a frailty model, which allows for correlations in distresses, augmented with regression splines. This model demonstrates competitive performance in terms of ranking firms by their riskiness, while providing accurate risk measures.

Resume
Nøjagtige konkursmodeller er nødvendige for tilsynsmyndigheder, virksomheder og investorer, som har brug for at evaluere konkursrisikoen i en låneportefølje. En række papirer har vist, at "machine learning"-modeller er bedre til at rangere virksomheder efter deres kreditrisiko end traditionelle statistiske modeller. Men det er stadig et åbent spørgsmål, om de avancerede modeller kan fange korrelation i konkurser, hvilket traditionelle modeller har svært ved. Vi implementerer en machine learning-model, som generelt har vist sig at være god til at forudsige konkurser, og finder, at konkursestimater på virksomhedsniveau forbedres, mens de estimerede aggregerede konkursrater forbliver næsten uændrede i forhold til de traditionelle konkursmodeller. Vores resultater taler altså for, at komplekse machine learning-modeller ikke eliminerer nødvendigheden af at inkludere en latent variabel, der fanger korrelation i konkurser. Som et alternativ implementerer vi en "frailty"-model, som direkte introducerer korrelation i konkurser. Modellen er ydermere udvidet med "regression splines", hvilket medfører, at den er god til at rangere virksomheder efter deres kreditrisiko, samtidig med at den vurderer risikoen i en låneportefølje korrekt.

Key words

Credit risk; Risk management

JEL classification
C55, G17, G33, M41.

Acknowledgements
The authors are grateful to Mads Stenbo Nielsen (discussant), David Lando, Søren Feodor Nielsen, seminar participants at Copenhagen Business School, and colleagues at Danmarks Nationalbank for helpful comments.

The authors alone are responsible for any remaining errors.

Can Machine Learning Models Capture Correlations in Corporate Distresses?*

Benjamin Christoffersen

Center for Statistics, Copenhagen Business School, DK-2000 Frederiksberg, Denmark, bch.fi@cbs.dk

Rastin Matin

Danmarks Nationalbank, DK-1093 Copenhagen, Denmark, rma@nationalbanken.dk

Pia Mølgaard

Danmarks Nationalbank, DK-1093 Copenhagen, Denmark, pim@nationalbanken.dk

Thursday 25th October, 2018

Abstract

Accurate probability-of-distress models are central to regulators, firms, and individuals who need to evaluate the default risk of a loan portfolio. A number of papers document that recent machine learning models outperform traditional corporate distress models in terms of accurately ranking firms by their riskiness. However, it remains unanswered whether advanced machine learning models can capture correlation in distresses, which traditional distress models struggle to do. We implement a regularly top-performing machine learning model and find that prediction accuracy of individual distress probabilities improves while there is almost no difference in the predicted aggregate distress rate relative to traditional distress models. Thus, our findings suggest that complex machine learning models do not eliminate the need for a latent variable that captures correlations in distresses. Instead, we propose a frailty model, which allows for correlations in distresses, augmented with regression splines. This model demonstrates competitive performance in terms of ranking firms by their riskiness, while providing accurate risk measures.

Keywords: corporate default prediction, discrete hazard models, frailty models, gradient boosting
JEL: C55, G17, G33, M41

*We are grateful to Mads Stenbo Nielsen (discussant), David Lando, Søren Feodor Nielsen, and seminar participants at Copenhagen Business School and Danmarks Nationalbank for helpful comments.

1 Introduction

Estimating accurate corporate distress probabilities is of particular interest to central banks in the European Union the coming years. Following the regulation on the collection of credit risk data of the European Central Bank (ECB), members of the euro area are obliged to establish central credit registers and to participate in a joint analytical credit database ("AnaCredit") shared between the member states. The database will contain detailed information on lending by commercial banks to corporate borrowers. Consequently, central banks can closely study the credit risk of a particular bank's loan portfolio. For that purpose, it is essential to model the probability of default of a group of individual borrowers jointly accurate in order to estimate portfolio risk measures, just as modelling jointly accurate corporate distress probabilities is important to any entity with portfolio risk.

A growing literature focuses on the application of machine learning models, i.e. complex models with highly non-linear dependency structures between the covariates and the outcome, to predict corporate bankruptcies (see e.g., Min and Lee 2005, Tinoco and Wilson 2013, Jones et al. 2017). These papers show applications of one or more complex statistical models which are commonly benchmarked against a logistic regression. Model performance is then evaluated by rank- or binary-based performance metrics comparing the models' ability to classify or predict the distress of a firm. However, the models' ability to accurately estimate the aggregated percentage of firms that will default in the next period remains uninvestigated. Nor is the models' ability to provide accurate portfolio risk measures addressed.

Another string of literature, pioneered by Duffie et al. (2009), shows that traditional hazard models (e.g., logistic regression models) yield too narrow confidence intervals of the aggregated default rate due to the model assumption that observations are conditionally independent. Duffie et al. (2009) then advocate for the need for unobservable temporal effects – or frailty – in the models, which add correlations in defaults after conditioning on covariates, thereby easing the conditional independence assumption. The conditional independence assumption is also implicitly made in most complex statistical models. However, whether this affects such models' ability to accurately estimate the distress rate as well as the risk of a loan portfolio is an open question.

In this paper we investigate whether complex statistical models, via their sophisticated dependency structures, can capture the correlation in corporate distress from firm level data alone and thereby eliminate the need for unobservable temporal effects. We implement a gradient boosted tree model which has displayed superior performance in both bankruptcy prediction and other fields.[1] We find that the model is as unable to capture the yearly heterogeneity in distress rates as traditional distress model. The gradient boosted

[1] See Caruana and Niculescu-Mizil (2006) for a comparison in many other fields and Zięba et al. (2016), Jones et al. (2017) who have applied gradient tree boosting to firm distress or bankruptcy prediction with success.

tree model is also unable to provide appropriate estimates for the level of uncertainty in a loan portfolio. Comparing results of the gradient boosted tree model to results of a model with frailty, which models confidence intervals and risk measures accurately, we show that loan portfolios of, in particular, large banks can be viewed as too safe in the eyes of the regulator and/or risk manager, if he or she relies on a gradient boosted tree model.

Our sample consists of annual financial accounts published between 2003 and 2018 of all non-financial Danish firms both traded and non-traded. Considering both traded and non-traded firms yields a large sample which allows us to include many covariates and add non-linear effects. The models in the main body of the paper are based solely on micro level data. In a robustness test we show that models including macro level data perform better in some periods. However, estimating a model that generalizes well may be hard with limited amount of cross-sections. Lastly, the unobserved temporal effect is still economically and statistically significant after the inclusion of the macro variable.

We start the analysis by benchmarking the gradient boosted tree model against a multiperiod logit model (as in Shumway 2001, Chava and Jarrow 2004, Beaver et al. 2005, Campbell et al. 2008) and a generalized additive model, which allows for a non-linear relationship between the covariates and the probability of entering into a distress on the logit scale. Like others before us, we observe improvements in out-of-sample ranking of firms by their distress probability as we use more complex models, going from an average out-of-sample area under the receiver operating characteristic curve (AUC) of 0.798 to an AUC of 0.822. Thus, we find that the more complex model is 2.4 percentage points more likely to predict a higher distress probability for a random distressed firm than for a random non-distressed firm in each year on average. However, the gains we find of complex modelling is more than 4 times smaller than what recent papers find.[2] Thus, one may prefer the simpler models if interpretability is desired with only a minor loss of accuracy. Our finding suggests that earlier papers have used poor baseline models when evaluating the gains of applying complex machine learning models.

Next, we address the models' ability to predict the percentage of firms that will enter into a distress in the following period. We find that all models fail to capture the yearly fluctuations in distress rates and provide too narrow confidence bounds. In particular, only very few of the 90% confidence intervals contain the realized percentage of firms entering into distress in the 10 years that we can backtest. We formally test the models' ability to provide accurate confidence intervals by backtesting estimated value-at-risk figures of the distress rates for different portfolios that mimic bank exposures. All three models fail the test at a 1% significance level with a null hypothesis that the value-at-risk figures have the correct coverage. Thus, none of the models have wide enough confidence intervals or provide accurate risk measures.

[2]See Zięba et al. (2016) and Jones et al. (2017).

The too narrow confidence bounds have several implications. First, they result in a downward bias in risk measures for a portfolio of exposures to different firms. Secondly, they suggest that the assumption of conditional independence given the covariates is not satisfied. Violation of the conditional independence assumption suggests that there may exist an unobservable macro effect that creates correlation in distresses. That is, the gradient boosted tree model is not sufficiently able to capture correlation in distresses from firm level data alone.

To relax the conditional independence assumption we estimate a generalized linear mixed model (a frailty model) with a random intercept which allows for correlation in distresses beyond the correlation introduced by the covariates. We contribute to the current literature on frailty models by adding non-linear dependencies between the covariates and the outcome variable. This gives us a frailty model which provides out-of-sample rankings that are almost as good as the gradient boosted tree model. We show that the random intercept in the frailty model is both statistically and economically significant.

2 Related Literature

This paper combines two strings of literature in the field of predicting corporate defaults. The first string focuses on frailty (and/or contagion) or time-varying effects (e.g., see Duffie et al. 2009, Koopman et al. 2011, Giesecke and Kim 2011, Duan and Fulop 2013, Lando et al. 2013, Nickerson and Griffin 2017, Kwon and Lee 2018, Azizpour et al. 2018). These papers generally show that models with a simple relationship between observable covariates and distress fail to capture the yearly fluctuations in default rates, i.e. a violation of the conditional independence assumption. Various forms of unobservable effects are then introduced which account for the yearly fluctuations. Our contribution to this line of work is a frailty model, where non-linear dependencies are introduced between some covariates and the outcome variable on the linear predictor scale. Furthermore, we compare the frailty model to a statistical model that allows for complex dependencies between covariates and the outcome variable and find that the frailty model shows almost as good ranking and better coverage of the confidence bounds. Moreover, we provide evidence that the need for frailty effects is not due to a too simple dependency structure.

The second string of literature that we relate to applies complex statistical models to improve probability of default estimates (e.g., see Min and Lee 2005, Kim and Kang 2010, Sun et al. 2011, Lin et al. 2012, Tinoco and Wilson 2013, Zięba et al. 2016, Jones et al. 2017). These papers generally use considerably more covariates in their models and use methods which allow for more complex relationships compared to typical frailty models. The main focus of these papers is on ranking or binary classification of firms and not on whether the models capture the temporal fluctuations. The complex models are typically benchmarked

against a logistic regression (among other models) with automated model selection and little focus on model diagnostics. In our paper we use a logistic model as benchmark as well, but we carefully set up the model using both statistical and economic sense. We add to this literature by evaluating the ability of the complex model to capture the yearly fluctuation in default rates. We show that the improvements in the forecasts for each firm do not outweigh the strict conditional independence assumption when one is interested in portfolio risk.

3 Statistical Models for Predicting Corporate Distress

In this section we go through the four discrete hazard models used in this paper to predict corporate distress. The discrete hazard models we use are estimated using a panel data set where each observation contains a set of covariates (financial ratios, age, sector etc.) and an indicator of whether the firm has a distress event or not in the given year. We will cover the distress event definition and discrete hazard models further in Section 4.1.

First, we briefly describe the well known multiperiod logit model. The notation introduced in this section will serve as the basis for the more general models. Secondly, we describe the generalized additive model which allows for a non-linear dependence between the covariates and the probability of distress on the logit scale. Thirdly, we describe the gradient boosted tree method we use. Finally, we introduce the generalized linear mixed model which relaxes the conditional independence assumption.

3.1 Generalized Linear Models

We will use so-called multiperiod logit models, where we employ a logistic regression in the discrete hazard model described in Section 4.1. Estimation in the multiperiod logit model can be done with maximum likelihood with iteratively re-weighted least squares. Let $R_t \subseteq \{1, \cdots, n\}$ denote the active firms at time t, y_{it} denote whether firm i has an event in year t, d denote the number of years, and $\boldsymbol{x}_{it}$ denote the covariates for firm i in year t. Then the maximum likelihood estimates of the coefficients, $\boldsymbol{\beta}$, are

$$\arg\max_{\boldsymbol{\beta}} \sum_{t=1}^{d} \sum_{i \in R_t} y_{it} \boldsymbol{\beta}^\top \boldsymbol{x}_{it} - \log\left(1 + \exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_{it}\right)\right) \tag{1}$$

where $\boldsymbol{x}_{it}$ includes a constant 1 for the intercept, industry covariates, and potentially macro covariates. Furthermore, we will refer to R_t as the *risk set* and let $\mathbf{X}_t$ denote the matrix with rows equal to the covariate vectors $\boldsymbol{x}_{it}$ with $i \in R_t$. Multiperiod logit models are a common choice for distress models since the work of Shumway (see Shumway 2001). See Chava and Jarrow (2004), Beaver et al. (2005), Campbell et al. (2008)

for other examples. We will refer to multiperiod logit models as generalized linear models (GLMs) since estimation is done with regular estimation methods for GLMs.

3.2 Generalized Additive Models

The GLM in Section 3.1 may pose too strict assumptions on the relationship between the covariates and whether a firm enters into distress. In particular, the assumption that the covariates are linearly related to the logit of the probability of distress may be too strict for some of the covariates. Generalized additive models (GAMs) relax this assumption by assuming that some of the covariates have a continuous and non-linear relationship with the distress probability on the logit scale.

We employ a GAM where non-linear effects are accounted for through natural cubic splines with a penalty on the second order derivative. The maximization problem with q non-linear effects and with given penalty parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^\top$ is

$$\boldsymbol{\beta}(\boldsymbol{\lambda}) = \arg\max_{\boldsymbol{\beta}} \sum_{t=1}^{d} \sum_{i \in R_t} y_{it}\eta_{it} - \log\left(1 + \exp\left(\eta_{it}\right)\right) - \boldsymbol{\beta}^{(s)\top}\mathbf{S}(\boldsymbol{\lambda})\boldsymbol{\beta}^{(s)} \tag{2}$$

where

$$\eta_{it} = \boldsymbol{\beta}^{(f)\top}\boldsymbol{x}_{it}^{(f)} + \sum_{j=1}^{q} \boldsymbol{\gamma}_j^\top \boldsymbol{f}_j(x_{itj}^{(s)}), \quad \boldsymbol{\beta}^{(s)} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_q \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^{(f)} \\ \boldsymbol{\beta}^{(s)} \end{pmatrix} \tag{3}$$

$\boldsymbol{f}_j$s are functions that return a basis vector for a natural cubic spline, $x_{itj}^{(s)}$ is firm i's covariate j with a non-linear effect at time t, $\boldsymbol{x}_{it}^{(f)}$ are the covariates with a linear effect for firm i at time t, and $\mathbf{S}(\boldsymbol{\lambda})$ is a penalty coefficient matrix which yields a second order penalty on each spline $j = 1, \ldots, q$. The knots for the natural cubic spline basis are chosen as empirical quantiles. Equation (2) can be solved with penalized iteratively re-weighted least squares if $\boldsymbol{\lambda}$ is known.

The penalty coefficient matrix, $\mathbf{S}(\boldsymbol{\lambda})$, depends linearly on the unknown penalty parameters, $\boldsymbol{\lambda}$. The penalty parameters, $\boldsymbol{\lambda}$, have to be estimated. This is done by minimizing the generalized cross-validation criterion which can be quadratically approximated as

$$\mathcal{V}(\boldsymbol{\lambda}) = \frac{n\left\|\sqrt{\mathbf{W}}\left(\boldsymbol{z} - \mathbf{X}\boldsymbol{\beta}(\boldsymbol{\lambda})\right)\right\|^2}{\left(n - \operatorname{tr}\left(\mathbf{F}_{\boldsymbol{\lambda}}\right)\right)^2}$$

where

$$\mathbf{F}_{\boldsymbol{\lambda}} = \left(\mathbf{X}^\top\mathbf{W}\mathbf{X} + \mathbf{S}(\boldsymbol{\lambda})\right)^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{X} \tag{4}$$

$n_t = |R_t|$ is the number of active firms at time t, $n = \sum_{t=1}^{d} n_t$ is the total number of observations, and $\text{tr}(\cdot)$ denotes the trace of a matrix. Furthermore, $\boldsymbol{y}$, $\mathbf{X}$, $\boldsymbol{z}$, and $\mathbf{W}$ denote the stacked matrices and vector from each year (e.g., $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_d^\top)^\top$). The columns of $\mathbf{X}$ include the evaluated basis functions, $\boldsymbol{f}_j$s, for the non-linear effects. $\mathbf{W}$ and $\boldsymbol{z}$ are the diagonal matrix with working weights and vector of pseudo-responses from iterative re-weighted least squares, respectively. They implicitly depend on $\boldsymbol{\beta}(\boldsymbol{\lambda})$, $\boldsymbol{y}$ and $\mathbf{X}$.[3] The maximization is done with the so-called performance-oriented iteration. See Wood et al. (2015), Wood (2017) for further technical details.

The final model also includes tensor product splines to allow for smooths in two dimensions. These are formed by taking an outer product of two spline basis functions, $\boldsymbol{f}_j$s, and is more general than the model in Equation (3). The extension to two-dimensional smooths is straightforward, but not covered in Equation (3) to keep the notation simple. GAMs have received limited attention in the corporate default literature (see e.g., Berg 2007).

The advantage of the GAM is that the researcher has control over the complexity of the model. For example, he or she can decide which covariates have a non-linear effect and which do not. Moreover, it is easy to validate whether the final model makes sense through standard diagnostic plots, to obtain marginal effects of covariates, to compute confidence intervals, etc. However, the researcher has to consider the effect and interactions of the covariates which may be hard, especially with higher order non-linear interactions.

3.3 Gradient Tree Boosting

Gradient tree boosting (GB) is a greedy function approximation method that can approximate very complex models. GB has gained much attention possibly due to its flexibility and easy usability. The researcher has only few and simple model choices relative to the GAM described in Section 3.2. Furthermore, GB has shown superior performance in many fields, see e.g., Caruana and Niculescu-Mizil (2006), where an empirical study is presented on different data sets where GB performs best on average on many metrics. However, the advantages of GB come at a cost of limiting the researcher's ability to set the complexity of the effect of each covariate. Furthermore, it is not clear how to perform inference such as testing significance of partial effects, and evaluating if the final model is "sensible" for various combinations of covariates may be difficult if one allows for higher-order interactions (i.e. deep trees). Lastly, figuring out why a given observation gets the predicted probability is not as easily done as with the GLM and GAM. This is a drawback for a financial institution that is required to provide an explanation of why a certain probability of distress is predicted.

[3] Let g denote the link function which maps from the probability of an event to the linear predictors, η_{it}, in Equation (3), let $\widehat{p}_{it} = g^{-1}(\eta_{it})$ be the expected probability of an event at the current iteration doing estimation or at convergence, and let $V(p) = p(1-p)$ denote the map from the probability of an event to the variance. Then $z_{it} = \eta_{it} + g'(\widehat{p}_{it})(y_{it} - \widehat{p}_{it})$ and $w_{iit} = 1/g'(\widehat{p}_{it})^2 V(\widehat{p}_{it})$.

We will cover gradient tree boosting in the context of classification with the logit link function. The interested reader is referred to Friedman (2001), Bühlmann and Hothorn (2007) for more details, Natekin and Knoll (2013) for a brief tutorial, and Chen and Guestrin (2016) for the software implementation of GB we use. We use Newton boosting, but in the following we will refer to it as gradient boosting as commonly done in literature. The estimation is done as follows: Denote the estimated mean probability of a distress by

$$\bar{p} = \frac{1}{n} \sum_{t=1}^{d} \sum_{i \in R_t} y_{it}$$

Initialize the linear predictors as $\eta_{it}^{(0)} = f^{(0)}(\boldsymbol{x}) = \text{logit}(\bar{p})$, where $\text{logit}(p) = \log(p/(1-p))$ is the logit function. Let $\mathbf{X}$, $\boldsymbol{y}$, and $\boldsymbol{\eta}^{(i)}$ denote the stacked matrix and vectors such that e.g., $\boldsymbol{\eta}^{(i)} = (\boldsymbol{\eta}_1^{(i)\top}, \ldots, \boldsymbol{\eta}_d^{(i)\top})^\top$. Define the loss function, L, as

$$L(\boldsymbol{\eta}) = \sum_{t=1}^{d} \sum_{i \in R_t} l(\eta_{it}; y_{it})$$

$$l(\eta; y) = -y\eta + \log\left(1 + \exp(\eta)\right)$$

Then for $i = 1, \ldots, k$

1. compute the first and second order derivatives using the linear predictors from the previous iteration and denote these by

$$g_{it} = -y_{it} + \left(1 + \exp\left(-\eta_{it}^{(i-1)}\right)\right)^{-1}$$
$$h_{it} = \exp\left(-\eta_{it}^{(i-1)}\right)\left(1 + \exp\left(-\eta_{it}^{(i-1)}\right)\right)^{-2}$$

2. fit a regression tree denoted by $a^{(i)}(\boldsymbol{x})$ which is an approximation to

$$\arg\min_{a \in \mathcal{C}} \sum_{t=1}^{d} \sum_{i \in R_t} h_{it} \left(-\frac{g_{it}}{h_{it}} - a(\boldsymbol{x}_{it})\right)^2$$

   where $\mathcal{C}$ is the set of trees we consider (e.g, trees with a given maximum depth).
3. update the model such that $f^{(i)}(\boldsymbol{x}) = f^{(i-1)}(\boldsymbol{x}) + \rho a^{(i)}(\boldsymbol{x})$, where $\rho \in (0, 1]$ is a predetermined shrinkage parameter.
4. update the linear predictors by computing $\eta_{it}^{(i)} = f^{(i)}(\boldsymbol{x}_{it})$.

The final GB model is the function $f^{(k)}$. There are three main parameters in the above algorithm: The shrinkage parameter ρ, the maximum depth of the trees in step 2, and the number of trees k. We select

these with 5-fold cross-validation where we sample the firms (not the financial statements) and evaluate the AUC which is introduced in Section 5. In general, it is preferable to decrease the shrinkage parameter, ρ, while increasing the number of trees, k, to get a better approximation of the true dynamics. However, one has a finite budget in terms of computational power, which limits k and thus forces one to select ρ to get the optimal number of trees around k. We fix k to around 250 and at most 300. We find ρ with cross-validation on the full sample from 2003-2016 which we will describe in Section 4.1. We find only very small improvements of decreasing the learning rate and using more trees. This only leaves us with a choice for the maximum depth of trees.

Usually so-called 'weak learners' (biased methods) are used in step 2 above. In our case, this amounts to shallow trees (trees with a low maximum depth). The weak learners are then combined through gradient boosting yielding one "good" model with a substantially lower bias than any of the learners while not affecting the variance much. See Bühlmann and Hothorn (2007) for some simpler examples with theoretical results. For the aforementioned reasons, we have tried maximum depths of 2-6 in preliminary testing. We used 5-fold cross-validation as described above. We find little difference in model performance when going from tree depths of 3 to 6. Thus, we choose a maximum depth of 3.

Given the fixed learning rate and maximum depth of 3, we estimate the optimal number of trees each year when we run our out-of-sample tests. The estimations are done again with 5-fold cross-validation and by sampling firms and not financial statements. We note that the estimation of the optimal number of trees is done on the estimation sample and not the test set.

3.4 Generalized Linear Mixed Models

We can extend the GLM from Section 3.1 to relax the conditional independence assumption by generalizing to a generalized linear mixed model (GLMM). This can be done by changing the conditional mean in the GLM from

$$\mathrm{E}\left(Y_{it} \middle| \boldsymbol{x}_{it}\right) = g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{x}_{it}), \quad g^{-1}(\eta) = \mathrm{logit}^{-1}(\eta)$$

to

$$\mathrm{E}\left(Y_{it} \middle| \boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \epsilon_t\right) = g^{-1}\left(\boldsymbol{\beta}^\top \boldsymbol{x}_{it} + \epsilon_t\right) \tag{5}$$

where $\epsilon_t \sim N\left(0, \sigma^2\right)$ is the random effect at time t, $\epsilon_i \perp \epsilon_j$ for $i \neq j$. Thus, the optimization problem becomes

$$\arg\max_{\boldsymbol{\beta}, \sigma^2} \sum_{t=1}^{d} \int_{(-\infty,\infty)} \left(\sum_{i \in R_t} y_{it} \eta_{it} - \log\left(1 + \exp\left(\eta_{it}\right)\right) \right) \varphi\left(\epsilon_t; \sigma^2\right) \partial \epsilon_t \tag{6}$$

$$\eta_{it} = \boldsymbol{\beta}^\top \boldsymbol{x}_{it} + \epsilon_t \tag{7}$$

where $\varphi(x, \sigma^2)$ is the density of a normal distribution with zero mean and variance σ^2. The log likelihood in Equation (6) has no closed form solution in general, but can be approximated with a Laplace approximation. Furthermore, the computational cost of the approximation can be greatly reduced if one exploits the sparsity of the matrices which are decomposed doing the estimation. See Bates and DebRoy (2004), Bates et al. (2015) for further details about the estimation method. The linear predictor in Equation (7) is easily modified to include splines by changing the $\boldsymbol{\beta}^\top \boldsymbol{x}_{it}$ part such that

$$\eta_{it} = \boldsymbol{\beta}^{(f)\top} \boldsymbol{x}_{it}^{(f)} + \sum_{j=1}^{q} \boldsymbol{\gamma}_j^\top \boldsymbol{f}_j(x_{itj}^{(s)}) + \epsilon_t, \quad \boldsymbol{\beta}^{(s)} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_q \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^{(f)} \\ \boldsymbol{\beta}^{(s)} \end{pmatrix}$$

which is similar to Equation (3). We call the random effect, ϵ_t, frailty though it is not a frailty in the original sense of frailty as popularized in Vaupel et al. (1979). The random effect variable in Vaupel et al. (1979) and Duffie et al. (2009) is a multiplicative factor on the hazards. Our random effect is multiplicative on the odds rather than the hazard since we can factorize Equation (5) when g is the logit function as

$$\frac{\mathrm{E}\left(Y_{it} \middle| \boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \epsilon_t\right)}{1 - \mathrm{E}\left(Y_{it} \middle| \boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \epsilon_t\right)} = \exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_{it}\right) \exp\left(\epsilon_t\right)$$

Thus, firms have a higher frailty if ϵ_t is large in a given year yielding an $\exp(\epsilon_t)$-factor higher odds of distress. The case $\epsilon_t = 0$ can be seen as a 'standard' year. Random effect models have received a lot of attention in the literature. The focus of these papers is on the structure of the random effects. E.g., Duffie et al. (2009) and Koopman et al. (2011) let the probability of distress depend on an unobservable order-one autoregressive process. However, contrary to Duffie et al. (2009), Koopman et al. (2011) assume that groups of firms depend differently on the unobservable process. We are limited in terms of how many random effects we can estimate as we only have 14 years of data. Thus, we will only estimate a single random intercept, where we assume that the ϵ_t's are iid as in Equation (6). An autocorrelation plot of estimated ϵ_ts does not show signs of autocorrelation.

4 Data

Our main data set consists of all non-consolidated financial statements filed by Danish private and public limited companies in the period 2003 to 2018.[4] The financial statements are supplemented with firm characteristics such as age, sector, and legal status from the Danish Central Business Register (CVR). As we are conducting a prediction exercise we utilize financial statements as of their publication date and not the accounting period end date. In our sample, financial reports are typically made public 5 months after the accounting period end date.[5] We use only the most recent published accounting data for each firm from year $t-1$ in year t in our models.

We apply standard filters to focus the analysis on the core of the Danish corporate sector. First, we exclude financial firms. Financial firms are different from other corporate companies in their size and the complexity of their assets, the accounting standards they comply with, and their special status in the eyes of the regulators. Furthermore, we exclude holding companies. Holding companies are usually set up with the sole purpose of extracting and dividing revenue from the firm to one or more owners of the firm, thus they are structured differently than other firms. We exclude them in order to avoid distorting the model estimation, but include the companies held by the holding companies. Finally, we exclude a small number of financial statements which are filed in Denmark in other currencies than DKK, EUR, GBR, USD, or SEK.[6] We do not impose any restrictions on firm size as we want to capture the whole economy in the analysis.

One could argue that the analysis should focus merely on "large" firms, as they hold the majority of the total assets and debt. But instead of estimating models on just large firms we allow for interaction between firm size and other variables in the GAM and GB model, thereby creating different models for firms of different sizes. Among the interactions tested we find that the interaction between scaled net profit and the log of the size variable we introduce later as well as between scaled liquid assets and log size are significant in the GAM. Including small firms increases our sample size which is important in order to estimate the non-linear effects in the GAM and GB model. The GLM we estimate does not include any interaction terms between firm size and other variables. The performance with respect to large firms is improved when we estimate a separate GLM for large firms, but remains inferior to the performance of GAM and GB model. For consistency, all results will be of the models estimated on the full sample.

[4] Financial statements are delivered to us by Bisnode and Experian.

[5] The exact publication date is used for statements filed from 2012 and onward. Unfortunately, we do not have access to the publication date of statements filed before 2012. For these statements we set the publication date to 6 months after the accounting period end date. We have two reasons for doing this. First, Danish law requires that the majority of firms in our sample must publish their financial statements within 5 months of the accounting period end date. We use 6 months instead of 5 to be conservative. Secondly, we find that 96% of financial statements are published within 6 months of the accounting period end date in the sub-sample where we have the publication date.

[6] Accounting variables reported in other currencies than DKK are converted to DKK as follows. All stock variables are converted using the end of accounting period exchange rate. All flow variables are converted using the daily average exchange rate over the accounting period.

The filtered data set includes 198 929 individual firms and 1.3 million firm years in the 2003 to 2016 period. Of the 198 929 firms, 43 674 enter into a distress period at least once. The seemingly high rate is due to a larger distress rate for small firms. An interesting aspect of our sample is that it includes non-traded firms, which are less studied in the literature.

4.1 Event Definition and Censoring

We obtain information on the full history of each firm's status from the Danish Central Business Registry (CVR). The CVR categorizes firm status into 21 categories. We combine categories into three groups: "normal", "in distress" and "other".[7] The "in distress" category includes firms in bankruptcy, firms that went bankrupt, firms under compulsory dissolution, or firms that have ceased to exist due to compulsory dissolution.

Our definition of "in distress" implies that firms that are "in distress" can become active again. Thus, we model recurrent events. We choose this framework as creditors are likely to suffer losses when a firm enters into a distress period, even if the firm becomes active again, due to delayed payments or a write-down of the debt. 3.4% of the firms in our sample have experienced a prior distress (some before 2003) and have recovered. Furthermore, 1 352 of these firms enter into more than one distress period during our sample period.

Distress dates are highly seasonal and reflects a potentially delayed processing time of the authorities.[8] Thus, we limit the models to be on a yearly basis. Each year includes all firms that:

1. had a "normal" status at the end of the previous year.
2. published a financial statement within the previous year.
3. (a) enter into "in distress" the following year or
   (b) do not publish a new financial statement the following year and enters into the "in distress" status within two years of the publication date of the latest financial statement or
   (c) are still "normal" at the end of the year (i.e. are not censored).

Firms that fulfil all of the above conditions are denoted *active* at the beginning of the given year. Among these firms, we say that a firm has an *event* if it satisfies condition 3a or 3b, or that the firm is a *control* if it satisfies condition 3c. Condition 3b is similar to the event definition in Shumway (2001), who defines a firm as going bankrupt if the firm delists the following year and "files for any type of bankruptcy within 5

[7] The "other" group includes firms that are under liquidation, liquidated, merged and split.

[8] Every year, there is a large "peak" in reported distress events in a single month in the fall and this peak does not fall on the same month every year. This arbitrary peak in reported distress events makes it questionable whether there is any meaning in the exact reporting month.

years of delisting". The difference to our data set is that firms do not delist, but instead do not publish a new financial statement. We also include a few firms that satisfy 3a or 3b as events if they enter into the "other" status between the "normal" and the "in distress" status.

In our event definition we have chosen a window of 2 years between the publication date of the last financial statement and the declaration date of "in distress". Most distresses in our sample are declared approximately 1.5 years after the publication date of the last financial statement but some occur later. We find, across years, that 95% to 99% of all "in distress" statuses are declared within the 2 year window we have chosen.

4.2 Covariates

It is commen to scale most of the financial statement variables by total assets to get all financial statement variables on a common scale. However, a non-trivial fraction of the firms in our sample have negative equity at some point. Thus, using total assets as the denominator will yield extreme covariates which may not fit well in a GLM. As Campbell et al. (2008) we define a more suitable metric to capture the firm size. We define firm size as

$$\text{size}_{it} = \max\{\text{debt}_{it}, \text{total assets}_{it}\} \tag{8}$$

where debt_{it} and total assets_{it} refer to the debt and total assets of firm i on the balance sheet from the financial statement published between year $t-1$ and t, respectively. Thus, size_{it} equals the total debt of the firm when equity is negative and otherwise total assets. We use size_{it} in the denominator of all the ratios where we would otherwise use total assets.

Besides the financial statement variables we include some variables that we have constructed ourself. Most interestingly, we include an industry specific covariate as in Chava et al. (2011) by computing the average net profit divided by the size variable each year for each leading four digit standard industrial classification (SIC) group. Unlike Chava et al. (2011), we do not have the stock return so we use the net profit divided by the size variable. Moreover, Chava et al. (2011) include a dummy for whether median stock return in the industry is below -20%. We do not believe that the variable has a discrete effect upon exceeding a pre-specified threshold. Therefore, we include the average value and estimate a slope. We winsorize[9] all covariates at 5% and 95% quantile as in Campbell et al. (2008). Preliminary results showed influential observations and poorer fits in the GLM when more extreme quantiles were used.

We end up with 44 numerical and 6 categorical covariates. We use the Thresholded Lasso estimator described in the Appendix A to select the covariates we will use in the GLM, GAM, and GLMM. 3 of the

[9]Cap values at a given high level quantile and floor at a given low level quantile. We winsorize ratios and not the numerator and denominator separately for ratio covariates.

covariates are excluded. All covariates are included in the GB model. The GB model tends to be robust against redundant covariates (e.g., this model do not have the same issues with multicollinearity as the other models). Another advantage of the GB model is that the regression trees used in the model are invariant to monotone transformations of the covariates. Consequently, we include both the non-winsorized ratios and the original (non-ratio) figures from the financial statements in the GB model. Descriptive statistics of all of the covariates can be seen in Appendix A.

5 Performance of the GLM, the GAM, and the GB Model

In this section, we perform out-of-sample tests of the GLM, GAM, and GB model presented in Section 3.1, 3.2, and 3.3 respectively. We will use an expanding window of data to estimate the models and forecast the probability of the firms entering into distress two years after the estimation window closes. As an example, we use models estimated on 2003 to 2007 data to predict default probabilities on 2009 data. The two-year gap mimics the true forecasting situation as the definition of the distress event requires a lag of two years.

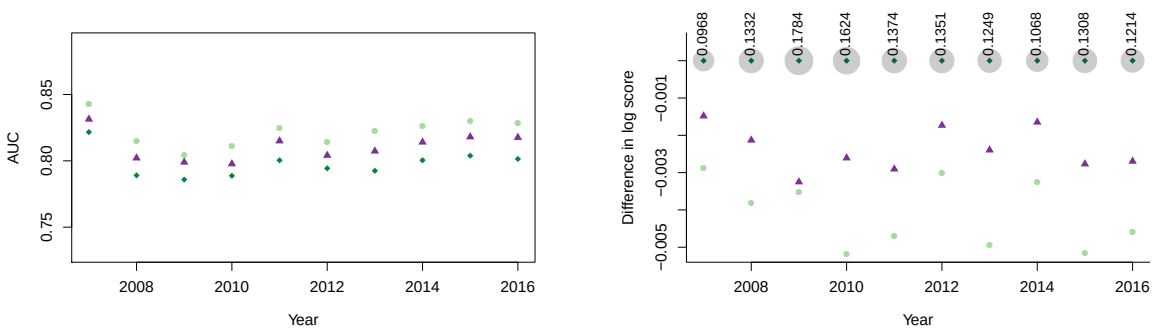
We measure performance on several different metrics. First, we consider the accuracy of the individual probability of distress estimates by comparing the AUC and the log score of the individual models. Next, we consider the performance of the models at an aggregated level by examining the models' ability to predict next year's aggregated percentage of firms in distress as well as the aggregated debt in distress. Finally, we look at the models' ability to estimate portfolio risk.

In-sample results on the 2003 to 2016 data set are presented in Appendix B. The appendix also includes some details of the final model specifications, illustrations of the estimated models, and comparisons between the models. The in-sample results are left as an appendix to allow the paper to focus on the forecasting ability of the models.

5.1 Evaluating Individual Distress Probabilities

We start by evaluating the models by their respective AUC. The AUC is a commonly used metric in prediction models. It measures the probability that a model places a higher risk on a random firm that experiences an event in a given year than a random firm that does not experience an event in a given year. Hence, 0.5 is random guessing and 1 is a perfect result.

Figure 1(a) shows the out-of-sample AUCs. In all years we find that the GB model gives the highest AUC and therefore is best at ranking firms by their distress risk, followed by the GAM and the GLM. This observation is consistent with the findings in Zięba et al. (2016) and Jones et al. (2017) in the sense that they also find that GB models are superior in terms of AUC. However, the differences we measure in AUCs



(a) Area under the ROC curve

(b) Log score

Figure 1: More complex models have higher AUC and better log scores. The figure shows performance measures of the three models (GLM ♦; GAM ▲; GB ●). Panel (a) shows out-of-sample area under the receiver operating characteristics curve (AUC) for the different models. Panel (b) illustrates the out-of-sample log scores of the three models. The figures above the center of the grey circles are the log scores for the GLM and the areas of the circles are proportional to the figures. The points show the log score of the model minus the log score of the GLM. That is, $\mathcal{L}_{tj} - \mathcal{L}_{t\text{GLM}}$ where $\mathcal{L}_{tj}$ is defined in Equation (9) and $j \in \{\text{GLM}, \text{GAM}, \text{GB model}\}$. The models are estimated on an expanding window of data with a 2-year gap to the forecasted data set. E.g., the models which are used to forecast the 2011 distresses are estimated on 2003-2009 data.

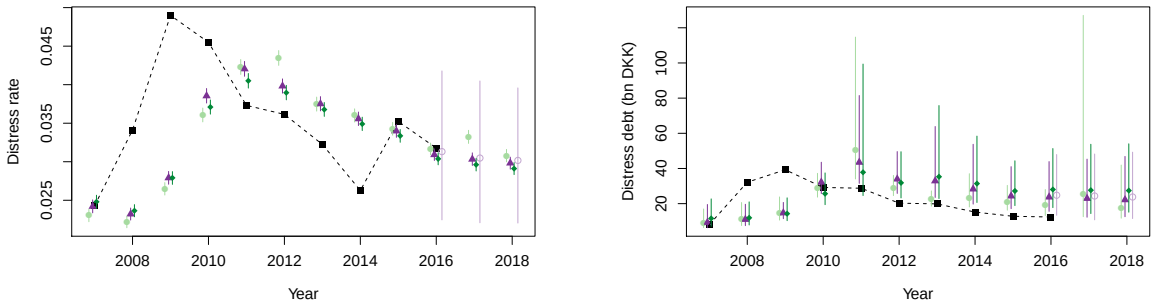
are much smaller than reported in the aforementioned papers. We find that the average AUC across years are 0.798, 0.811, and 0.822 for the GLM, GAM, and GB model respectively. Hence, there is an improvement in AUC between the GLM and the GB model of only 0.024. Comparably, Zięba et al. (2016) and Jones et al. (2017) find improvements in the AUC between a benchmark logistic regression and boosted tree models above 0.1. We reckon that the greater improvement in AUC is, to a large extent, due to the GLM used in Zięba et al. (2016) and Jones et al. (2017).[10]

As mentioned above, the AUC is only a ranking measure. A model may rank the firms well, but perform poorly in terms of the level of the predicted probabilities. We are interested in well calibrated probabilities as well. Thus, we look at the log score which is computed by

$$\mathcal{L}_{tj} = -\frac{1}{n_t} \sum_{i \in R_t} \left(y_{it} \log\left(\widehat{p}_{itj}\right) + (1 - y_{it}) \log\left(1 - \widehat{p}_{itj}\right) \right), \tag{9}$$

where $\mathcal{L}_{jt}$ is the log score of model j in year t, y_{it} is a dummy equal to 1 if firm i has an event in year t, $\widehat{p}_{itj}$ is the predicted probability of distress of firm i in year t by model j, R_t is the sample of active firms in year t, and n_t is the number of firms in R_t. A perfect score is zero. The out-of-sample log scores are illustrated in Figure 1(b). We find that the GB model outperforms the other models in all years. However, as with the AUCs, we find that the improvements in the log scores with more complex models are relatively small. The

[10]The data set used in Zięba et al. (2016) is publicly available. We can confirm that the results for the GLM can be greatly improved with limited effort.



(a) Realized and predicted distress rate

(b) Realized and predicted debt in distress

Figure 2: Models without frailty are unable to predict aggregated distress levels. The figures compare realized percentage of firms in distress (panel (a)) and realized debt in distress (panel (b)) to model predicted values (realized ■; GLM ♦; GAM ▲; GB ●; GLMM ○). The models are estimated on an expanding window of data with a 2-year gap to the forecasted data set. E.g., the models which are used to forecast the 2011 distresses are estimated on 2003-2009 data. The bars indicate simulated 90% confidence interval where outcomes are simulated using the predicted probabilities for each model.

numbers above the GLM figures are the log scores of the GLM model and illustrate that all models perform worst during the crisis in 2009-2010.

To summarize, we find evidence that the GB model is the best model at estimating individual default probabilities. However, the improvements are not large compared to the GAM. Thus, one may prefer the GAM model if interpretability is important.

5.2 Evaluating Aggregated Distress Probabilities

In this section, we look at the models' ability to predict the distress risk of the aggregated sample. Figure 2(a) shows the realized percentage of firms entering into distress as well as the out-of-sample predicted percentage of firms that will enter into distress each year for each of the models. All four models are included in the figure for later comparison, but for now we will only discuss results of the GLM, GAM, and GB model. It is clear that none of the models capture the distress level. Furthermore, none of the models' 90% confidence intervals have close to 90% coverage, which indicates that the assumed conditional independence assumption is violated, i.e. there is some correlation in distress events which is not accounted for in any of the models. That is, the complex GB model is just as bad at capturing the aggregated distress level as the more simplistic GLM. We run a formal test of the models' ability to estimate risk measures in Section 5.3.

The aggregated distress rate of the GB model in 2012 and 2017 is higher, and in 2012 further away from the realized value than the distress rate of the other models. This raises the question whether the GB model suffers from overfitting. However, it does not as we use cross validation to select the number of trees. Furthermore, the out-of-sample aggregate distress rates of the GB model are virtually the same as

the distress rates of the other models in all the other years, suggesting that the GB model is on aggregate similar to the other models. Finally, and perhaps most convincingly, we find no improvements in in-sample results of the GB model compared to the other models in terms of aggregate distress rates. An improvement would be expected in-sample in the case of overfitting.

The amount of debt varies greatly from firm to firm. The largest 21% of the firms have a size greater than 10 million DKK and account for 91% of the total debt in the economy. Thus, the percentage of firms in distress and the amount of debt in distress may differ greatly. Therefore, we also test how the models predict the amount of debt in distress. We compute the debt in distress each year as

$$\text{DiD}_t = \sum_{i \in R_t} y_{it} \left(\text{short debt}_{it} + \text{long debt}_{it} \right)$$

and the predicted debt in distress each year for all models as

$$\widehat{\text{DiD}}_{tj} = \sum_{i \in R_t} \widehat{p}_{itj} \left(\text{short debt}_{it} + \text{long debt}_{it} \right)$$

where DiD_t is an abbreviation for "debt in distress" in year t and $\text{short debt}_{it} + \text{long debt}_{it}$ is the total debt of firm i at time t.

Figure 2(b) shows results for the realized and out-of-sample predicted debt in distress. Similarly to the distress level results shown in Figure 2(a), we find that none of the models get near the actual level or have 90% confidence intervals with 90% coverage. However, the results here depend highly on a few number of firms. The 25 firms with the largest debt on their balance sheet in 2018 account for 28.47% of the debt. Thus, Figure 2(b) essentially reflects a non-trivial probability of default for some of these firms. As seen by Figure 2(b), frailty (the GLMM) has little impact with such unequal distributions of exposures. However, we do not expect such unequal distribution of exposures in, say, a bank's loan portfolio.

5.3 Measuring Portfolio Risk Without Frailty

Above we illustrated that all models fail to capture the percentage of firms entering into distress in the next period. In this section we explore this further by examining the models' ability to evaluate portfolio risks. Specifically, we compare the models' predicted 95% Value-at-Risk (VaR), which is the upper bound in the 90% confidence intervals, to the realized distress rate. If the 95% VaR figures are accurate, we will find that the realized distress rate is below the VaR figure about 95% of the cases and above about 5% of the cases.

We use bank connections reported by the firms themselves to construct portfolios for each year and bank. If a firm indicates two bank connections, the firm will appear in the portfolio of both banks. We only include

Table 1: Likelihood ratio test for coverage of the out-of-sample VaR figures. We form four portfolios of firms representing bank exposures for each calender year yielding 40 portfolios in total. For each portfolio we compute the 95% out-of-sample VaR figure for the distress rate in each of the three different models and perform a test where the null hypothesis is that the VaRs have the correct coverage level. The "asymptotic p-value" is the p-value from the test in Kupiec (1995) and the "MC p -value" is the monte carlo corrected p-values used in Berkowitz et al. (2011).

Model	Likelihood ratio	Asymptotic p-value	MC p-value
GLM	49.670	< 0.0000001	< 0.0000001
GAM	25.901	0.0000004	0.0000004
GB	18.005	0.0000220	0.0000190

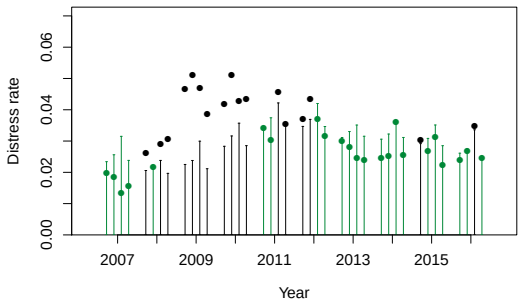
banks with at least 500 connections to ensure that the portfolio is somewhat diversified. Four banks fulfill this requirement. The smallest and largest number of connections for a given bank and year are 534 and 5 063 firms and the mean number of connections is 2 196. We track the four banks through 10 years resulting in a total of 40 portfolios.

The portfolios we have constructed are only a rough proxy for the exposure of the banks in the Danish economy. Thus, this exercise should be seen as an example of none-random portfolios rather than as representing the lending risk of the Danish banks.

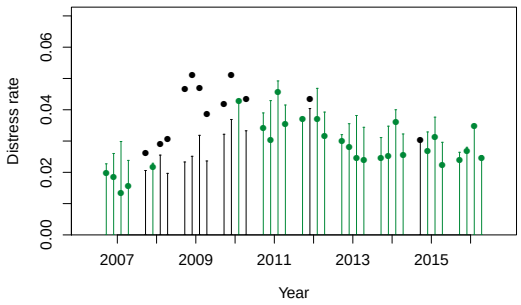
We estimate the out-of-sample 95% VaR figure of the distress rate in each of the portfolios assuming the GLM, GAM, and GB model respectively and test the coverage of the VaR figures. Table 1 reports results of the VaR coverage test introduced by Kupiec (1995) and the Monte Carlo correction from Berkowitz et al. (2011). We reject the null hypothesis that the coverage has the correct level for all models at a 1% significance level with both the asymptotic p-values and finite sample Monte Carlo corrected p-values. That is, we can statistically reject that any of the models including the GB model are able to estimate accurate risk measures.

Figure 3 illustrates when the realized values are above the VaR figures for each of the portfolios. The vertical lines represents VaR figures. The lines are green when the realized distress rate is below the VaR figure and black when the realized value is above the VaR figure. The GLM has 17 VaR breaches, the GAM has 12 VaR breaches, and the GB model has 10 VaR breaches. Most breaches occur in 2008-2009.

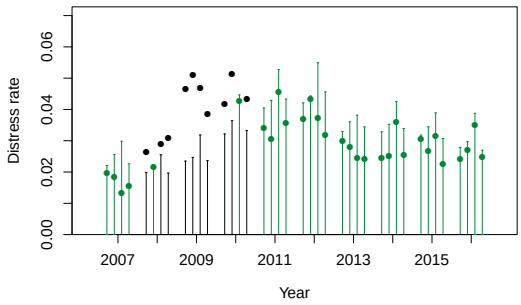
The models' inability to capture the time-varying distress level and the lack of coverage of the VaR figures is a sign that the models are misspecified. In order to mitigate this we implement a mixed model in the next section which allows for a random intercept.



(a) 95% VaRs of the distress rate in the GLM



(b) 95% VaRs of the distress rate in the GAM

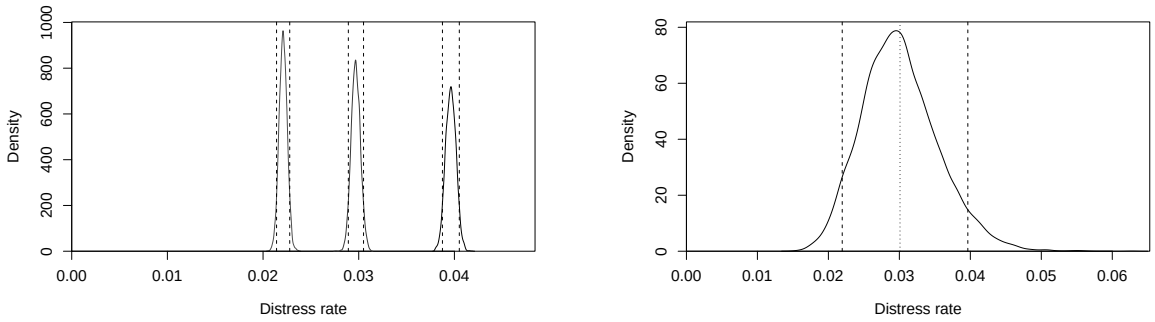


(c) 95% VaRs of the distress rate in the GB model

Figure 3: Models without frailty estimate too low VaR figures. We form bank portfolios based on self-reported bank connections in the firms' financial statements. For each portfolio we compute the 95% VaR figure by simulation, using the out-of-sample predicted firm probabilities of distress. Panel (a), (b), and (c) show the VaRs for the GLM, GAM, and the GB model respectively. The dots show the realized level, bars show the VaR figures. Black bars and dots indicate years where the realized level is not covered by the confidence interval.

6 Modelling Frailty in Distresses with a Generalized Linear Mixed Model

In this section we estimate a generalized linear mixed model (GLMM) introduced in Section 3.4 with a random intercept to relax the conditional independence assumption we have assumed so far. That is, the model allows for an unobservable macro effect and thus creates correlation in distresses beyond the observed covariates. Furthermore, we add non-penalized natural cubic regression splines to the model given the higher AUC and lower Akaike information criterion (AIC) of the GAM compared to the GLM (see Appendix B.1 for the latter). While several others have implemented GLMM with random intercept (e.g., see Duffie et al. 2009), we differ by including non-linear effects. We use non-penalized splines as software allowing for penalized splines in a GLMM is not readily accessible to us. Furthermore, we expect a minor difference between a penalized and a non-penalized model due to our large sample.



(a) Density plot with fixed random effect

(b) Density plot without fixed random effect

Figure 4: Density plots of the GLMM forecasted 2018 distress rate. We estimate the GLMM on 2003-2016 data and simulate densities of the predicted cross-sectional distress rate in 2018. In panel (a) the random effect is fixed at the 5%, 50%, and 95% quantile. The three quantiles can be seen as a "good", "middle", and "bad" future state of the unobservable macro effect in 2018. The tall density curves and narrow confidence intervals are consistent with what a model without a random intercept would predict. Panel (b) shows a density curve estimate where we simulate both the random intercept term and the outcomes. The outer dashed lines are 5% and 95% quantiles and the inner line is the mean.

The estimated standard deviation of the random intercept is $\widehat{\sigma} = 0.196$ when estimated on the 2003–2016 data set. That is, a change of one standard deviation in the random intercept implies an $\exp(0.196) = 1.217$ times higher odds of entering into distress for all firms. Thus, there is a non-negligible random effect. A conservative likelihood ratio test for $H_0 : \sigma = 0$ is rejected with a test statistics of 1 483 which should be compared to a χ^2 distribution with 1 degree of freedom.[11] Thus, we can reject the conditional independence assumption. We end this section by illustrating what can go wrong if one relies on a model that does not account for the observed correlation in distresses.

6.1 Predictive Results of the GLMM

Figure 4 shows a forecast for the 2018 distress rate and illustrates how adding a random intercept to the model affects the confidence bounds of the distress rate. Panel (a) of the figure shows the 2018 forecasts of the distress rate assuming that the random effect is fixed at three different quantiles of its estimated distribution, the 5%, 50%, and 95% quantile. The three quantiles can be seen as a "good", "middle", and "bad" future state of the unobservable macro effect in 2018. Panel (b) of the figure shows the unconditional 2018 forecast density of the distress rate (i.e. without fixing the random intercept). The width of the confidence interval is much wider than that of the GLM, GAM, and GB model (see Figure 2(a)). Whereas the width of the confidence interval, when the random effect is assumed to take a specific value, is of the same magnitude as in the GLM, GAM, and GB model. The large effect of the random intercept on the confidence bounds is

[11] The p-value is likely conservative (e.g., see the simulations in Pinheiro and Bates 2000). Though, it does not matter in this case since the p-value is essentially zero already.

similar to what Duffie et al. (2009) find[12] and reflects the large estimated standard deviation of the random effect.

The GLMM requires a relatively long estimation period, hence we can only backtest results of the GLMM in 2016 as in Section 5. We will compare these results to results of the GAM in the following, though similar conclusions can be made for the GLM and GB model. In 2016 we find an AUC of 0.815 in the GLMM, which is close to the 0.818 AUC of the GAM we find in 2016. Hence, we find evidence that the GLMM is equally good at ranking the firms in terms of riskiness.

The 90% confidence interval of the distress rate in 2016 is $[0.0220, 0.0396]$ in the GLMM, while we estimated the same interval to be $[0.0302, 0.0317]$ in the GAM. The realized distress rate in 2016 was 0.0318. That is, the realized distress rate is not included in the confidence interval of the GAM while it is included in the confidence interval of the GLMM. Furthermore, the 2016 confidence intervals of the GLMM predicted debt in distress and the GAM predicted debt in distress are $[11.52, 49.28]$ and $[15.60, 43.93]$ respectively. The realized debt in distress in 2016 is 12.40 billion DKK.

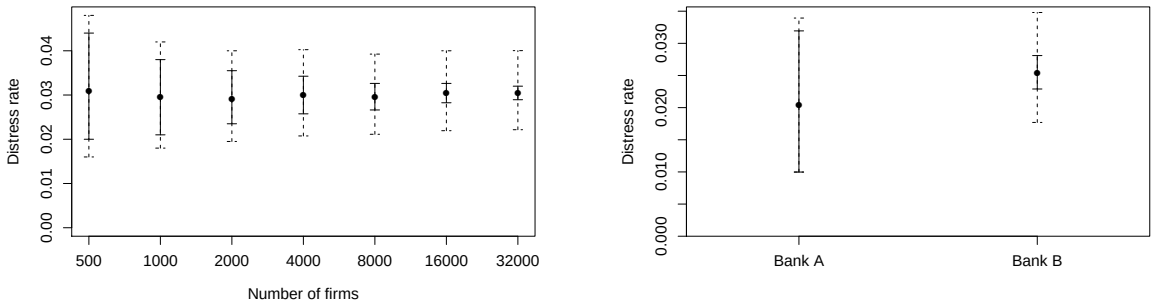
The confidence intervals of the GLMM and GAM are both illustrated in Figure 2. The confidence intervals of the two models are much more similar in the debt in distress example than in the distressed rate example. Again, this is due to a few firms in the sample with large debt, implying that a portfolio of firm debt is less diversified. The connection between portfolio diversification and the confidence intervals is explained in the following section.

6.2 Frailty Models and Portfolio Risk

Accounting for frailty is more important for some portfolios with distress risk than others. Particularly, adding a frailty to a model matters more for portfolios with many exposures of equal size. To illustrate this point, we randomly sample firms that are active on January 1, 2018 (as defined in Section 4.1) into portfolios of sizes ranging from 500 to 32 000. Thus, some portfolios are much more diversified than others, which means that confidence intervals of the predicted distress rate will vary.

For each portfolio, we then compute the distress rate using the estimated GLMM and simulate 90% confidence intervals of the distress rate. First, we ignore the frailty component by integrating out the random effect in the firm-specific distress probabilities and draw the firm-outcomes independently using these probabilities. Secondly, a simulation is done where we account for the frailty component by first drawing the random effects from its estimated distribution, compute the firm probabilities conditional on the drawn random effect, and then draw the firm outcomes conditional on these probabilities. The second method is the same as the one used for the simulated confidence intervals in Figure 4(b), and the width of

[12] See Figure 5 of the paper.



(a) Random portfolio example

(b) Example with reversed ranking

Figure 5: Frailty matters more for large portfolios. In panel (a), we split the sample of firms that are active on January 1, 2018 into portfolios of size $500, 2 \cdot 500, \ldots, 32\,000$ and compute the distress probability based on the GLMM estimated on the 2003-2016 sample. The dots are the expected unconditional distress rate of the portfolios. The solid lines are the simulated 90% confidence interval where we integrate the random effect out on a firm-by-firm level and then simulate the outcomes independently. The dashed lines are the simulated confidence interval when we do account for frailty. Panel (b) shows 90% confidence intervals of the distress risk of loan portfolios of two banks. Bank A has 501 clients with distress probabilities evenly distributed on the interval $[0.10, 0.30]$ in the case of the GLMM where the random effect is equal to zero. Likewise, Bank B has 10 001 clients with distress probabilities evenly distributed on the interval $[0.15, 0.35]$ when the random effect is zero. The solid and dashed lines are confidence intervals simulated in a model without and with frailty respectively.

the confidence intervals of the model without frailty is very similar to the width of the confidence intervals in Figure 4(a), which again is very similar to the confidence intervals of the GB model.

The results are shown in Figure 5(a). The figure illustrates that the tail risk is generally underestimated when we do not account for frailty. However, the discrepancy between the two models is much more pronounced for the large portfolios than for the small. This is because the model without frailty drastically shrinks the confidence intervals of the more diversified large portfolios. The confidence intervals of the model with frailty are also affected when the portfolio becomes more diversified, but to a much smaller extent. This is because the frailty model takes into account the correlation in distresses and thereby treats the large portfolios as less diversified. An economist relying on a model without frailty could then easily conclude that a well diversified portfolio is much safer than what it is in reality. How this can lead to misperception of portfolio risk of two banks with different strategies is illustrated in the following example.

Assume that we have two banks: Bank A has a few safe clients and Bank B has many relatively more risky clients. Specifically, we let Bank A have 501 clients with distress probabilities evenly distributed on the interval $[0.10, 0.30]$ in the case of the GLMM where the random effect is equal to zero. Bank B has 10 001 clients with distress probabilities evenly distributed on the interval $[0.15, 0.35]$ when the random effect is zero. Appendix C provides further details regarding the simulation of the two bank portfolios. The confidence intervals of the distress rate with and without accounting for frailty are illustrated in Figure 5(b).

The VaR figures of Bank A and Bank B are 0.0319 and 0.0281 respectively, if we do not account for frailty. Thus, Bank A appears more risky by this metric. However, the correct figures – the ones where we account for frailty – are 0.0339 and 0.0348 respectively. Hence, Bank B has the highest risk in reality. Thus, if one relies on a model without frailty, one might wrongly assume that a large bank is exposed to relatively little risk.

7 Robustness

The GLM, GAM, and GB models implemented in this paper are all unable to produce confidence intervals of next period's default rate that accurately capture realized values. In this section, we show that the low coverage of the confidence intervals are not due to: (1) not accounting for parameter uncertainty in the estimated confidence bounds and (2) not including macro variables in the models.

7.1 Accounting for Parameter Uncertainty

We have not considered parameter uncertainty in the confidence bounds of the distress rate or the debt-in-distress we have shown up to this point. To do so, we return to the GLM in Section 3.1 estimated with the model on the 2003–2016 data set and make a forecast for the debt in distress in 2018. We compute the observed Fisher information matrix $I(\boldsymbol{\beta}) = \mathbf{X}^\top \mathbf{W}(\boldsymbol{\beta})\mathbf{X}$, where $\mathbf{W}$ is the diagonal matrix with working weights as in Section 3.2,[13] we have made the dependence on $\boldsymbol{\beta}$ explicit, and $\mathbf{X}$ is the stacked design matrices. Then we use the large sample approximation

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N\left(\mathbf{0}, I(\boldsymbol{\beta})^{-1}\right)$$

to sample $\widetilde{\boldsymbol{\beta}} \sim N(\widehat{\boldsymbol{\beta}}, I(\widehat{\boldsymbol{\beta}})^{-1})$ and simulate the outcomes as in Section 5 but using different $\widetilde{\boldsymbol{\beta}}$ instead of the single estimate $\widehat{\boldsymbol{\beta}}$. See Wood (2017) for a similar approach for GAMs.

The 90% confidence interval is $[15.10, 53.60]$ billion DKK when we simulate without accounting for parameter uncertainty as in Section 5, and $[15.10, 53.64]$ if we account for parameter uncertainty. Similar figures for the distress rate are $[0.02831, 0.02990]$ and $[0.02826, 0.02999]$. That is, the confidence intervals remain practically unchanged when parameter uncertainty is accounted for. The uncertainty of the parameters is estimated assuming that observations are conditionally independent, which is already assumed in the GLM, GAM, and GB model. The uncertainty could potentially increase if the conditional independence assumption was relaxed as we do in the GLMM.

[13] The matrix $\mathbf{X}$ does not include columns for the spline functions as in Section 3.2 since we use a GLM.

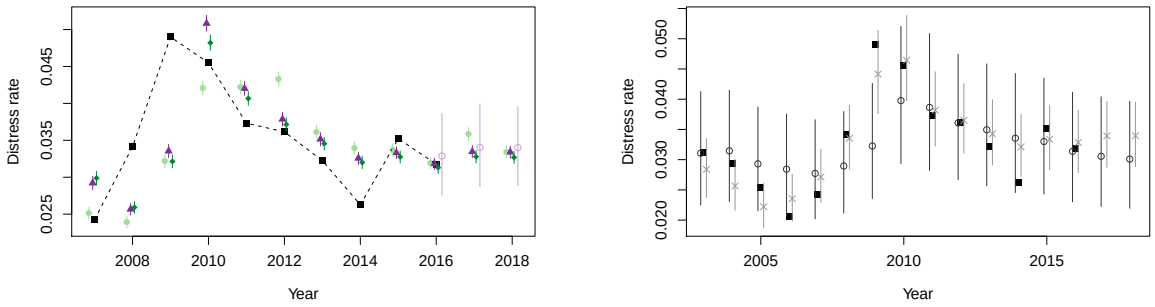
Besides estimating uncertainties of the slopes we also estimate the uncertainty of the standard deviation of the random intercept, $\widehat{\sigma}$, in the GLMM. The uncertainty of $\widehat{\sigma}$ is largely due to the short time series. A 95% profile likelihood-based confidence interval for $\widehat{\sigma}$ is $[0.155, 0.332]$. The large uncertainty of $\widehat{\sigma}$ implies that the confidence bounds of the distress rate in the GLMM could be considerably more narrow or wider.

7.2 Including Macro Variables in the Models

The models implemented so far are based solely on micro level data. While the models are good at ranking the individual firms by riskiness they are far from good at estimating the aggregated distress rate in the next period. This raises the question whether the models could be improved by including some macro variables. In this section, we show results of models including macro variables, we argue that we may have a potential candidate for macro variable but estimating an effect of the covariate may be hard with the few number of cross-sections we have. Lastly, we show that the random effect is still needed in-sample even after the inclusion of the random effect.

Some common macro variables in the existing literature are return of the S&P 500 index, 3-month treasury rate, 10-year treasury rate, inflation, GDP growth, and unemployment rate (e.g., see Duffie et al. 2007, Das et al. 2007, Duffie et al. 2009, Chava et al. 2011, Duan et al. 2012, Lando et al. 2013). We include the Danish equivalent of these variables in our models, except for the stock index return since the majority of firms in our sample are non-traded, and test if the models' predictions improve. We lag all macro variables to ensure predictability. Furthermore, we use a swap rate, a short-term, and a long-term interbank rate instead of treasury rates, since the Danish government bond market is much smaller than the U.S. Finally, we include the GDP gap instead of the GDP growth as GDP gap has been included in earlier versions of the Danish central bank's internal corporate distress model. While some of the aforementioned papers track events on a quarterly or monthly basis, we choose to do so only on a yearly basis. This is because the start date of the "in distress" status can be somewhat arbitrary and reflects a potentially delayed processing time of the authorities. Thus, we end up with relatively few observations in the time dimension, implying that we can include at most one macro variable in our models.

We run separate logistic regressions including each of the macro variables one at a time and find that the model with the unemployment rate has the lowest AIC. We then include the unemployment rate in all four models and run predictive tests. The inclusion of the unemployment rate in the GB model is done by estimating a logistic regression with two covariates: the unemployment rate and the linear predictor from the estimated GB model without the unemployment rate. The motivation for the two-step model is that we can control the complexity of the unemployment rate. This turned out to be an issue in some preliminary



(a) Realized and predicted distress rate

(b) In-sample predicted distress rate

Figure 6: Results of models with the unemployment rate. We re-estimate all four models adding the unemployment rate as a covariate. Panel (a) shows the realized distress rate together with the out-of-sample predicted values (realized ■; GLM ♦; GAM ▲; GB ●; GLMM ○). The bars indicate the simulated 90% confidence interval where outcomes are simulated using the predicted probabilities for each model respectively. Panel (b) shows the predicted distress with and without the unemployment rate along 90% confidence intervals for the GLMM with parameters estimated on the full sample (realized ■; GLMM without unemployment rate ○; GLMM with unemployment rate ×).

results where a GB model including the unemployment rate as a covariate generalized poorly.

We find improvements in the out-of-sample forecasts when the unemployment rate is included in the models (see Figure 2(a) and Figure 6(a)). However, Figure 6(a) still shows too narrow confidence bounds for the GLM, GAM and GB model. The standard deviation of the random intercept of the GLMM estimated in the period 2003 to 2016 is reduced from 0.196 in a model without the unemployment rate to 0.106 in a model with the unemployment rate. The reduction shows that the unemployment rate explains some of the yearly fluctuations. This is also evident from Figure 6(b) which shows a much better in-sample predicted distress rate for the GLMM with the unemployment rate. However, the random intercept remains significant with a chi-square test-statistic of 297 with 1 degree of freedom.

The estimated slope on the unemployment rate is negative and statistically significant, which may seem counter-intuitive. Furthermore, the slope estimate varies a lot during the first out-of-sample forecasts, which is not surprising given the low number of cross-sections included in this sample. One major question is whether we will see the same in the future. I.e. if the association we estimate now will generalize. This is particularly questionable given that we have already considered five potential macro variables with only 14 cross-sections. However, we also estimate a negative slope for the unemployment of the same size on aggregate defaults for which we have data going back to 1980.[14] This provides evidence that the effect we estimate may generalize. Whether the estimated slope on unemployment generalizes or not does not change the fact that the random intercept remains significant, i.e. we cannot avoid a frailty component.

[14]We use the number VAT registered firms (Danmarks Statistik 2018a) as the denominator and the number of defaults (Danmarks Statistik 2018b) as the numerator in a binomial regression model.

8 Conclusion

We have shown that gradient tree boosting performs better in out-of-sample ranking of firms in terms of riskiness compared to more traditional statistical models in a sample containing the majority of Danish limited liability firms. However, the improvement is only minor compared to what recent papers find. Furthermore, the out-of-sample tests yield too narrow confidence bounds of the aggregated distress rate for both traditional statistical models and the gradient boosted tree model. That is, the more complex model is not better at capturing correlation in defaults across the cross-section of firms. The lack of correlation leads to too small risk measures for individuals, firms, or regulators who evaluate the riskiness of a portfolio exposed to multiple firms. We show how to relax this assumption with a generalized linear mixed model. We including non-linear dependency structures between some of the covariates in the model and the dependent variable, thereby obtaining competitive firm level performance.

References

Azizpour, S., K. Giesecke, and G. Schwenkler (2018). Exploring the sources of default clustering. *Journal of Financial Economics 129*(1), 154–183.

Bates, D. M. and S. DebRoy (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis 91*(1), 1–17.

Bates, D. M., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Beaver, W., M. McNichols, and J. Rhie (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies 10*(1), 93–122.

Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry 23*(2), 129–143.

Berkowitz, J., P. Christoffersen, and D. Pelletier (2011). Evaluating value-at-risk models with desk-level data. *Management Science 57*(12), 2213–2227.

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science 22*(4), 477–505.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Campbell, J. Y., J. Hilscher, and J. Szilagyi (2008). In search of distress risk. *The Journal of Finance 63*(6), 2899–2939.

Caruana, R. and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, pp. 161–168.

Chava, S. and R. A. Jarrow (2004). Bankruptcy prediction with industry effects. *Review of Finance 8*(4), 537–569.

Chava, S., C. Stefanescu, and S. Turnbull (2011). Modeling the loss distribution. *Management Science 57*(7), 1267–1287.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 785–794. ACM.

Danmarks Statistik (2018a). Fiks9: Firmaernes køb og salg, historisk sammendrag efter beløb. Retrieved September 2018, http://www.statistikbanken.dk/FIKS9.

Danmarks Statistik (2018b). Konk9: Erklærede konkurser (historisk sammendrag). Retrieved September 2018, http://www.statistikbanken.dk/KONK9.

Das, S. R., D. Duffie, N. Kapadia, and L. Saita (2007). Common failings: How corporate defaults are correlated. *The Journal of Finance 62*(1), 93–117.

Duan, J.-C. and A. Fulop (2013). Multiperiod corporate default prediction with the partially-conditioned forward intensity. Working paper.

Duan, J.-C., J. Sun, and T. Wang (2012). Multiperiod corporate default prediction—a forward intensity approach. *Journal of Econometrics 170*(1), 191–209.

Duffie, D., A. Eckner, G. Horel, and L. Saita (2009). Frailty correlated default. *The Journal of Finance 64*(5), 2089–2123.

Duffie, D., L. Saita, and K. Wang (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics 83*(3), 635–665.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*(5), 1189–1232.

Giesecke, K. and B. Kim (2011). Systemic risk: What defaults are telling us. *Management Science 57*(8), 1387–1405.

Jones, S., D. Johnstone, and R. Wilson (2017). Predicting corporate bankruptcy: An evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting 44*(1-2), 3–34.

Kim, M.-J. and D.-K. Kang (2010). Ensemble with neural networks for bankruptcy prediction. *Expert Systems with Applications 37*(4), 3373–3379.

Koopman, S. J., A. Lucas, and B. Schwaab (2011). Modeling frailty-correlated defaults using many macroeconomic covariates. *Journal of Econometrics 162*(2), 312–325.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives 3*, 73–84.

Kwon, T. Y. and Y. Lee (2018). Industry specific defaults. *Journal of Empirical Finance 45*, 45–58.

Lando, D., M. Medhat, M. S. Nielsen, and S. F. Nielsen (2013). Additive intensity regression models in corporate default analysis. *Journal of Financial Econometrics 11*(3), 443–485.

Lin, W. Y., Y. H. Hu, and C. F. Tsai (2012). Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42*(4), 421–436.

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*(4), 603–614.

Natekin, A. and A. Knoll (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics 7*, 21.

Nickerson, J. and J. M. Griffin (2017). Debt correlations in the wake of the financial crisis: What are appropriate default correlations for structured products? *Journal of Financial Economics 125*(3), 454–474.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-effects models in S and S-PLUS*. Springer-Verlag New York.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business 74*(1), 101–124.

Sun, J., M. Y. Jia, and H. Li (2011). Adaboost ensemble for financial distress prediction: An empirical comparison with data from chinese listed companies. *Expert Systems with Applications 38*(8), 9305–9312.

Tinoco, M. H. and N. Wilson (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis 30*, 394–419.

Vaupel, J. W., K. G. Manton, and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*(3), 439–454.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (2 ed.). Chapman and Hall/CRC.

Wood, S. N., Y. Goude, and S. Shaw (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 64*(1), 139–155.

Zhou, S. (2010). Thresholded lasso for high dimensional variable selection and statistical estimation. Working paper.

Zięba, M., S. K. Tomczak, and J. M. Tomczak (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications 58*, 93–101.

A Variable Selection with Lasso

We use the so-called Thresholded Lasso estimator to perform variable selection. The Thresholded Lasso estimator is found by the following steps.

1. Standardize the covariates.
2. Perform K-fold cross-validation to find the penalty variable, λ_{init}, that maximizes

$$\lambda_{\text{init}} = \arg\max_{\lambda} \max_{\boldsymbol{\beta}} \sum_{t=1}^{d} \sum_{i \in R_t} y_{it} \boldsymbol{\beta}^\top \boldsymbol{x}_{it} - \log\left(1 + \exp\left(\boldsymbol{\beta}^\top \boldsymbol{x}_{it}\right)\right) - \lambda \|\boldsymbol{\beta}\|_1$$

   where $\|\cdot\|_1$ is the L1 norm. This is the log likelihood from the multiperiod logit model in Equation (1) with an added L1 penalty.
3. Denote $\widehat{\boldsymbol{\beta}}_{\text{init}}$ as the estimated coefficients with penalty parameter λ_{init} and define the sets $S(\delta) = \{j : |\beta_{\text{init},j}| > \delta\}$. Then use K-fold cross-validation to find a threshold value, $\widehat{\delta}$, in the range that maximizes

$$\arg\max_{\delta} \max_{\boldsymbol{\beta}_{S(\delta)}} \sum_{t=1}^{d} \sum_{i \in R_t} y_{it} \boldsymbol{\beta}_{S(\delta)}^\top \boldsymbol{x}_{S(\delta),it} - \log\left(1 + \exp\left(\boldsymbol{\beta}_{S(\delta)}^\top \boldsymbol{x}_{S(\delta),it}\right)\right)$$

   where $\boldsymbol{\beta}_{S(\delta)}^\top \boldsymbol{x}_{S(\delta),it}$ is the linear predictors which only include the covariates in the index set $S(\delta)$. This amounts to fitting a GLM with a subset of the covariates.

Step 1 and 2 yield the common Lasso estimator $\widehat{\boldsymbol{\beta}}_{\text{init}}$. That is, we add an L1 penalty which shrinks parameters and discards variables where the coefficient is shrunk to 0. Step 3, in addition to the previous, yields the Thresholded Lasso estimator, where we discard any variables where the coefficient is below the threshold $\hat{\delta}$. The final estimates are no longer shrunk as we do not apply a penalty. The motivation to use the Thresholded Lasso estimator rather then the Lasso estimator is to addresses the bias problems with $\widehat{\boldsymbol{\beta}}_{\text{init}}$. See Zhou (2010), Bühlmann and Van De Geer (2011) for properties of the Thresholded Lasso estimator.

We end with the 6 categorical and 44 numerical covariates listed in Table 2. The numerical variables are divided by firm size when appropriate and all are winsorized at the 5% and 95% quantile. We exclude 3 numeric covariates in the Lasso estimation while none of the categorical covariates considered are dropped.

Table 2: Summary statistics for covariates in the data set from 2003 to 2016. Variables divided by size are in percentages. Size is the maximum of total asset and total debt. The statistics are computed after winsorizing. There is 1.3 million firm year observations. Panel A shows the numerical covariates that are left after variable selection with the Thresholded Lasso method and the estimated coefficients where stars indicate the significance of the effect with a Wald test (*** is 1% significance, ** is 5%, and * is 10%). Panel B shows the numerical variables that are excluded after variable selection. Panel C shows the binary and categorical covariates included in the model. Panel D shows variable descriptions of some of the covariates.

Covariate	Mean	Median	St. Dev.	Min	Max	GLM coefficient estimates
Panel A: Numerical covariates included after variable selection						
Accounts payable / size	8.06	2.83	10.99	0.00	38.00	0.0246^{***}
Accounts receivable / size	12.75	3.34	16.88	0.00	54.00	-0.0097^{***}
Change in log size	0.03	0.00	0.24	-0.46	0.58	-0.0901^{**}
Corporation tax / size	1.12	0.00	2.18	0.00	7.60	0.0708^{***}
Current assets / size	58.41	66.07	35.50	1.00	100.00	-0.0025^{***}
Deferred tax / size	1.16	0.00	2.31	0.00	8.10	-0.0688^{***}
Depreciation / size	-3.08	-1.11	4.18	-14.00	0.00	0.0140^{***}
EBIT / size	4.19	3.44	17.44	-36.00	40.00	-0.0036^{***}
Equity / invested capital	6.16	2.27	10.05	-3.90	38.00	-0.0255^{***}
Equity / size	33.78	32.32	38.10	-48.00	96.00	0.0032^{***}
Expected dividends / size	1.59	0.00	4.08	0.00	15.60	-0.0968^{***}
Financial assets / size	6.10	0.00	14.71	0.00	58.00	-0.0117^{***}
Financial income / size	0.99	0.17	1.63	0.00	5.80	-0.0531^{***}
Financing costs / size	2.22	1.54	2.25	0.00	7.40	0.0479^{***}
Fixed costs / size	-44.96	-25.09	52.16	-175.00	0.00	0.0002
Immaterial fixed assets / size	1.73	0.00	4.82	0.00	19.00	-0.0169^{***}
Ind. EW avg. net profit / size	2.03	2.11	2.94	-39.00	34.00	-0.0279^{***}
Interest coverage ratio	0.02	-0.71	21.75	-47.00	48.00	-0.0003
Inventory / size	9.09	0.00	16.45	0.00	56.00	-0.0052^{***}
Invested capital / size	20.16	9.40	25.77	0.90	97.00	0.0009^{**}
Land and buildings / size	16.04	0.00	31.17	0.00	95.00	-0.0076^{***}
Liquid assets / size	14.94	3.51	21.69	0.00	75.00	-0.0131^{***}
log(age)	1.98	2.08	1.16	0.00	4.60	-0.2965^{***}
log(size)	7.85	7.82	1.61	4.95	10.91	0.0133^{*}
Long-term bank debt / size	2.60	0.00	6.98	0.00	26.00	0.0110^{***}
Long-term debt / size	11.65	0.00	20.49	0.00	66.00	0.0026^{***}
Long-term mortgage debt / size	5.20	0.00	13.24	0.00	47.00	0.0077^{***}
Net profit / size	2.04	2.16	16.52	-39.00	34.00	-0.0066^{***}
Other operating expenses / size	-2.24	0.00	6.08	-23.00	0.00	-0.0094^{***}
Other receivables / size	4.33	0.97	7.15	0.00	26.00	-0.0050^{***}
Other short debts / size	13.79	8.58	15.01	0.00	53.00	0.0114^{***}
Personnel costs / size	-34.28	-10.05	45.82	-151.00	0.00	0.0015^{***}
Prepayments / size	0.52	0.00	0.96	0.00	3.40	-0.0659^{***}
Provisions / size	1.34	0.00	2.60	0.00	9.20	0.0096^{*}
Quick ratio	2.35	0.98	3.86	0.00	16.00	-0.0040
Receivables from related parties / size	5.61	0.00	12.99	0.00	49.00	0.0014^{**}
Relative debt change	0.10	0.00	0.48	-0.62	1.50	-0.0621^{**}
Retained earnings / size	6.20	6.73	38.71	-91.00	72.00	-0.0049^{***}
Return on equity (pct.)	-1.05	0.12	4.95	-19.40	3.60	-0.0175^{***}
Short-term bank debt / size	7.32	0.00	13.19	0.00	44.00	0.0109^{***}
Short-term mortgage debt / size	0.12	0.00	0.38	0.00	1.50	-0.0223
Tangible fixed assets / size	26.17	9.44	32.59	0.00	96.00	-0.0040^{***}
Tax expenses / size	-1.68	-0.44	3.51	-10.30	3.80	-0.0080^{***}
Total receivables / size	26.91	18.69	26.82	0.00	90.00	0.0037^{***}
Panel B: Numerical covariates excluded after variable selection						
Current ratio	2.58	1.21	3.87	0.00	16.00	
max(equity + provisions, 0) / size	39.74	34.79	31.88	0.00	100.00	
Short-term debt / size	47.94	45.87	32.11	1.80	100.00	
Panel C: Categorical covariates						
Is non-stock based	0.73	1.00	0.45			0.3277^{***}
Has prior distress	0.03	0.00	0.16			0.9542^{***}
Large debt change	0.08	0.00	0.27			0.1936^{***}
Negative equity	0.15	0.00	0.36			0.1770^{***}
Region						
Sector						

Continued on next page

Table 2 – *Continued from previous page*

Panel D: Variable description	
Change in log size	The log of firm size as reported in the current financial account minus the log of firm size as reported in the financial account from the previous year. We use the size definition in Equation (8). The variable is set to zero if the firm did not hand in a financial account the previous year.
Current ratio	Current assets divided with short-term debt. If the short-term debt is zero or below 10 000 DKK we divide by 10 000 instead to avoid dividing with zero.
Is non-stock based	A dummy variable equal to 1 if the firm is non-stock based ("Anpartsselskab"). The alternative is a stock-based firm ("Aktieselskab").
Has prior distress	A dummy variable equal to 1 if the firm has previously been "in distress".
Ind. EW avg. net profit	We group firms by their 3-digit SIC code and compute the equally weighted average net profit of each group each year.
Interest coverage ratio	Net profit divided by net financial revenue. If the net financial revenue is zero, we divide by 1 instead.
Large debt change	A dummy variable equal to 1 if the total debt grew more than 100% in the past year. It is zero if the firm did not hand in a financial account the previous year.
Negative equity	A dummy variable equal to 1 if equity is negative.
Region	The firms are grouped based on the location of their headquarter into 5 geographical regions, going from the most to the least densely populated areas.
Relative debt change	The firm's total debt of the current financial account divided by the total debt of the financial account from the previous year. The variable is set to zero if the firm did not hand in a financial account the previous year.
Return on equity	Net profit divided by equity. If equity is zero or below 10 000 DKK we divide by 10 000 instead to avoid dividing with zero.
Sector	The firms are grouped into 7 general sectors: Construction; industrial; farming and fishing; trade; transport; information; real estate; other.
Quick ratio	Current assets minus inventories divided with short-term debt. If the short-term debt is zero or below 10 000 DKK we divide by 10 000 instead to avoid dividing with zero.

B Model Estimation

The estimated coefficients of the covariates included in the GLM after variable selection are listed in Table 2. Figure 7 shows the largest absolute standardized coefficients. Most noticeably, we find a large age effect unlike Shumway (2001). This is not surprising given that we use the age since incorporation for some potentially small and risky firms whereas Shumway (2001) uses the age since listing for large corporate firms. The industry specific covariate mentioned in Section 4.2 has a coefficient estimate of -0.02791 with a standard error of 0.00176. The negative sign is consistent with the results in Chava et al. (2011) who find a higher likelihood of a default when the median stock performance in an industry is below 20%.

The GLM uses 61 degrees of freedom whereas the GAM uses 254.7. Hence, the GAM is much more complex than the GLM.[15] In-sample estimations on the full sample period (2003-2016) yield Akaike information criterion (AIC) of the GLM and GAM at 327 625 and 321 388 respectively, which shows an improvement from the GLM to the GAM in spite of the increased complexity of the model.

B.1 Non-linear Effects in the GAM and the GLMM

The GAM has 11 non-linear effects of which nine have interactions. We include 6 non-linear effects in the GLMM model with only 2 non-linear interaction. Table 3 shows which non-linear effects are included. Furthermore, in the GAM, we use 6 to 20 dimensional basis for each (marginal) spline while we only use 5 in

[15] We use the effective degrees freedom, which is $\mathrm{tr}\left(\mathbf{F}_{\boldsymbol{\lambda}}\right)$ from Equation (4) for the GAM.

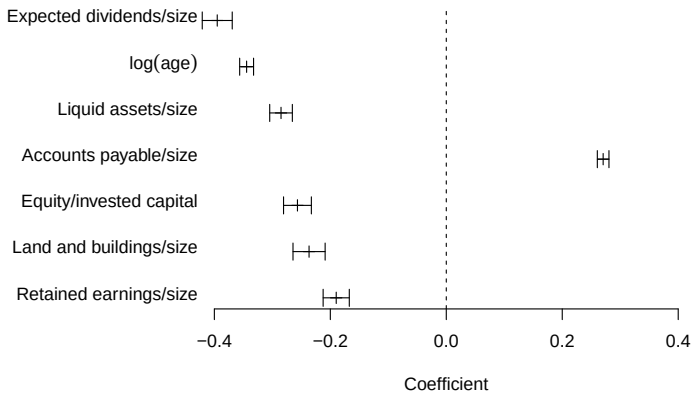
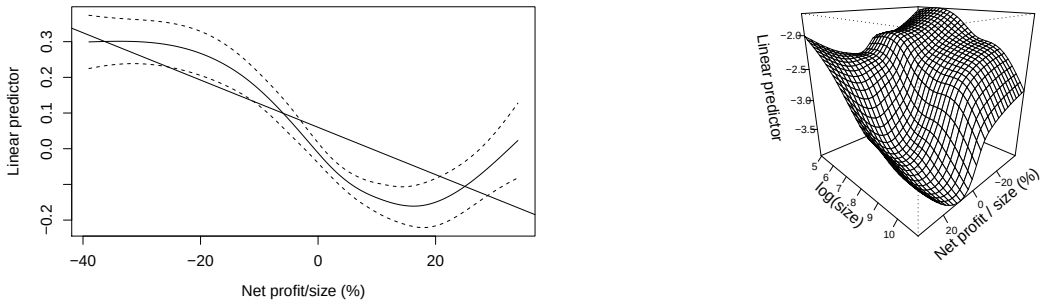


Figure 7: Standardized coefficients in the GLM. The plot shows the effect on the linear predictor from a one standard deviation move in the covariates. Only the 7 largest standardized estimates of non dummy variables are included in the plot. The outer lines are 95% Wald confidence intervals and the inner lines are the estimated coefficients.



(a) Non-linear net profit ratio main effect

(b) Non-linear net profit ratio and the log size effect

Figure 8: Example of non-linear effects in GAM. Panel (a) shows how the logit of the probability of entering into distress depends on the net profit divided by the size variable defined in Equation (8). The straight line is the effect in the GLM and the curve is the main effect in the GAM. The dashed lines are ± 2 standard deviations conditional on the estimated penalty variables in $\boldsymbol{\lambda}$. The spline is subject to an identification constrain (sum-to-zero constraint) so only the relative difference of the y-values along the curve is of interest. Panel (b) shows how the logit of the probability of entering into distress depends on changes in the net profit divided by the size variable and on changes in the log size variable. All other numerical covariates are set equal to their median value and categorical covariates are chosen to be the most common category.

the GLMM. We have reduced the dimension of each spline in the GLMM compared to the GAM as we do not penalize the splines in the GLMM.

One of the non-linear terms in the GAM and the GLMM is the net profit to size ratio. Figure 8(a) shows the main effect[16]of the net profit to size ratio in the GAM and GLM. Allowing for non-linearity, we see that the distress rate has a strong association to changes in the net profit ratio in the range from -10% to 0%,

[16]We refer to $\boldsymbol{\beta}_1^\top \boldsymbol{f}_1(x_1)$ as the main effect of x_1 in the model $\eta = \boldsymbol{\beta}_1^\top \boldsymbol{f}_1(x_1) + \boldsymbol{\beta}_2^\top \boldsymbol{f}_2(x_2) + \boldsymbol{\beta}_3^\top (\boldsymbol{f}_1(x_1) \otimes \boldsymbol{f}_2(x_2))$ where $\otimes$ denotes the Kronecker product and $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are subject to "sum-to-zero" constraint. See Wood (2017) for details.

Table 3: Non-linear effects in the GAM and the GLMM. All covariates except for log(size), log(age), and return on equity are divided by the size variable defined in Equation (8). "Varying coefficient" implies an interaction with a spline basis function for the first covariate and a linear effect for the second covariate. Hence, the slope of the second covariate "varies" as a function of the first covariate. The included non-linear effects in the GLMM are those which we deemed "the most non-linear" in the GAM.

First covariate	Second covariate	Type of term
Non-linear effects in the GAM		
Retained profit		Spline
Return on equity		Spline
Net profit	log(age)	Varying coefficient
Net profit	log(size)	Tensor product spline
Net profit	Liquid assets	Tensor product spline
Net profit	Other receivables	Tensor product spline
log(size)	Short-term bank debt	Tensor product spline
log(size)	Financial costs	Tensor product spline
log(size)	Tangible fixed assets	Tensor product spline
Liquid assets	log(size)	Tensor product spline
Liquid assets	Fixed costs	Tensor product spline
Non-linear effects in the GLMM		
Retained profit		Spline
Return on equity		Spline
Liquid assets		Spline
Other receivables		Spline
Financial costs		Spline
Net profit	log(size)	Tensor product spline
log(size)	Short-term bank debt	Tensor product spline

while the dependents of the linear predictor is flat in other regions on the net profit ratio scale. Also, firms with a very high net profit ratio tend to have a slightly higher rate of distress. A potential explanation is that firms with relatively large profits are growing fast and may be more volatile firms. The plot highlights that the association is far from linear as assumed in the GLM. Figure 8(b) shows an example of a smooth in two dimensions (a tensor product spline). The plot shows that the association between net profit and size is weaker for small firms. Furthermore, the figure shows that there is weak association between distress rate and firm size for firms with a large loss relative to their size. Overall there is a clear non-linear interaction effect.

C Details of the Two Bank Portfolio Example of Section 6.2

We will provide details regarding the simulation example in Section 6.2 in the following. Assume that we have two banks: one with few low-risk loans (Bank A) and one with many high-risk loans (Bank B). The

probability of a default for each firm j in the bank portfolio i is

$$p_{ij} = g^{-1}(\eta_{ij} + \epsilon), \qquad \epsilon \sim N(0, \sigma^2), \qquad \eta_{ij} = g\left(\frac{\bar{p}_i(j-1) + \underline{p}_i(n_i - j)}{n_i - 1}\right)$$

where ϵ is a random effect which we cannot observe and g is the logit function. We fix σ to 0.2 and let Bank A have $n_A = 501$ clients and Bank B have $n_B = 10\,001$ clients. Further we set Bank A's risk parameters to $(\underline{p}_A, \bar{p}_A) = [0.10, 0.30]$ and Bank B's risk parameters to $(\underline{p}_B, \bar{p}_B) = [0.15, 0.35]$. Thus, the latter bank has more clients which are more risky on average. We use the mean firm probabilities when we simulate the firms' outcome in the model that does not account for frailty. These probabilities are given by

$$\tilde{p}_{ij} = \int_{\mathbb{R}} g^{-1}(\eta_{ij} + \epsilon)\varphi(\epsilon; \sigma^2)d\epsilon > g^{-1}(\eta_{ij})$$

where $\varphi(\cdot; \sigma^2)$ is the normal distribution density function with zero mean and variance σ^2 and the inequality follows from a Jensen's inequality and holds when $\eta_{ij} < 0$. The firm outcomes are then simulated independently using $\tilde{p}_{ij}$ to produce confidence bounds for the portfolios similar to what we do for the GLM, GAM, and GB model in e.g., Figure 2.

DANMARKS NATIONALBANK
HAVNEGADE 5
DK-1093 COPENHAGEN K
WWW.NATIONALBANKEN.DK

DANMARKS
NATIONALBANK