ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

Predicting bank distress: testing the additional predictive power of an equity market indicator with machine learning algorithms

Master Thesis Accounting, Auditing and Control

Johannes van der Drift

382577

Supervisor: Dr. F.M. Elfers

Second assessor: Dr. M.H.R. Erkens

Disclaimer: The content of this thesis is the sole responsibility of the author and does not reflect the view of either the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Table of Contents

1. Introduction

More than ten years after the global financial crisis of 2008, we still observe the effects of regulatory reforms to improve the stability of financial institutions and the debt and equity markets. The financial crisis showed that regulations and oversight were ineffective in preventing the financial crisis (Dewatripont, Rochet and Tirole, 2010). When oversight by regulatory bodies or government intervention fails, market discipline can be a powerful governance mechanism to monitor and signal excessive risk taking by banks. Market discipline can be defined as the monitoring by and actions of depositors and shareholders to signal excessive risk in banks (World Bank, 2019). Two components of market discipline can be identified: monitoring and influencing (Bliss and Flannery, 2002). I will focus on the monitoring of banks to assess the additional predictive power of a market indicator for predicting bank distress. Earlier research has centered around the distance to default[1] as market indicator (Gropp, Vesala and Vulpes, 2006; Auvray and Brossard, 2012). However, their research methodology is based on logistic regression. Logistic regression fails to capture non-linear effects and advanced machine learning methods are recognized to consistently outperform logistic regression in classification problems (Dumitrescu, Hué, Hurlin and Tokpavi, 2021). Machine learning can be used to analyze large datasets, uncover the best variables and combinations of variables to explain and predict an outcome variable (Bertomeu, Cheynel, Floyd and Pam, 2020). Advanced machine learning techniques have not yet been used to ascertain the capabilities of a market indicator to predict banking distress. The models to predict defaults have improved substantially since the seminal works of Beaver (1966) and Altman (1968). Early research mainly comprised discriminant analysis. The work of Martin (1997) provided a shift in default research, by employing a logistic regression. Research using machine learning techniques started in the 1990s, when Odom and Sharda (1990) found Neural Network (NN) outperformed a multivariate discriminant analysis based on Altman (1968) for prediction bankruptcies. I will assess the additional predictive power of the distance to default market indicator on multiple distress events, which leads to the following research question as the focal point of this research:

Can the addition of a market indicator to established risk indicators improve predictions of bank distress?

To answer this question multiple analyses are conducted based on advanced machine learning algorithms. I formulate one hypothesis in the alternative form, to help answer the research question.

H1a: Market indicators improve machine learning models based on established risk indicators when predicting bank distress.

The findings of this paper give an indication that the market indicator distance to default can improve models predicting bank distress. These results however are not adequate enough to indicate an immediate shift in prediction models is needed. The results also indicate that the importance of a market indicator in bank distress prediction models becomes greater when a longer prediction horizon is employed. When the implied cost of capital is used as market indicator, there are significant improvements in predictions of bank distress one year ahead of the event.

The contribution of this paper is to assess the predictive power of market indicators on bank distress using machine learning. Prior research covering this approach on financial institutions is limited. This research corroborates the findings of Miller, Olson and Yeager (2015) of the small improvement a market indicator

[1] The distance to default (DD) is a measure of distance in the number of asset value standard deviations from a point of default where the value of assets and debt are equal.

has on bank distress predictions, but expands on the methodology by using advanced machine learning algorithms. When predicting bank distress over a larger horizon the importance of a market indicator as additional predictor becomes larger. Additionally, this paper shows the improvement in bank distress predictions when using implied cost of capital or equity beta as market indicator. The results of this study can be applied by regulatory bodies, as well as investors and banks, to assess warning signs of excessive risk taking by banks or predict financial instability.

The paper continues as follows. Section 2 will detail the relevant literature on banking supervision, risk indicators for predicting bank distress, machine learning, and hypothesis development. Section 3 will address the data used in this research. Section 3 will explain the distress events used as a dependent variable for the research, the risk and market indicators for predicting the distress events, and the machine learning methodology to test the hypothesis. Section 4 will show the results from four machine learning algorithms on various models for predicting bank distress. Section 5 will elaborate on additional modelling and robustness testing. Section 6 will state the conclusion of the research, and will address limitations of the paper and possible further research.

2. Literature review

This part of the paper will highlight previous research on banking supervision and bank distress predictions.

Previous literature in the use of market signals for bank supervision has mainly focused on debt markets, as opposed to equity market information (Curry, Fissel and Elmer, 2003). However, Levonian (2001) finds that the information content of debt markets is similar to that of equity markets with regard to market discipline. Moreover, equity markets are thought to process information more efficiently than bond markets (Saunders, 2001), and there are more banking institutions with publicly traded equity than those with publicly traded debt (Curry et al., 2003). The potential problems from using debt signals as market discipline indicator arise from their implementation into prediction models. Hancock and Kwast (2001) show that multiple bonds issued by a single U.S. bank can result in multiple spread estimates. Contradictory to the positive relation between yields and ratings, Bliss (2001) shows spreads have low predictive power when estimating ratings. In addition, spreads can incorporate time-varying liquidity premia, and this diminishes their usefulness in predicting bank distress. Elton (2001) shows that the premium in corporate rates over treasuries is in a surprisingly small fraction explained by expected default. This is corroborated by Huang and Huang (2012), who conclude that credit risk accounts for only a small fraction of yield spreads for investment-grade bonds.

There are also difficulties in using market indicators like the distance to default, and related expected default frequency. These difficulties are threefold; opacity, option value effect and moral hazard due to the public safety net (Auvray and Brossard, 2012). The first problem is embedded in the opacity of some bank assets, meaning they are not easily monitored by outside shareholders and creditors. The screening and monitoring of these assets must therefore be done by bank employees, which the outside stakeholders have to rely upon (Diamond, 1984; Freixas and Rochet, 1999). This opacity effect can be controlled for in models by using accounting variables. Another stated drawback of equity based signals is that upside gains for equity holders stemming from increased risk-taking also lead to increased asset volatility. In the case of high default probability, shareholders take more risk, because the option value is larger than the charter value (Park and Peristiani, 2007). Gropp, Vesala and Vulpes (2006), as well as this research, therefore use the distance to default indicator, because this combines equity price information, leverage and asset volatility in one indicator. Gropp and Vesala (2004) show that distance to default can indicate bank distress when asset volatility increased. The last issue is the rise of moral hazard with the presence of a public safety net for certain banks, also known as the "too-big-to-fail" banks (Dewatripont and Tirole, 1993). Effective market discipline can in this case be diminished because government in effect replaces market monitoring with government supervision. Risk is not fully incorporated in the cost of uninsured funding for these banks with systemic importance[2], as evidence shows (Acharya, Anginer and Warburton, 2017). This moral hazard effect can hinder the improvement of prediction models and needs to be controlled for (Distinguin, Rous and Tarazi, 2006). Other papers merely mention this problem (Gropp et al., 2006) or circumvent the problem by choosing a sample that does not contain "too-big-to-fail" banks (Curry, Elmer and Fissel, 2007).

Berger, Davies and Flannery (2000) find that the information gathered on the condition of large U.S. bank holding companies by supervisors and bond rating agencies can complement each other. They also find that supervisory assessments are less accurate in predicting future performance changes when compared to bond and equity market assessments. Their market indicators are based on the abnormal returns, insider or institutional holdings, and rating downgrades crossing the investment grade threshold. Gunther, Levonian and Moore (2001) find that an expected default frequency (EDF) market signal provides incremental information during the period between bank inspections. Their findings state EDF's as statistically significant in predicting supervisory rating downgrades. Other research by Krainer and Lopez (2004) concluded that equity market variables are useful for assessing the condition of bank holding companies. While they did not find an improvement in out-of-sample forecast accuracy when including equity market variables, this conclusion might be biased by their sample period of 1990-1999, which contained less bank distress than the years preceding their sample period. Gropp et al. (2006) provide evidence that market signals predict bank distress, based on a sample of European banks between 1990 and 2001. They find the distance to default can predict bank issuer rating downgrades between 6 and 18 months before the downgrade, but closer to the downgrade it performs quite poorly. Furthermore, the predictive power of spreads diminishes beyond the horizon of 12 months before a downgrade. Their results also suggest that spreads are only useful predictors for banks that are not insured against default by the government, while the predictive power of the distance to default is not affected by public support. They conclude that spreads and distance to default indicators complement each other, and that equity market information in addition to accounting data improves forecasting. Another paper using the distance to default indictor to predict bank distress was written by Auvray and Brossard (2012). They concluded that, for European banks between 1997 and 2005, the distance to default indicator had more predictive power in the case of concentrated ownership.

Two studies have focused on the distance to default indicator around the financial crisis period of 2008. Milne (2014) studied the distance to default for the 41 largest global banking institutions between the second half of 2006 and the second half of 2011. His findings suggest the distance to default indicator failed to predict either failure or bank share decline. Only for the latter half of 2008 did the indicator have statistically significant predictions of failure for the banks. He also finds the 'option value' of the bank safety net remained small, and bank shareholders were largely unaware of the exposed risk, suggested by the failure

[2] If there is no explicit guarantee by the government, the security prices should reflect the financial condition of a bank. However, Acharya, Anginer and Warburton (2017) find that bond premiums do not fully reflect the risk taking of banks because bond holders believe the government will prevent the adverse consequences of failure of these systemically important banks.

of the distance to default indicator as predictor for bank distress. He found little indication that the bank safety net was being used by bank shareholders to shift risk onto the taxpayers. Another paper by Miller, Olson and Yeager (2015) studies the contribution of equity and subordinated debt signals as predictors of bank distress during the financial crisis. The sample of their research contained bank holding companies (BHC) between the fourth quarters of 2006 and 2012. They concluded the expected default frequency, derived from the distance to default, did not improve predictions relative to accounting-based indicators. The yield spreads on subordinated debt also did not improve bank distress predictions, because the "too-big-to-fail" subsidies distorted the risk rankings of the largest BHC's. For large BHC's, the tier 1 leverage ratio was the most accurate distress indicator, this holds for the period during the crisis as well as outside of that period.

The accounting-based signals used in this paper are proxies developed by Thomson (1991). He provided five variables that could accurately predict bank defaults in the period 1984-1989. The variables were proxies for capital adequacy, asset quality, management quality, earnings and liquidity. These factors combined make up the acronym CAMEL, which is a system used to rate financial institutions by regulatory banking authorities. In 1997 sensitivity to market risk was included to expand the system to CAMELS. More recently, Cole and White (2012) tested proxies for CAMELS and found they are accurate predictors of bank failures in both periods of 1985-1992 and following the financial crisis of 2008. Their results also suggest real-estate loans play an important role in the strength and weakness of a bank. With more construction loans, commercial mortgages and multi-family mortgages, banks have a higher probability of failure, while banks with more loans allocated to the residential single-family market are either neutral or have higher probability of survival. Jin, Kanagaretnam and Lobo (2011) also test various accounting variables to predict banking failure in the period leading up to the crisis of 2008. They identify the following ten predictors: auditor type, Tier 1 capital ratio, proportion of securitized loans, nonperforming loans, loan loss provisions, growth in commercial loans, growth in real estate loans, growth in overall loans, loan mix, and a dummy if the bank is traded publicly.

The recent developments in advanced machine learning techniques have made them a viable option for many empirical researchers. Machine learning algorithms can detect complex patterns in large sets of data (Bertomeu, Cheynel, Floyd and Pan, 2020). The algorithm selects variables that best explain an outcome variable, and can find appropriate combinations of variables to make accurate predictions out-of-sample. Odom and Sharda (1990) were among the first to apply this approach for distress prediction, using neural networks (NN) to predict corporate defaults. This was later expanded on by Tam (1991) and Tam & Kiang (1992) as an application for bank default prediction. Their findings suggest NN outperformed other prediction models. Heo and Yang (2014) found that adaptive boosting outperformed other models, when classifying bankruptcy predictions for Korean construction companies. They especially found higher predictive power for larger companies. Although their research only focuses on one industry, this paper will focus on distress in the banking sector. Danenas (2015) evaluated various machine learning models for credit risk evaluation and default assessment for US firms. He found an overall high classification accuracy but stated support vector machines are less stable than other classification models. Kim, Kang and Kim (2015) provide valuable insight into the problem of data imbalance, when predicting classifications with a majority class. The issue of data imbalance has two sides. Firstly, the performance of classification models is mostly based on the arithmetic accuracy. With imbalanced samples, the model's predictive performance is highly skewed towards the majority class. Banks in distress, for instance, are relatively less abundant in most samples than banks not in distress. This holds even more for default instances. Basically the model learns to predict the majority class, but not so much the minority class, which in most cases is the instance at the center of research. In this case a model with high accuracy can be argued to be meaningless. Like Kim et al., this paper will use the area under the receiver operating characteristic (ROC) curve to assess classification accuracy of both majority and minority classes. The second problem arising from data imbalance is the distortion of the decision boundaries. With largely imbalanced samples, the decision boundary for the majority class tends to gradually expand, while the decision boundary for the minority class is gradually reduced. This results in decreased accuracy for minority class classification. The proposed solution for this issue is boosting, which gives more learning opportunities to the minority class. More recently Barboza, Kimura and Altman (2017) researched various machine learning models for corporate bankruptcy prediction. They found that machine learning models show approximately 10 percent more accurate predictions in relation to traditional models, on average. Their result also suggested the machine learning technique related to random forest performs the best out of all tested machine learning techniques. They also touch on the drawback of machine learning in reduced explanatory validation for the models. The goal, however, is to correctly predict and not explain it. In that case the estimation of prediction error is more important than the relative contribution (Efron and Hastie, 2016).

2.1 Hypothesis development

The focal point of this paper is whether market signals provide an improvement in predicting bank distress in addition to other public indicators. The market signal used to test this is the distance to default. Earlier research has shown that this indicator captures three important aspects of the condition of a bank: first is the total market value of assets, second the debt level of the firm as indication of risk which has to be paid off by the total market value of assets, and third is the volatility of assets (Gropp et al., 2006; Miller et al., 2015). To properly test the predictive power of this indicator, a machine learning method approach will be used. Distress in banks can result in high costs for all stakeholders, this shows the need for accurate indicators to assess a banks' condition. Therefore, the hypothesis tested in this paper is as follows:

H1a: Market indicators improve machine learning models based on established risk indicators when predicting bank distress.

3. Data and methodology

This section will detail the distress events tested, the book- and market signals to classify those distress events, and the machine learning methods to achieve the best possible classification. The data used in this research was collected from multiple sources. The quarterly data for US banks for the period 1990-2020 was collected from Compustat Bank Fundamentals Quarterly. For the same sample period, the Center for Research in Security Prices (CRSP) provided market data, like stock information. The equity volatility indicated by the beta was collected from the WRDS Beta Suite.

3.1 Distress events

To test the hypothesis that market information improves the prediction of a bank in distress, two different distress events are applied in this paper: a Texas ratio greater than one hundred percent and a bank failure. Table 1 gives an overview of all distress events.

The main distress event in this paper is recorded using the Texas ratio. This ratio is the sum of non-performing assets (loans more than 90 days outstanding) and real estate other than bank premises owned, which is divided by the sum of tangible common equity (TCE) and loan loss reserves. Tangible common equity is calculated by subtracting intangible assets and preferred stock from the banks' total equity. The Texas ratio was developed as a result of the large amount of bank failures in Texas in the 1980s. Gerard Cassidy and fellow analysts at RBC Capital Markets found that banks with a Texas ratio over one hundred percent have a high probability of default. A Texas ratio greater than one hundred percent indicates the non-performing assets are larger than the available resources to cover potential losses on those assets. Only the first observations of a Texas ratio over one hundred percent are included as a distress event. To properly test the predictive power of the models, the Texas ratio dependent variable will be a dummy with value of 1 in the quarter preceding a recorded Texas ratio over one hundred percent. This way, a distress event in the following quarter can be predicted based on the information available in the current quarter. Predictions will also be made for a one year and two year prediction horizon. The sample includes 63 bank quarter observations of a Texas ratio over one hundred in the following quarter, 33 observations of this distress event one year ahead, and 10 observations for predictions two years prior to the distress event.

A bank failure is also classified as a distress event. The Federal Deposit Insurance Corporation (FDIC) provides a list of failed banks since October 2000. This list was linked with quarterly data through the collection of Federal Reserve Bank ID's for the Bank Holding Company (RSSD ID) which could be linked with the CRSP PERMCO numbers. The resulting observations were cross-referenced with CRSP delist events for all banks in the sample. For this research the focus was on delisting events where a bank was in distress. Bank failure observations were added based on liquidation (delist code 450), insufficient assets (delist code 561), bankruptcy (delist code 574), and for the protection of investors and the public interest (delist code 585). The distress event for bank failure is the last quarter recorded within one year of the bank failure or delist date. The dependent variable to test the predictive power of the models will again be a dummy with value 1 in the quarter preceding the bank failure quarter. The sample to test this includes more observations than the sample tested for the Texas ratio, because this sample includes subsequent observations of a Texas ratio over one hundred. Also in the case of bank failures, no bank quarter observations are included after the distress event. The observation of a bank failure is also highly correlated with a Texas ratio over one hundred. Of the 12 observations with a bank failure in the following quarter, 11 of these observations coincided with a Texas ratio over one hundred in the following quarter. Predictions are made on the 12 observations in the following quarter, 7 observations one year ahead, and 5 observations for a two year prediction horizon.

A third distress event can be indicated by the credit rating a bank receives, if that rating drops below a certain threshold. However, after collecting all relevant data and issuer level credit ratings for the banks in the sample, the amount of distress observations was not adequate for robust testing.[3] Therefore, contrary to earlier research (Auvray and Brossard, 2012; Miller et al., 2015), this distress event will not be included in this paper.

[3] The credit ratings were collected from the Eikon database and include both Fitch and Moody's long term issuer credit ratings. After all necessary calculations for the ratios used in testing, just two bank distress observations remained.

Table 1

Panel A: Distress events

$Texas\ ratio_{q+1}$ $> 100\%$	This is a dummy dependent variable taking value 1 in the quarter preceding an observation of a Texas ratio over 100%. This ratio is the sum of non-performing assets (loans more than 90 days outstanding) and real estate other than bank premises owned, which is divided by the sum of tangible common equity (TCE) and loan loss reserves.
$Bank\ failure_{q+1}$	This is a dummy dependent variable taking value 1 in the quarter preceding an observation of a bank failure as per FDIC or CRSP delisting event related to bank distress.

Panel B: Distress event frequencies

Year	$Texas\ ratio_{q+1}$	$Bank\ failure_{q+1}$	$Texas\ ratio_{q+4}$	$Texas\ ratio_{q+8}$	$Bank\ failure_{q+4}$	$Bank\ failure_{q+8}$
	$> 100\%$		$> 100\%$	$> 100\%$		
1993	0	0	0	0	0	0
1994	0	0	0	0	0	0
1995	0	0	0	0	0	0
1996	0	0	0	0	0	0
1997	0	0	0	0	0	0
1998	0	0	0	0	0	0
1999	0	0	0	0	0	0
2000	0	0	0	0	0	0
2001	0	0	0	0	0	0
2002	0	0	0	0	0	0
2003	0	0	0	0	0	0
2004	0	0	0	0	0	0
2005	0	0	0	0	0	0
2006	0	0	0	1	0	0
2007	0	0	1	1	0	0
2008	2	0	1	0	0	0
2009	43	5	23	7	3	3
2010	13	2	7	1	3	2
2011	4	4	1	0	1	0
2012	1	1	0	0	0	0
2013	0	0	0	0	0	0
2014	0	0	0	0	0	0
2015	0	0	0	0	0	0
2016	0	0	0	0	0	0
2017	0	0	0	0	0	0
2018	0	0	0	0	0	0
2019	0	0	0	0	0	0
2020	0	0	0	0	0	0
Total	63	12	33*	10*	7*	5*
Total n	15391**	16065	15391**	15391**	16065	16065

* These totals are less then the totals for models with q+1 as prediction horizon because data is not available. Some banks enter the sample at a later stage than year 1993. If a bank is in distress at year 2009 for example, but only data one year prior is available, the bank will have no observation when the distress event is q+8 away.

** The number of observations in this sample is reduced from the sample for bank failures. This is due to deletion of subsequent Texas ratio distress observations after the first recorded distress observation for robust testing.

3.2 Distress signals

The main focus of this research is the added value of market signals when classifying a bank as in distress. An overview of all dependent and independent variables is shown in table 2. The Merton-KMV distance to default (DD) provides a risk indicator based on market prices (Merton, 1974). Several researches have shown the added value of this indicator, because it takes the leverage of the bank and asset volatility into account (Auvray & Brossard, 2012). The DD can be derived from the modelling of equity as a call option on the assets of a company (Crosbie and Bohn, 2003). The value of equity and value of assets are connected by the following equation:

$$V_e = V_a N(d_1) - e^{-rT} D N(d_2) \tag{1}$$

for $d_1 = \frac{\log\left(\frac{V_a}{D}\right)+\left(r+\frac{\sigma_a^2}{2}\right)T}{\sigma_a\sqrt{T}}$ and $d_2 = d_1 - \sigma_a\sqrt{T}$

Where $V_e(V_a)$ is the value of equity (assets), D is the level of debt, σ_a is the volatility of assets, $N(d_1)$ and $N(d_2)$ are the normal distributions of d_1 and d_2, and r is the risk-free rate. Then, following the Black-Scholes option pricing model, the level and volatility of the market value of assets can be determined using the market value of equity, equity volatility and debt level (Merton, 1974). The volatility of equity can be derived from the beta. To reduce noise in observations, the average beta per quarter is taken from daily estimated beta's with a 252 day estimation window in the WRDS Beta Suite. The level of debt can be obtained from quarterly data. A common approach for the debt in this calculation, following Auvray & Brossard (2012) and Moody's KMV, is to take a one year horizon using the sum of short-term debt and half of the long-term debt. This is because of the reduced impact of long term debt on the probability of default when there is a trend of growing assets. The value of equity is stock price multiplied by outstanding shares. To obtain the value and volatility of assets, one solves the system of two non-linear equations to minimize sum of squared errors (Shah, Singh and Aggarwal, 2013): $eq^2 = eq_1^2 + eq_2^2$

for $eq_1 = V_e - V_a N(d_1) - e^{-rT} D N(d_2)$ and $eq_2 = V_e \sigma_e - N(d_1) V_a \sigma_a$

When V_a and σ_a are known, one can calculate the distance to default:

$$DD = \frac{\log\left(\frac{V_a}{D}\right) + (r - \sigma_a^2)T}{\sigma_a\sqrt{T}} \tag{2}$$

The DD is a measure of distance in the number of asset value standard deviations from a point of default where the value of assets and debt are equal. A higher DD can be the result from an increase in valuation of the assets. A higher DD can also come from a decrease in volatility reflecting less uncertainty about the asset values.

To determine a benchmark for classifying the financial stability of a bank, the approach in this paper will be based on multiple established risk indicators similar to the CAMELS rating system for financial institutions (Auvray & Brossard, 2012; Miller et al., 2015). This results in creating variables reflecting Capital adequacy, Asset quality, Management quality, Earnings, Liquidity and Sensitivity. Capital adequacy reflects the amount of capital that serves as protection against potential losses. Banks should have adequate capital reflecting the risk of operations. Capital adequacy ratios (C1, C2, C3) are expected to be negatively related with bank distress. Asset quality measures the amount of risk a bank is exposed to relating to various asset portfolios (e.g. loans, other real estate owned). Asset quality is impaired when earning assets are exposed to higher risk. Asset quality ratios (A1, A2, A3, A4) are expected to be positively related to bank distress. Adequate management of the bank also forms an important part of its stability. However, it is very difficult to correctly capture the ability of management to control for risk exposure of a bank. This study captures management quality by calculating total operating expenses over income before provisions[4]. The variable for management quality (M1) is expected to be positively related to bank distress. Next is earnings quality, which can determine the long term viability of a bank. Earnings should be stable, not reliant on one-time gains, and cover the credit risk exposure of a bank. Earnings quality variables (E1, E2, E3, E4, E5, E6) are expected to be negatively related to bank distress. Following earnings quality, the liquidity of a bank is measured. The liquidity of a bank is especially important to assess a bank's financial stability. A bank should have adequate liquidity to meet its obligations. Liquidity ratios (L1, L2, L3) are expected to be positively related to bank distress. The last part of the CAMELS rating is sensitivity to market risk, which was a recent addition to the rating system. Regulators regard this as the degree to which the earnings and financial stability of a bank are affected by changes in interest rate or other macroeconomic factors. Following earlier research (Kerstein and Kozberg, 2013), the proxy used for market sensitivity in this paper is the ratio of interest bearing deposits to total assets. As this measure increases, the bank will be more sensitive to interest rate changes, and therefore be positively related to distress. Besides proxies for the CAMELS variables, I will further apply variables used in other bankruptcy prediction research. One important factor stressed by Barboza et al. (2017) is to include one or more variables reflecting a change in certain indicators. I will therefore construct two variables reflecting changing conditions. The first variable indicates liquidity change (X1), and is expected to be negatively related to bank distress. It is calculated as the sum of change in cash, cash equivalents and receivables, divided by the sum of change in deposits and current debt. The second variable indicates change in loans receivable to total assets. This second change variable (X2) is expected to be negatively related to bank distress. The last variable to account for size of the banks is a log transformed total amount of assets. No expectation is set on the relation between this variable and bank distress. An overview of the descriptive statistics for all variables can be seen in table 4, panel A. The formula which forms the basis of testing is as follows:

$$\begin{aligned} D_{t+i} = \; & C1_t + C2_t + C3_t + A1_t + A2_t + A3_t + A4_t + M1_t + E1_t + E2_t + E3_t + E4_t \\ & + E5_t + E6_t \; + L1_t + L2_t + L3_t + S1_t + X1_t + X2_t + X3_t \end{aligned} \tag{3}$$

Where D_{t+i} denotes bank distress (dummy value 1) $t + i$ quarters away and all independent variables are taken from the current quarter. To clarify, a model with D_{t+4} as a dependent variable predicts bank distress one year ahead. The formula used to test the additional predictive power of a market indicator for predicting bank distress is as follows:

$$\begin{aligned} D_{t+i} = \; & DD_t + C1_t + C2_t + C3_t + A1_t + A2_t + A3_t + A4_t + M1_t + E1_t + E2_t + E3_t + E4_t \\ & + E5_t + E6_t \; + L1_t + L2_t + L3_t + S1_t + X1_t + X2_t + X3_t \end{aligned} \tag{4}$$

This formula follows the same logic as the first, with the addition of the distance to default as an independent variable.

[4] It should be noted this only captures management quality in the short term, and does not asses the long-term effect of management on the financial stability of a bank. However, as this proxy captures the cost-effectiveness of revenue realization it can give an indication as to how this cost-effectiveness was realized by previous management, both in the short-term and long-term.

Table 2
Predictive variables used to assess whether a bank will be in distress. For Texas ratio distress and bank failure the current quarter values are used as input to predict distress in the following quarter.

Variable	Calculation method
DD	$\frac{\log\left(\frac{V_a}{D}\right)+\left(r-\sigma_a^2\right)T}{\sigma_a\sqrt{T}}$
C1	$\frac{Tier\ 1\ capital_q}{Total\ risk\ weighted\ assets_q}$
C2	$\frac{Common\ equity_q}{Total\ assets_q}$
C3	$\frac{Common\ equity_q}{Total\ long\ term\ debt_q}$
A1	$\frac{Loan\ loss\ provisions_y}{Net\ interest\ revenue_y}$
A2	$\frac{Net\ charge\ offs_q}{Gross\ loans\ income_q}$
A3	$\frac{Non\ perfroming\ assets_q}{Loan\ loss\ provisions_q}$
A4	$\frac{Real\ estate\ other\ then\ bank\ premises\ (OREO)_q}{Total\ assets_q}$
M1	$\frac{Current\ operating\ expenses_q}{Net\ income\ before\ loan\ loss\ provisions_q}$
E1	$\frac{Net\ income\ after\ extraordinary\ items_q}{Total\ assets_q}$
E2	Earnings per share, diluted, excluding extraordinary items and 12 month moving average
E3	Net interest $\text{margin}_q = \frac{Interest\ revenue - Interest\ expenses}{Average\ earning\ assets}$
E4	$\frac{Retained\ earnings_q}{Total\ assets_q}$
E5	$\frac{Current\ operating\ earnings\ before\ tax_q}{Total\ assets_q}$
E6	$\frac{(Operating\ cash\ flow_y + Financing\ cash\ flow_y + Investing\ cash\ flow_y)/4}{Long\ term\ debt_q}$
L1	$\frac{Cash\ and\ due\ from\ banks_q + Federal\ funds\ sold_q}{Deposits_q + Short\ term\ borrowings_q}$
L2	$\frac{Commercial\ paper_q}{Total\ assets_q}$
L3	$\frac{Fixed\ assets\ (ppe)_q}{Liquid\ assets\ (cash\ and\ due\ from\ banks + federal\ funds\ sold)_q}$
S1	$\frac{Interest\ bearing\ deposits_q}{Total\ assets_q}$
X1	$\frac{\Delta\ Cash\ and\ equivalents_y + \Delta\ Receivables_y}{\Delta\ Deposits_y + \Delta\ Current\ debt_y}$
X2	$\frac{(\Delta\ Loans\ receivable_y)/4}{Total\ assets_q}$
X3	$\log(Total\ assets_q)$

3.3 Machine learning

As technology and archival financial research evolves, we are continuously introduced to new streams of data. Machine learning algorithms can discover complex patterns in the data, pick the best variables that explain a certain outcome variable, and detect appropriate combinations of variables to make out-of-sample predictions as accurate as possible (Bertomeu et al, 2020). In this research, several machine learning algorithms are used to classify a bank as in distress or not in distress, based on the aforementioned variables. First, the general process of applying machine learning will be explained, after which we go into detail about the specific algorithms.

3.3.1 Dealing with missing data and class imbalance

After collecting the data, new variables are created to use in testing. Missing data also needs to be addressed. There are multiple ways to impute missing data. Deleting observations leads to a robust model, but works poorly if there is a large portion of data missing. Mean/median imputation is an easy solution, however this does not take the covariance between variables into account. A powerful tool to deal with missing values is the MICE[5] package available in R statistical software. The MICE package works under the assumption that data is missing at random. This means the probability of missing data is reliant on observed data, therefore observed data values can be used to make predictions of missing data. Through linear regression and “Predictive Mean Matching” a distribution is formed per missing data point, from which plausible replacements are drawn for the missing values (Van Buuren and Groothuis-Oudshoorn, 2011). Another potential problem are outliers in the data. The machine learning algorithms employed in this research are more robust when dealing with outliers, than a logistic regression for example, so this will pose less of a problem. With the input of imputed data, a machine learning algorithm learns to predict the classification of a bank as being in distress or not. To correctly test this, the data will be split into a training set and a validation set to validate the predictions of the algorithms against known values of the outcome variable. To ensure the algorithms get equal opportunity to learn the features of each of the classes in the dataset, and subsequently test what it has learned on the same number of instances of each class, a stratified split would commonly be applied. Stratified splitting is important when predicting a binary outcome variable. However, following earlier research on machine learning models with yearly or quarterly observations, the data is split at a certain point in time (Barboza et al., 2017; Bertomeu et al., 2020). This while making sure the algorithms have sufficient data to both train and test on the outcome variable of interest. By setting the split at year 2010, the training set contains 45 Texas ratio distress observations and the validation set contains 18. A potential issue with predicting bank distress is that it does not happen often. This is seen by the low amount of distress events in table 1 when compared with the total amount of quarter observations. This creates an imbalance in the data, meaning there is a clear majority class (no distress) and minority class (distress), with a ratio of 37:1. Machine learning algorithms will then create biased predictions because they assume balanced classes, and aim to minimize the error of the whole set for which a minority class has little effect. There are several methods to solve this and create a balanced training dataset. These methods are known as resampling methods and they modify the training data to create a balance between the two classes. The first method is oversampling. This method duplicates values from the minority class to create balance in the data. However, this will lead to overfitting and possibly lead to worse out-of-sample predictions. The second method is undersampling, which works the opposite of oversampling. This method chooses random observations from the majority class to be deleted until the data is balanced. A problem with this method is

[5] Multivariate Imputation via Chained Equations (MICE)

that deleting observations will in effect delete important information to classify the majority class. One can combine over- and undersampling as well to balance the data. The last method is synthetic data generation, this is a type of oversampling, but instead it creates new observations rather than duplicating them from the minority class. One such technique is called ROSE, this technique applies bootstrapping; duplicating observations in the training data as assessment set and resampling the training data. The most widely used technique for synthetic data generation is called ‘synthetic minority oversampling technique’ (SMOTE). This works with bootstrapping and K-Nearest Neighbor. The SMOTE algorithm takes the distance between the variable vector of a certain minority observation and its nearest neighbor, and multiplies this distance with a random number between 0 and 1. This is added to the variable vector and creates random points on the space between two variables of observations. In effect, the algorithm creates new observations that are similar to other observations in the minority class, instead of just duplicating them or creating complete random values. To justify choosing one of the aforementioned resampling techniques, the results of a k-nearest neighbor and random forest algorithm predicting Texas ratio distress using all these techniques are shown in table 3. The best method is one that maximizes true positive predictions (1,1 in confusion matrix) and true negative predictions (0,0 in confusion matrix). As shown in the table ROSE provides the best results for predicting positive instances, however the negative instances are the worst. The next best technique is undersampling, with fifteen out of twenty true positive predictions. The number of false positives is however still substantial. The best performing resampling technique is SMOTE, with a very high accuracy and maximizing both true positives and true negatives. SMOTE will be used for all subsequent testing in this research. The three best performing resampling techniques (SMOTE, oversampling, and both over- and undersampling) will also be addressed in robustness testing.

Table 3

Panel A: Confusion matrix results from resampling training data using a k-nearest neighbor model to predict bank distress (Texas ratio q+1).

Method		Oversampling		Undersampling		Both (Under & Over)		ROSE		SMOTE	
		Actual		Actual		Actual		Actual		Actual	
		1	0	1	0	1	0	1	0	1	0
Prediction	1	4	269	14	1977	4	558	18	5080	10	1212
	0	14	13275	4	11567	14	12986	0	8464	8	12332
Accuracy		0.9791		0.8539		0.9578		0.6254		0.91	

Panel B: Confusion matrix results from resampling training data using a randomforest model to predict bank distress (Texas ratio q+1).

Method		Oversampling		Undersampling		Both (Under & Over)		ROSE		SMOTE	
		Actual		Actual		Actual		Actual		Actual	
		1	0	1	0	1	0	1	0	1	0
Prediction	1	0	5	13	440	2	27	16	11611	3	73
	0	18	13539	5	13104	16	13517	2	1933	15	13471
Accuracy		0.9983		0.9672		0.9968		0.1437		0.9935	

3.3.2 Performance metrics

After imputation and creating a balanced train set, all models can be trained using the risk indicators as input variables and the different distress events as outcome variables. Classifiers can calculate the binary outcome variable based on predicted values for the positive instance or the negative instance. Based on a threshold for the predicted probability, the actual binary outcome will then be selected. To evaluate the performance of the models multiple metrics are used. These metrics include the true positive rate (TPR, sensitivity, recall, 1- Type I Error) and the true negative rate (TNR, specificity, 1- Type II Error). True positive is when a bank in distress is correctly classified, and true negative is when a bank not in distress is correctly classified. Of these two, the TPR is more important, as higher Type I errors result in higher losses when predicting a bank will not be in distress while it actually will be. Another measure is the precision. This is the ratio of true positive prediction to all positive predictions. The most basic performance metric is accuracy, calculated as the number of correct predictions divided by the total amount of observations in the validation set. As mentioned earlier, this is a very crude performance metric. A model with none of the banks in distress correctly classified can have a high accuracy, but basic intuition tells us this is not a desirable model. There is also a trade-off between precision and sensitivity (recall) based on the threshold (Calders and Jaroszewicz, 2007). A high threshold will result in high precision but low sensitivity, and the opposite holds as well. This research will follow the approach that all predicted probabilities of the outcome variable above 0.5 will be classified as 1, and 0 otherwise. The last performance metric covered is also the most comprehensive because it can measure the performance of a model without fixing the threshold, which is the AUC (area under the ROC curve). The Receiver Operating Characteristics (ROC) curve is a plot of the sensitivity on the y-axis and specificity on the x-axis for different thresholds. An AUC of 0,5 means a random prediction, so the models must at least perform better than that, and the performance increases as the AUC approaches a maximum of 1. The AUC follows a Wilcoxon-Mann-Whitney equation (Hanley and Mcneil, 1982) in the following form:

$$AUC = \frac{\sum_{1}^{N_{D0}} \sum_{1}^{N_{D1}} S(D_0, D_1)}{N_{D0} * N_{D1}} \tag{5}$$

Here N_{D0} denotes the total number of banks not in distress in the set and N_{D1} denotes the total number of banks in distress, and $S(D_0, D_1)$ is 1 if $P(D_0) < P(D_1)$. The AUC measures the probability that a random chosen positive instance will be ranked ahead of a randomly chosen negative instance. To test the central hypothesis of this research, the distance to default variable will be added as input variable for the models, and the potential improvement in model performance will be measured.

3.3.3 Machine learning algorithms

The machine learning techniques applied in this research include K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest (RF), and XGBoost. Data transformation and modelling will be done with R statistical software. All R software packages used, are not modified before introducing the learning algorithms. The first algorithm is K-Nearest Neighbor. This algorithm forms clusters in the data based on similar observations. The advantage is that no assumptions are made about the data distribution. A disadvantage however, is that KNN is not suited for large datasets and is sensitive to irrelevant variables. Also, if variables have different scaling units, the variables have to be normalized so all distances have the same range of values. We move on to Naïve Bayes, which is particularly suited for classification with large datasets and a large number of variables. However, the drawback of this approach is it assumes rigid independence between the variables to predict the outcome, and this does not hold in most cases. The next algorithm is Random Forest, which is an ensemble model based on Decision Trees. The Decision Tree splits data based on decision classifications, whether a person is above or below a certain age for example. The base of a tree is the root node which gives the highest information gain in classification. The data is then further split through decision nodes until all data is classified. A single decision tree has multiple drawbacks; complex trees through overfitting, non-optimal solutions when a decision node is not optimal (greedy trees), and instability because of high variance in the data. An alternative is Random Forest. It consists of multiple decision trees and averages the outcomes across all trees. The classification outcome for RF is the class selected by the majority of trees. The final algorithm we apply is XGBoost. This stands for “Extreme Gradient Boosting”, and is similar to the random forest algorithm but adds gradient boosting. The technique of boosting refers to ensemble models that add new models to correct an existing model. Specifically with gradient boosting, a new model (tree) is added to predict the error (residual) of the previous model (existing forest). The name gradient boosting comes from the gradient descent algorithm that is used to minimize training losses with the addition of a new model. This training loss indicates how well the model predicts the training data, shown by the mean squared error. For all algorithms the hyper parameters will be tuned to improve model performance.

4. Empirical results and analysis

The descriptive statistics for all variables are shown in table 4. Panel A shows the descriptive statistics before imputation of missing values, and panel B shows the descriptive statistics after imputation. The potential issues of missing data and class imbalance are already addressed. Another potential issue of outliers does not apply to this research because all machine learning algorithms used are less sensitive to outliers then other algorithms like a logistic regression.

Table 4

Panel A: Descriptive statistics Texas ratio testing sample.

Variable	DD	C1	C2	C3	A1	A2	A3	A4	M1	E1	E2	E3	E4	E5	E6	L1	L2	L3	S1	X1	X2	X3
Min	-24.133	-1.710	-0.033	-0.240	-1.391	-0.068	-41,857.000	0.000	-959.929	-0.071	-166.900	-212.470	-0.359	-0.084	-2,515.824	0.003	0.000	0.000	0.283	-5,620.152	-0.037	4.487
1st Quartile	2.444	10.850	0.088	0.740	0.021	-0.001	5.800	0.000	1.008	0.001	0.540	3.190	0.023	0.002	-0.026	0.030	0.000	0.166	0.463	-0.156	0.001	7.001
Median	4.425	12.380	0.103	1.240	0.054	0.000	13.750	0.001	1.220	0.002	1.210	3.550	0.047	0.003	0.007	0.053	0.000	0.340	0.514	0.087	0.005	7.900
Mean	15.142	12.870	0.108	26.110	0.118	-0.001	34.910	0.003	1.798	0.002	1.432	3.537	0.044	0.002	-0.227	0.074	0.003	0.458	0.509	-0.382	0.007	8.170
3rd Quartile	9.167	14.320	0.123	2.280	0.124	0.000	29.860	0.003	1.586	0.003	2.100	3.900	0.068	0.004	0.084	0.090	0.000	0.621	0.566	0.595	0.012	9.001
Max	14,363.169	98.400	0.868	75,937.000	6.382	0.018	148,066.000	0.075	5,130.783	0.065	47.500	11.260	0.189	0.061	1,071.384	0.975	0.072	6.030	0.708	596.909	0.088	15.035
Missing	-	523	1	-	667	155	1065	1	86	86	32	118	4	5	942	1,189	14,100	1,020	13,335	9,420	11,372	1
Total n	15,391																					

Panel B: Descriptive statistics Texas ratio testing sample after MICE imputation.

Variable	DD	C1	C2	C3	A1	A2	A3	A4	M1	E1	E2	E3	E4	E5	E6	L1	L2	L3	S1	X1	X2	X3
Min	-24.133	-1.710	-0.033	-0.240	-1.391	-0.068	-41,857.000	0.000	-959.929	-0.071	-166.900	-212.470	-0.359	-0.084	-2,515.824	0.003	0.000	0.000	0.283	-5,620.152	-0.037	4.487
1st Quartile	2.444	10.840	0.088	0.740	0.022	-0.001	5.800	0.000	1.009	0.001	0.540	3.190	0.024	0.002	-0.028	0.030	0.000	0.162	0.529	-0.049	0.002	7.001
Median	4.425	12.400	0.104	1.240	0.055	0.000	13.980	0.001	1.221	0.002	1.210	3.550	0.047	0.003	0.007	0.053	0.000	0.331	0.580	0.116	0.007	7.900
Mean	15.142	12.900	0.108	26.110	0.124	-0.001	44.280	0.003	1.813	0.002	1.430	3.534	0.044	0.002	-0.102	0.073	0.000	0.449	0.576	0.996	0.010	8.169
3rd Quartile	9.167	14.380	0.123	2.280	0.131	0.000	31.150	0.003	1.590	0.003	2.100	3.900	0.068	0.004	0.086	0.090	0.000	0.605	0.634	1.008	0.016	9.001
Max	14,363.169	98.400	0.868	75,937.000	6.382	0.018	148,066.000	0.074	5,130.783	0.065	47.500	11.260	0.189	0.061	1,071.384	0.975	0.072	6.030	0.709	596.909	0.088	15.035
Missing	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Total n	15,391																					

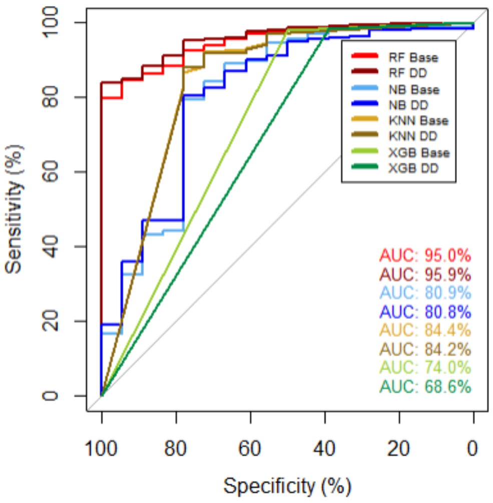

Figure 1: ROC curves for models with Texas ratio over one hundred as distress event in q+1. The corresponding area under curve (AUC) is also printed per model.

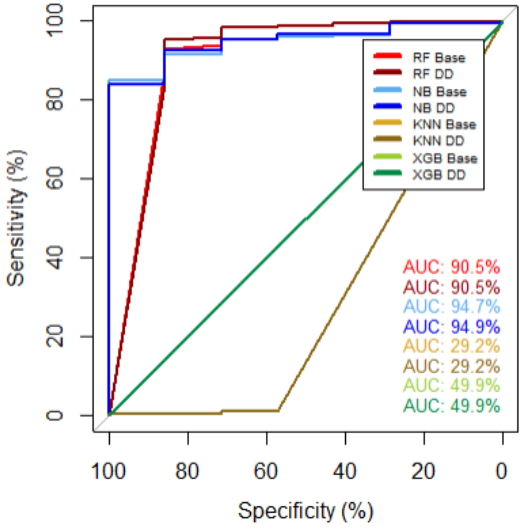

Figure 2: ROC curves for models with a bank failure as distress event in q+1. The corresponding area under curve (AUC) is also printed per model.

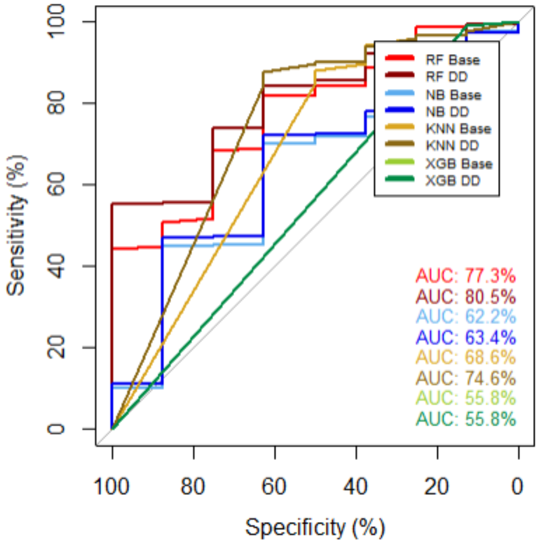

Figure 3: ROC curves for models with Texas ratio over one hundred as distress event in q+4. The corresponding area under curve (AUC) is also printed per model.

4.1 Machine learning algorithm results

Table 5 panel A shows the performance metrics for classifying bank distress for all base models and all models including the DD as an independent variable. Figure 1 shows the ROC curves for the models predicting $Texas\ ratio_{q+1} > 100\%$, figure 2 shows the ROC curves for testing $Bank\ failure_{q+1}$, and figure 3 shows the ROC curves for testing $Texas\ ratio_{q+4} > 100\%$. First, the results from the classifiers of Texas ratio over one hundred percent in the next quarter will be evaluated. Thereafter the results of bank failure in the following quarter will be addressed, followed by the results for a one year and two year prediction horizon. In the next section, variable importance will be addressed. For predicting a Texas ratio distress event one quarter ahead, the base model that had the lowest performance metrics is Naïve Bayes (NB). Although the AUC estimate of 80.91% is not the worst, the sensitivity is the highest (77.78%), the low precision (0.22%) indicates this model has a large amount of errors when predicting positive instances. Looking at the difference between performance when including the DD indicator, the NB algorithm performed worse by 0.08% in the AUC. Possible explanations for this performance lie in the assumptions of Naïve Bayes models. NB assumes independence of all variables that predict the outcome. This assumption does not hold in most cases, as variables are often correlated in some way. Next is the K-Nearest Neighbor base model. Again this model has a high AUC estimate (84.41%), but a moderate sensitivity (61.11%). Meaning, out of all true bank distress observations, KNN correctly identified just more than half of them. This is combined with a precision rate of 1.18%, meaning out of all predicted distress observations only 1.18% were true positives. If the DD indicator is added, the AUC performance of the KNN model changes by just 0.16% in the negative direction. A drawback from KNN is that it does not learn, but memorizes distances, and it is sensitive to irrelevant variables. The third model we evaluate is XGBoost. The AUC estimate is the lowest out of all models (73.98%). The sensitivity of 50.0% indicates XGBoost correctly identified 50% of all true positive distress observations. The precision rate of 3.17% means XGBoost performed better when predicting positive instances than both NB and KNN. When the DD indicator is included, there is an ambiguous change in performance metrics. The model predicts less false positives, however it also identifies less true positive instances. This results in a drop in the AUC estimate of 5.35%. The final model is Random Forest, which produced the highest AUC metric. The AUC for the base model is measured at 95.03%. The RF model predicted less true positive instances of bank distress than the XGBoost model, but the RF model also predicted less false positives than the XGBoost model. Taking the DD indicator into account, the RF model shows a slight improvement when compared to the RF base model. With the DD added as an independent variable, the RF AUC increased by 0.88%. There are some caveats to the results, which will be discussed in a later section. These findings are not an indication that the addition of a market variable improves models when predicting bank distress, so they do not support the hypothesis. It could be expected that variables containing information not yet covered by other variables would improve the predictive accuracy of the outcome variable. However, the economic effect of adding the DD indicator for this model is questionable.

Table 5
Panel A: Results of machine learning predictions shown by distress event and model.

Distress event	**Model***	**TP**	**FP**	**TN**	**FN**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
$Texas\ ratio_{q+1} > 100\%$	KNN Base	11	925	12619	7	0.8441	0.6111	0.9317	0.0118	0.9313
	KNN DD	11	967	12577	7	0.8425	0.6111	0.9286	0.0112	0.9282
	NB Base	14	6335	7209	4	0.8091	0.7778	0.5323	0.0022	0.5326
	NB DD	14	6332	7212	4	0.8083	0.7778	0.5325	0.0022	0.5328
	RF Base	4	60	13484	14	0.9503	0.2222	0.9956	0.0625	0.9945
	RF DD	5	61	13483	13	0.9591	0.2778	0.9955	0.0758	0.9945
	XGB Base	9	275	13269	9	0.7398	0.5000	0.9797	0.0317	0.9791
	XGB DD	7	221	13323	11	0.6863	0.3889	0.9837	0.0307	0.9829
$Bank\ failure_{q+1}$	KNN Base	1	74	14037	6	0.7080	0.1429	0.9948	0.0133	0.9943
	KNN DD	1	73	14038	6	0.7081	0.1429	0.9948	0.0135	0.9944
	NB Base	0	16	14095	7	0.9470	0.0000	0.9989	0.0000	0.9984
	NB DD	0	18	14093	7	0.9491	0.0000	0.9987	0.0000	0.9982
	RF Base	0	9	14102	7	0.9047	0.0000	0.9994	0.0000	0.9989
	RF DD	0	8	14103	7	0.9052	0.0000	0.9994	0.0000	0.9989
	XGB Base	0	32	14079	7	0.4989	0.0000	0.9977	0.0000	0.9972
	XGB DD	0	32	14079	7	0.4989	0.0000	0.9977	0.0000	0.9972

* KNN is k-nearest neighbor, NB is naive bayes, RF is randomforest and XGB is XGBoost.

Panel B: Model performance metrics

Metric	Explanation	Formula
AUC	AUC stands for the area under the ROC curve. It is the probability that a random chosen positive instance will be ranked ahead of a randomly chosen negative instance.	$\frac{\sum_1^{N_{D0}} \sum_1^{N_{D1}} S(D_0, D_1)}{N_{D0} * N_{D1}}$
Sensitivity	Sensitivity is 1 - Type I error. It is the rate at which a positive prediction is indeed positive.	$\frac{TP}{TP + FN}$
Specificity	Specificity is 1 - Type II error. It is the rate at which a negative prediction is indeed negative.	$\frac{TN}{TN + FP}$
Precision	Precision is also called the "positive predictive value". It is the ratio of true positive predictions to all positive predictions.	$\frac{TP}{TP + FP}$
Accuracy	Overall accuracy is the ratio of all correct predictions to all predictions made.	$\frac{TP + TN}{TP + FP + TN + FN}$

Table 5 also shows the results for the different algorithms concerning failed banks in the following quarter. For failed banks the true positive rates are lower than those resulting from the Texas ratio testing. The NB, KNN and RF models showed improvement in prediction performance measured by the AUC when the DD was included, while it remained unchanged for the XGBoost models. All algorithms showed change of less than one percent for this model. These results should be treated carefully, because there are few instances on which to test on. The imbalance in the training set is again addressed with SMOTE, but the small number of bank failure observations make it harder to draw significant results. The small number of observations also affect the changes in performance metrics. With only twelve true observations of bank failure in the following quarter for training and test set combined, the effect of one additional positive prediction is far greater than when more data is available. This means no definitive conclusion can be drawn whether these results support or reject the hypothesis.

Following these results, table 6 shows the results when a longer prediction horizon is applied to assess the additional predictive power of the DD indicator. Here $Texas\, ratio_{q+4} > 100\%$ was set as the dependent variable, denoting a distress event one year ahead of the independent variables used to predict this distress event. As the prediction horizon is larger, data from a single bank spanning a longer time span is needed. This data is not available for all banks in the sample. Therefore the amount of true positive instances in the sample is lower for this test than the test with just one quarter horizon. Out of the four tested algorithms, NB predicted the most true positive instances of bank distress one year in advance. This coincided with a large amount of false positives. The addition of the DD indicator showed improvement for all models. The AUC of the KNN improved by 5.93%, the AUC of the NB model improved by 1.20%, the AUC of the RF model improved by 3.22%, and the AUC of the XGBoost model improved by 0.08%. A market indicator like the DD can be expected to be a stronger indicator for an event more into the future, because the information embedded in market signals is forward-looking, instead of the backward-looking information embedded in established risk indicators based on accounting information. The results from these tests are an indication this is indeed the case, and support the hypothesis that adding a market variable improves predictions of bank distress. To further test the additional predictive power of the DD indicator, a longer than one year time horizon was taken. Table 6 also shows the results when predicting bank distress two years into the future. The AUC from the RF model is the highest. When including the DD indicator, three models improved in the AUC estimate. However, all changes were below 1%, therefore cannot be considered as economically significant.

A longer prediction horizon was also applied to algorithms predicting bank failure. For a one year horizon the RF and XGBoost models showed no improvement when the DD indicator was included. The NB model and KNN model improved their AUC estimate, but the difference was less than 1%. It should be noted that these improvements coincided with no true positive predictions, except for the KNN model, therefore questioning the performance of these models. A two year prediction horizon for bank failure showed stronger differences when including the DD indicator. XGBoost showed no change and the KNN AUC decreased by 2.11%. The NB and RF models showed improvements of 1.99% and 4.95% respectively. These predictions again coincided with no true positive predictions. The results from table 6 are an indication that the DD has additional predictive power with a longer predictive horizon, therefore supporting the hypothesis. The economic effect of this improvement is questionable, as not one algorithm showed one additional true positive prediction with the addition of the DD indicator. Figure 3 shows the ROC curves for $Texas\, ratio_{q+4} > 100\%$ as distress event, the ROC graphs for the other three models from table 6 are shown in appendix A.

Table 6
Results of machine learning predictions shown by distress event and model for a one and two year prediction horizon.

Distress event	Model*	TP	FP	TN	FN	AUC	Sensitivity	Specificity	Precision	Accuracy
$Texas\ ratio_{q+4} > 100\%$	KNN Base	3	796	12758	5	0.6864	0.3750	0.9413	0.0038	0.9409
	KNN DD	3	800	12754	5	0.7457	0.3750	0.9410	0.0037	0.9406
	NB Base	5	5891	7663	3	0.6223	0.6250	0.5654	0.0008	0.5654
	NB DD	5	5831	7723	3	0.6343	0.6250	0.5698	0.0009	0.5698
	RF Base	0	49	13505	8	0.7729	0.0000	0.9964	0.0000	0.9958
	RF DD	0	47	13507	8	0.8051	0.0000	0.9965	0.0000	0.9959
	XGB Base	1	134	13420	7	0.5576	0.1250	0.9901	0.0074	0.9896
	XGB DD	1	110	13444	7	0.5584	0.1250	0.9919	0.0090	0.9914
$Texas\ ratio_{q+8} > 100\%$	KNN Base	0	966	12595	1	0.8185	0.0000	0.9288	0.0000	0.9287
	KNN DD	0	989	12572	1	0.8139	0.0000	0.9271	0.0000	0.9270
	NB Base	0	5962	7599	1	0.0452	0.0000	0.5604	0.0000	0.5603
	NB DD	0	5897	7664	1	0.0464	0.0000	0.5652	0.0000	0.5651
	RF Base	0	57	13504	1	0.9889	0.0000	0.9958	0.0000	0.9957
	RF DD	0	60	13501	1	0.9890	0.0000	0.9956	0.0000	0.9955
	XGB Base	0	492	13069	1	0.4819	0.0000	0.9637	0.0000	0.9636
	XGB DD	0	452	13109	1	0.4833	0.0000	0.9667	0.0000	0.9666
$Bank\ failure_{q+4}$	KNN Base	1	887	13227	3	0.8127	0.2500	0.9372	0.0011	0.9370
	KNN DD	1	889	13225	3	0.8138	0.2500	0.9370	0.0011	0.9368
	NB Base	0	92	14022	4	0.4786	0.0000	0.9935	0.0000	0.9932
	NB DD	0	117	13997	4	0.4813	0.0000	0.9917	0.0000	0.9914
	RF Base	0	91	14023	4	0.9342	0.0000	0.9936	0.0000	0.9933
	RF DD	0	100	14014	4	0.9293	0.0000	0.9929	0.0000	0.9926
	XGB Base	0	179	13935	4	0.4937	0.0000	0.9873	0.0000	0.9870
	XGB DD	0	179	13935	4	0.4937	0.0000	0.9873	0.0000	0.9870
$Bank\ failure_{q+8}$	KNN Base	0	1063	13053	2	0.6660	0.0000	0.9247	0.0000	0.9246
	KNN DD	0	1194	12922	2	0.6449	0.0000	0.9154	0.0000	0.9153
	NB Base	0	512	13604	2	0.4248	0.0000	0.9637	0.0000	0.9636
	NB DD	0	578	13538	2	0.4447	0.0000	0.9591	0.0000	0.9589
	RF Base	0	38	14078	2	0.8415	0.0000	0.9973	0.0000	0.9972
	RF DD	0	42	14074	2	0.8910	0.0000	0.9970	0.0000	0.9969
	XGB Base	0	235	13881	2	0.4917	0.0000	0.9834	0.0000	0.9832
	XGB DD	0	235	13881	2	0.4917	0.0000	0.9834	0.0000	0.9832

* KNN is k-nearest neighbor, NB is naive bayes, RF is randomforest and XGB is XGBoost.

4.2 Variable importance

One important factor of empirical research is model interpretability. The decisions makers (regulators, investors) will want to understand how they can infer knowledge from underlying causal mechanisms. For randomforest models we can gain an understanding about the importance of the different variables by looking at the mean decrease in Gini index. First, one must understand that gini impurity is a measure for decision trees, and consequently tree ensemble models, that indicates how important variables are for predicting the dependent variable (distress in this case). The mean decrease in gini then measures the decrease in decision node impurity for a variable, and averages this decrease across all decision trees in the forest. A relatively high value of mean gini decrease means that a particular variable reduces the impurity more than other variables, and is more important in determining the outcome variable. Table 7 shows the mean gini decrease per variable for all models. To get a better feeling of the relative importance of all variables, the related ranks are shown for all variables. When looking at the $Texas\,ratio_{q+1}$ model, one can see the DD variable has a mean decrease in gini of 31.4957 and ranks 14th among the other variables for predicting bank distress. Particularly interesting in these importance ranks is the difference between $Texas\,ratio_{q+1}$ and the $Texas\,ratio_{q+4}$ and the $Texas\,ratio_{q+8}$ models. Here we can see that the DD indicator becomes more important, when predicting over a longer time horizon, indicated by the jump from rank fourteen to five and nine. For predicting bank failure, we can also see a clear progression when predicting over a longer horizon. As the DD for $Bank\,failure_{q+1}$ is ranked number eighteen, this rank jumps to thirteen for $Bank\,failure_{q+4}$, and thirteen for $Bank\,failure_{q+8}$. Also included is the sum of ranks to see which variable was most important over all the models. E1 and E2 have the lowest sum of ranks with a value of forty-two. This means the variable calculated by dividing net income after extraordinary items over total assets (E1) is on average the most important factors for predicting bank distress in the random forest models, together with the variable earnings per share, diluted, excluding extraordinary items and 12 months moving (E2).

Table 7
Variable importance for randomforest models, calculated with the Mean Decrease in Gini index. Per variable and model the mean decrease in gini index is shown. The higher numbers mean a larger decrease, indicating these independent variables are more important in predicting the dependent variable for that model. The ranks are shown to indicate how important the different variables are in relation to the other variables.

	$Texas\,ratio_{q+1} > 100\%$		$Bank\,failure_{q+1}$		$Texas\,ratio_{q+4} > 100\%$		$Texas\,ratio_{q+8} > 100\%$		$Bank\,failure_{q+4}$		$Bank\,failure_{q+8}$		
Variable	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Sum of ranks
DD	31.4957	14	0.1157	18	52.3608	5	13.1956	9	0.4526	13	0.5278	13	72
C1	33.4202	13	25.8376	2	19.1585	16	26.5083	2	0.8387	10	0.5505	12	55
C2	76.5280	5	14.7513	4	21.5954	13	7.5104	12	0.3258	15	0.4872	14	63
C3	25.0994	18	0.3119	14	19.8657	14	2.5218	21	0.0122	22	2.2508	5	94
A1	61.7777	6	1.0194	12	57.3079	3	7.0004	14	1.4627	9	1.9707	7	51
A2	45.4703	10	8.7413	5	17.6641	17	6.4435	15	3.9650	5	2.0400	6	58
A3	46.4200	9	0.3219	13	14.6238	20	5.3095	17	0.0934	19	0.4122	15	93
A4	253.4886	1	1.2449	10	77.4958	2	21.6499	5	2.9725	6	0.1192	21	45
M1	40.8553	11	2.2724	9	14.3284	21	8.1097	11	0.0740	20	0.2940	19	91
E1	155.5406	3	30.9189	1	77.8400	1	3.1360	20	8.5370	1	0.3917	16	42
E2	51.5350	7	6.6319	6	56.2036	4	3.5942	18	4.0133	4	3.4428	3	42
E3	25.8788	16	2.3867	8	41.1081	9	2.3844	22	0.1457	16	0.6170	11	82
E4	26.4673	15	6.4075	7	19.8158	15	12.8119	10	2.4052	7	1.2071	10	64
E5	182.4812	2	16.2765	3	31.0554	11	13.3028	8	6.3130	2	0.3098	18	44
E6	24.6508	19	0.1831	15	48.3405	7	14.4259	7	2.2217	8	3.5029	2	58
L1	13.9215	21	0.1280	17	23.0503	12	5.9856	16	0.7205	11	4.3883	1	78
L2	7.0090	22	0.0093	22	5.7805	22	41.2657	1	0.1065	17	0.0990	22	106
L3	25.7562	17	0.0184	21	43.9571	8	24.8618	3	0.3956	14	0.3665	17	80
S1	140.6045	4	0.0511	20	37.3641	10	7.2916	13	0.5856	12	2.5235	4	63
X1	17.6434	20	0.0733	19	50.7625	6	18.6885	6	0.0999	18	1.2988	9	78
X2	47.2374	8	1.0807	11	14.6858	19	3.1977	19	0.0680	21	0.2714	20	98
X3	38.3379	12	0.1754	16	17.4235	18	24.7265	4	5.1653	3	1.8413	8	61

Because XGBoost is similar to random forest as a decision tree ensemble model, the same variable importance can be extracted from the trained models. In table 8, gain is stated as the contribution to the model for a particular variable, relative to other variables in the same model. The higher values, similar to the random forest importance, mean a more important variable for predicting the outcome variable of bank distress. In correlation with the results from table 7, we see an increase in importance of the DD indicator when the prediction horizon moves further into the future. For the first model, $Texas\ ratio_{q+1}$, the DD counts for a gain of 2.54% in the model and is ranked number eleven. For the one year horizon, the DD jumps to a gain of 6.45% and is ranked number six, and for a two year prediction horizon the DD has a gain of 4.00% and has rank ten. When looking at bank failure predictions we see no improvement in rank for the DD indicator. For bank failure the DD indicator is excluded because the importance was below the threshold to be included. Table 8 also shows the sum of ranks. Similar to table 7, one can see E1 has the lowest sum of ranks, meaning net income after extraordinary items over total assets is on average the most important indicator for predicting bank distress in the XGBoost models.

Table 8
Variable importance for XGBoost models, calculated as the relative contribution of the corresponding variable to a model. The higher numbers indicate these independent variables are more important in predicting the dependent variable for that model. The ranks are shown to indicate how important the different variables are in relation to the other variables.

	$Texas\ ratio_{q+1} > 100\%$		$Bank\ failure_{q+1}$		$Texas\ ratio_{q+4} > 100\%$		$Texas\ ratio_{q+8} > 100\%$		$Bank\ failure_{q+4}$		$Bank\ failure_{q+8}$		
Variable	Gain	Rank	Gain	Rank	Gain	Rank	Gain	Rank	Gain	Rank	Gain	Rank	Sum of ranks**
DD	0.0254	11	*	16	0.0645	6	0.0400	10	*	14	*	19	76
C1	0.0168	14	0.1996	2	0.0395	10	0.0826	3	0.0064	8	0.0843	4	41
C2	0.0566	5	0.0055	5	0.0367	11	0.0079	16	*	19	*	21	77
C3	0.0136	16	*	19	0.0348	13	0.0017	19	*	21	0.1153	3	91
A1	0.0326	8	*	13	0.0491	8	0.0060	17	0.0068	7	0.0026	13	66
A2	0.0224	12	0.0028	6	0.0133	19	0.0525	7	0.0210	5	0.0623	5	54
A3	0.0350	7	0.0009	9	0.0363	12	0.0093	15	*	18	*	20	81
A4	0.3803	1	*	11	0.2154	1	0.0272	11	*	15	0.0035	12	51
M1	0.0260	10	*	15	0.0069	20	0.0649	4	*	11	*	16	76
E1	0.0691	4	0.7655	1	0.0853	3	0.0007	22	0.6596	1	*	15	46
E2	0.0059	21	0.0061	4	0.0232	14	0.0008	21	0.0219	4	0.1692	2	66
E3	0.0116	18	0.0025	7	0.0469	9	0.0030	18	*	20	0.0078	9	81
E4	0.0097	19	*	21	0.0027	21	0.1553	2	0.0084	6	0.0035	11	80
E5	0.0941	2	0.0003	10	0.0021	22	0.0577	5	*	12	*	17	68
E6	0.0097	20	0.0012	8	0.0671	4	0.0242	12	0.0830	3	0.0297	7	54
L1	0.0055	22	*	22	0.0137	18	0.0225	13	*	16	0.4497	1	92
L2	0.0145	15	*	18	0.0194	15	0.2795	1	*	10	0.0197	8	67
L3	0.0182	13	*	17	0.0599	7	0.0458	8	*	13	*	18	76
S1	0.0752	3	*	12	0.0156	16	0.0437	9	0.0006	9	0.0441	6	55
X1	0.0135	17	*	20	0.0875	2	0.0180	14	*	17	0.0063	10	80
X2	0.0358	6	0.0156	3	0.0142	17	0.0013	20	*	22	*	22	90
X3	0.0285	9	*	14	0.0657	5	0.0554	6	0.1922	2	0.0022	14	50

* Variable excluded because importance was below threshold to be included in output.

** Because some variables are not important predictors for multiple models, and due to random assignment of the corresponding low ranks, the high ranks (low sum of ranks) are more representative of the most important predictors than the low ranks (high sum of ranks) are accurate representations of the least important predictors.

5. Additional modelling and robustness tests

5.1 Implied cost of capital

The first form of additional modelling centers around the theoretical basis for the market variable. The main testing in this paper is done to assess the additional predictive power of the distance to default market variable. Another market variable that has gained more attention in academic literature, is the implied cost of capital (ICC) (Lee, So and Wang, 2011). *"The implied cost of capital (ICC) for a given asset can be defined as the discount rate (or internal rate of return) that equates the asset's market value to the present value of its expected future cash flows"* (Lee, So and Wang, 2011). This could provide a viable alternative to improve prediction results of bank distress. To construct the ICC variable, additional data is gathered from the IBES database with detailed analyst forecasts of future earnings for all US banks in the sample and expected long term growth rates of these earnings. Forecasted earnings are calculated as the average EPS forecasts of all individual analysts per bank quarter observation. Forecasts for the next five years are collected where possible, but only observations are included in the sample with at least forecasts for one year and two years ahead. Following Gebhardt, Daske and Klein (2006), missing forecasts are estimated as $feps_{t+1} = feps_t * (1 + g_t)$ where g_t is the long-term growth rate retrieved from IBES. If g_t is not known, the forecasts for year three, four and five are estimated as follows:

$$feps_3 = feps_2 + \frac{(feps_2 - feps_1)}{1} \qquad feps_4 = feps_3 + \frac{(feps_3 - feps_1)}{2} \qquad feps_5 = feps_4 + \frac{(feps_4 - feps_1)}{3}$$

In the calculation of two ICC estimates for this research, a growth rate is needed as input. This paper will use a rolling average (moving average) over the last two years of quarterly US gross domestic product growth, to proxy for the growth rate per bank quarter observation. Dividend forecasts are calculated by multiplying the earnings forecast with the dividend payout rate. The remaining data, like stock price, was already collected for earlier testing. Three models are then applied to estimate the ICC per bank quarter observation. The first model is known as the Gordon growth model (GGM) and provides a seminal approach in valuing the stock price of a company based on future dividend payments (Gordon and Shapiro, 1956). The model is estimated with the following equations:

$$\text{GGM:} \quad P_0 = \sum_{t=1}^{\infty} D_0 \frac{(1+g)^t}{(1+r)^t} \quad \text{which simplifies to} \quad P_0 = \frac{D_1}{r-g} \quad \text{solve for r gives} \quad r = \frac{D_1}{P_0} + g \qquad (6)$$

Here P_0 is the current stock price, D_0 and D_1 are dividend payments for the current period and one period ahead, g is the growth rate, and r is the implied cost of capital.

The second model is derived from the work of Ohlson and Juettner-Nauroth (2005), and is based on the same principle that current stock price is determined by discounting forecasted earnings. For estimating short term growth, this paper will follow the work of Gode and Mohanram (2003), that takes the average of near term growth $\left(\frac{feps_{t+3} - feps_{t+2}}{feps_{t+2}}\right)$ and five year growth $\left(\frac{feps_{t+5} - feps_{t+4}}{feps_{t+4}}\right)$. The Ohlson and Juettner-Nauroth model (OJM) calculates the ICC with the following equation:

$$\text{OJM:} \quad r = A + \sqrt{A^2 + \frac{feps_{t+1}}{P_0} * \left[\frac{1}{2} * \left(\frac{feps_{t+3} - feps_{t+2}}{feps_{t+2}} + \frac{feps_{t+5} - feps_{t+4}}{feps_{t+4}}\right) - (\gamma - 1)\right]} \qquad (7)$$

$$\text{where} \quad A = \frac{1}{2}\left[(\gamma - 1) - \frac{D_1}{P_0}\right] \qquad (8)$$

Here P_0 is the current stock price, D_1 are dividend payments for one period ahead, γ is the growth rate g, $feps_{t+i}$ are forecasted earnings and r is the implied cost of capital. The original approach of the Ohlson and Juettner-Nauroth model would only use $\left(\frac{feps_{t+2}-feps_{t+1}}{feps_{t+1}}\right)$ as short term growth, now five period forecasts can be implemented into the model.

The third model for estimating the ICC, the price earnings growth (PEG) model comes from the work of Easton (2004). He takes a forecast horizon of just two periods and a growth rate of zero to estimate the implied cost of equity capital. There are two equations that estimate the ICC, one that includes future dividends and the other in the case of no forecasted dividends.

PEG: $D_1 \neq 0 \rightarrow \ r = \sqrt{\frac{(feps_{t+2}+r*D_1-feps_{t+1})}{P_0}}$ solve for r gives $r = \frac{D_1+\sqrt{{D_1}^2-4*{P_0}^2*(feps_{t+1}-feps_{t+2})}}{2*{P_0}^2}$ (9)

(M)PEG[6]: $D_1 = 0 \rightarrow \ r = \sqrt{\frac{(feps_{t+2}-feps_{t+1})}{P_0}}$ (10)

Finally, all positive ICC estimates from the three models are averaged, to form an aggregate ICC estimation. Negative ICC estimates will be dropped for testing, since ICC is in essence a discount rate reflecting risk, and negative risk is nonsensical. This results in samples with 5005 observations for the average ICC estimate, 4436 observations for the GGM ICC, 4391 observations for the PEG ICC and 4872 observations for the OJM ICC, as shown in table 9. Other notable ICC calculation models stem from papers by Claus and Thomas (2001), and by Gebhardt, Lee, and Swaminathan (2002). These models are excluded from this research because they include book value per share in their calculation, besides earnings forecasts. This is a problem because data to calculate book value per share comes from Compustat which is based on GAAP principles and earnings forecasts are retrieved from IBES which is based on Non-GAAP principles. To combine these different methods of estimation would lead to distorted calculations. The additional predictive power of the market indicator (ICC in this case) will again be tested by comparing models that include and exclude the indicator as an independent variable when predicting a Texas ratio over one hundred percent. Two year ahead predictions are not included because the data was not sufficient. The results are shown in table 10. The corresponding ROC curves are shown in appendix H. Appendix I contains table 17 with results for one quarter predictions with ICC and table 18 with the corresponding variable importance for one quarter predictions.

[6] MPEG stands for the modified price earnings growth model.

Table 9
Descriptive statistics for the implied cost of capital and equity beta.

Variable	Average ICC	GGM ICC	PEG ICC	OJM ICC	Beta
Min	0.000	0.000	0.000	0.000	-1.689
1st Quartile	0.046	0.020	0.026	0.056	0.283
Median	0.058	0.026	0.072	0.073	0.684
Mean	0.070	0.031	0.091	0.074	0.706
3rd Quartile	0.072	0.038	0.107	0.088	1.046
Max	1.000	0.483	1.000	1.000	3.536
Total n	5005	4436	4391	4872	17034

The results from table 10 are an indication the ICC can improve models predicting bank distress. The timing of this added value is similar to the results from table 5 and table 6. As earlier results show the DD market indicator had little added value over a one quarter horizon prediction, the average ICC market indicator shows no significant improvement for this horizon. However, for a one year prediction horizon, the average ICC in table 10 shows more improvement for a single model than previously calculated. The NB AUC increases 0.17%, the XGBoost AUC increases 2.23%, the KNN AUC increases 5.11%, and the RF AUC increases by 9.72%. This shows improvement that could be considered as economically significant. When looking at the individual ICC estimates, we see a breakdown of the prediction improvement per ICC estimate. The most improvement in the AUC estimate when adding the ICC variable is shown by the OJM KNN and PEG KNN models. The AUC improves by 10.84% and 10.81% respectively. Table 11 shows the variable importance for the RF algorithms testing these ICC models, so the following statements only apply to the random forest algorithm. Table 11 contradicts previous results, because it shows the average ICC is more important for predicting bank distress one quarter ahead than predicting one year ahead. For one quarter ahead, the average ICC is the most important factor by a clear margin. For one year ahead, the ICC has rank number 2. The highest rank of the DD variable was five for predicting one year ahead. These results are an indication a market variable has added value when prediction bank distress, therefore supporting the hypothesis. Moreover, these results show the average ICC estimate could be superior to the distance to default as risk estimator. Table 11 also shows the average ICC is ranked highest for predicting one year ahead, compared with individual ICC model estimates. For the ICC models, variable A4 (real estate other than bank premises over total assets) is the most important, indicated by the lowest sum of ranks. One important drawback from ICC estimation models is their reliance on earnings growth. This, combined with the forecast bias of analysts (Gu and Wu, 2003), is an important consideration when evaluating these results.

Table 10

Results of machine learning predictions shown for Texas ratio distress event in q+1 and q+4 per method of ICC calculation.

Distress event	ICC*	**Model****	**TP**	**FP**	**TN**	**FN**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
$Texas\,ratio_{q+1} > 100\%$	Average	KNN Base	3	142	4579	3	0.8910	0.5000	0.9699	0.0207	0.9693
	Average	KNN ICC	3	120	4601	3	0.8935	0.5000	0.9746	0.0244	0.9740
	Average	NB Base	3	482	4239	3	0.8644	0.5000	0.8979	0.0062	0.8974
	Average	NB ICC	3	424	4297	3	0.8651	0.5000	0.9102	0.0070	0.9097
	Average	RF Base	1	8	4713	5	0.9729	0.1667	0.9983	0.1111	0.9972
	Average	RF ICC	0	5	4716	6	0.9673	0.0000	0.9989	0.0000	0.9977
	Average	XGB Base	1	68	4653	5	0.5761	0.1667	0.9856	0.0145	0.9846
	Average	XGB ICC	0	7	4714	6	0.4993	0.0000	0.9985	0.0000	0.9972
$Texas\,ratio_{q+4} > 100\%$	Average	KNN Base	0	1281	3443	3	0.4793	0.0000	0.7288	0.0000	0.7284
	Average	KNN ICC	0	1165	3559	3	0.5304	0.0000	0.7534	0.0000	0.7529
	Average	NB Base	0	104	4620	3	0.3968	0.0000	0.9780	0.0000	0.9774
	Average	NB ICC	0	103	4621	3	0.3985	0.0000	0.9782	0.0000	0.9776
	Average	RF Base	0	23	4701	3	0.7695	0.0000	0.9951	0.0000	0.9945
	Average	RF ICC	0	24	4700	3	0.8667	0.0000	0.9949	0.0000	0.9943
	Average	XGB Base	1	321	4403	2	0.6327	0.3333	0.9320	0.0031	0.9317
	Average	XGB ICC	1	110	4614	2	0.6550	0.3333	0.9767	0.0090	0.9763
$Texas\,ratio_{q+4} > 100\%$	GGM	KNN Base	0	1094	3630	3	0.4836	0.0000	0.7684	0.0000	0.7679
	GGM	KNN ICC	0	1063	3661	3	0.4907	0.0000	0.7750	0.0000	0.7745
	GGM	NB Base	0	273	4451	3	0.4151	0.0000	0.9422	0.0000	0.9416
	GGM	NB ICC	0	223	4501	3	0.4526	0.0000	0.9528	0.0000	0.9522
	GGM	RF Base	1	141	4583	2	0.7528	0.3333	0.9702	0.0070	0.9697
	GGM	RF ICC	0	96	4628	3	0.7589	0.0000	0.9797	0.0000	0.9791
	GGM	XGB Base	1	1161	3563	2	0.5438	0.3333	0.7542	0.0009	0.7540
	GGM	XGB ICC	1	1161	3563	2	0.5438	0.3333	0.7542	0.0009	0.7540
$Texas\,ratio_{q+4} > 100\%$	OJM	KNN Base	0	764	3960	3	0.4575	0.0000	0.8383	0.0000	0.8377
	OJM	KNN ICC	0	765	3959	3	0.5659	0.0000	0.8381	0.0000	0.8375
	OJM	NB Base	0	70	4654	3	0.4062	0.0000	0.9852	0.0000	0.9846
	OJM	NB ICC	0	77	4647	3	0.4061	0.0000	0.9837	0.0000	0.9831
	OJM	RF Base	0	28	4696	3	0.7635	0.0000	0.9941	0.0000	0.9934
	OJM	RF ICC	0	28	4696	3	0.7741	0.0000	0.9941	0.0000	0.9934
	OJM	XGB Base	0	758	3966	3	0.4198	0.0000	0.8395	0.0000	0.8390
	OJM	XGB ICC	0	758	3966	3	0.4198	0.0000	0.8395	0.0000	0.8390
$Texas\,ratio_{q+4} > 100\%$	PEG	KNN Base	1	565	4159	2	0.7229	0.3333	0.8804	0.0018	0.8801
	PEG	KNN ICC	2	586	4138	1	0.8310	0.6667	0.8760	0.0034	0.8758
	PEG	NB Base	0	226	4498	3	0.4359	0.0000	0.9522	0.0000	0.9516
	PEG	NB ICC	0	223	4501	3	0.4874	0.0000	0.9528	0.0000	0.9522
	PEG	RF Base	1	41	4683	2	0.8291	0.3333	0.9913	0.0238	0.9909
	PEG	RF ICC	1	40	4684	2	0.8694	0.3333	0.9915	0.0244	0.9911
	PEG	XGB Base	1	529	4195	2	0.6107	0.3333	0.8880	0.0019	0.8877
	PEG	XGB ICC	1	529	4195	2	0.6107	0.3333	0.8880	0.0019	0.8877

* Implied Cost of Capital (ICC) calculation models; Average is the average of positive ICC values per bank quarter observation (5005 total observations), GGM is the Gordon Growth Model (4436 total observations), OJM is the Ohlson-Jeuttner-Nauroth Model (4872 total observations) and PEG is the Price Earnings Growth Model (4391 total observations).

** KNN is k-nearest neighbor, NB is naive bayes, RF is randomforest and XGB is XGBoost.

Table 11

Variable importance for randomforest models, calculated with the Mean Decrease in Gini index, for different implied cost of capital models as market variable. Per variable and model the mean decrease in gini index is shown. The higher numbers mean a larger decrease, indicating these independent variables are more important in predicting the dependent variable for that model. The ranks are shown to indicate how important the different variables are in relation to the other variables.

	$Texas\ ratio_{q+1} > 100\%$		$Texas\ ratio_{q+4} > 100\%$		$Texas\ ratio_{q+4} > 100\%$		$Texas\ ratio_{q+4} > 100\%$		$Texas\ ratio_{q+4} > 100\%$		
	Average ICC		**Average ICC**		**GGM***		**OJM***		**PEG***		
Variable	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Sum of ranks
ICC	43.1891	1	4.3097	2	0.9054	11	0.1235	22	0.7253	12	48
C1	8.4252	7	1.0256	13	0.8735	12	0.6695	12	0.6434	13	57
C2	2.3486	14	2.0044	7	1.4430	6	4.8275	3	3.6101	3	33
C3	0.7819	22	0.0926	22	0.0568	22	0.2782	18	0.0837	22	106
A1	7.4910	9	1.1968	11	0.3186	17	0.7503	11	0.1217	21	69
A2	2.1092	17	0.2938	17	0.4170	14	0.1739	20	0.3356	19	87
A3	3.1248	13	3.0792	5	1.3400	7	3.5016	4	2.0720	6	35
A4	14.1411	6	6.6738	1	1.8076	4	5.5807	2	4.8314	2	15
M1	2.2120	16	3.2491	4	2.7477	3	1.5941	7	0.3806	18	48
E1	17.8156	5	0.6620	15	1.0306	9	0.4254	14	0.2237	20	63
E2	4.5810	10	0.1350	20	0.1967	21	0.1704	21	1.1105	11	83
E3	2.0792	19	4.1827	3	8.7269	1	9.4350	1	3.1052	4	28
E4	0.8022	21	1.9294	8	0.5566	13	1.3694	9	1.8433	7	58
E5	29.5873	2	1.4900	10	0.9991	10	0.8610	10	1.3745	10	42
E6	2.0920	18	0.1238	21	0.3858	16	0.4199	16	0.6322	14	85
L1	3.2719	12	1.0337	12	0.2966	18	0.5547	13	0.5379	16	71
L2	2.3110	15	0.2329	18	0.2916	19	0.3945	17	0.3971	17	86
L3	8.2333	8	0.8807	14	1.6648	5	1.4971	8	2.0784	5	40
S1	23.0686	4	0.3421	16	0.3896	15	0.4217	15	1.6320	9	59
X1	1.6116	20	0.1737	19	0.1971	20	0.2647	19	0.5749	15	93
X2	4.3212	11	1.5479	9	1.3218	8	2.0109	6	1.6979	8	42
X3	29.2876	3	2.3034	6	5.9650	2	2.6169	5	9.9587	1	17

* GGM is the Gordon Growth Model, OJM is the Ohlson-Jeuttner-Nauroth Model and PEG is the Price Earnings Growth

5.2 Equity volatility

So far two market variables have been tested in this paper as potential additions for improving bank distress predictions. The distance to default and implied cost of capital indicators are very comprehensive. This makes them a very sophisticated approximation of the underlying risk for a bank, on the other hand, they are difficult to calculate and are prone to measurement error. One way to solve this is to take a much more simple approach to estimate risk and find an elementary proxy to use for predicting bank distress. The volatility of equity, indicated by the beta, gives such a proxy. We will use the same beta as used in earlier calculations. This is the average beta per quarter is taken from daily estimated beta's with a 252 day estimation window in the WRDS Beta Suite. By comparing models with the beta as added market variable, its viability as predictive indicator can be assessed. The results from these tests are shown in table 12, the corresponding ROC graphs are shown in appendix J. Similar to earlier testing, the most improvement is calculated for a one year prediction horizon. The difference in AUC estimate for NB and XGBoost models is marginal. However, the RF AUC improves by 2.72% and the KNN AUC improves by 5.66%. Appendix J also shows the importance for random forest models including the beta variable. The importance rank of the beta increases for predicting one year ahead, compared to predicting one quarter ahead. However the importance drops again for a two year ahead prediction. These results are an indication a market variable can improve bank distress predictions, therefore supporting the hypothesis. It is questionable if the beta provides a superior risk estimator over the distance to default or implied cost of capital, as they produce similar results.

Table 12
Results of machine learning predictions shown for Texas ratio distress event with the equity beta as additional market variable.

Distress event	**Model***	**TP**	**FP**	**TN**	**FN**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
$Texas\ ratio_{q+1} > 100\%$	KNN Base	13	921	14131	7	0.8059	0.6500	0.9388	0.0139	0.9384
	KNN Beta	11	953	14099	9	0.8154	0.5500	0.9367	0.0114	0.9362
	NB Base	15	7222	7830	5	0.7329	0.7500	0.5202	0.0021	0.5205
	NB Beta	15	7230	7822	5	0.7327	0.7500	0.5197	0.0021	0.5200
	RF Base	6	70	14982	14	0.9512	0.3000	0.9953	0.0789	0.9944
	RF Beta	5	64	14988	15	0.9465	0.2500	0.9957	0.0725	0.9948
	XGB Base	7	183	14869	13	0.6689	0.3500	0.9878	0.0368	0.9870
	XGB Beta	7	213	14839	13	0.6679	0.3500	0.9858	0.0318	0.9850
$Texas\ ratio_{q+4} > 100\%$	KNN Base	4	1025	14039	4	0.8032	0.5000	0.9320	0.0039	0.9317
	KNN Beta	4	947	14117	4	0.8598	0.5000	0.9371	0.0042	0.9369
	NB Base	5	7357	7707	3	0.5345	0.6250	0.5116	0.0007	0.5117
	NB Beta	5	7338	7726	3	0.5339	0.6250	0.5129	0.0007	0.5129
	RF Base	0	59	15005	8	0.8262	0.0000	0.9961	0.0000	0.9956
	RF Beta	1	58	15006	7	0.8534	0.1250	0.9961	0.0169	0.9957
	XGB Base	0	137	14927	8	0.4955	0.0000	0.9909	0.0000	0.9904
	XGB Beta	0	127	14937	8	0.4958	0.0000	0.9916	0.0000	0.9910
$Texas\ ratio_{q+8} > 100\%$	KNN Base	1	1408	13663	0	0.9202	1.0000	0.9066	0.0007	0.9066
	KNN Beta	1	1530	13541	0	0.9137	1.0000	0.8985	0.0007	0.8985
	NB Base	0	6958	8113	1	0.0414	0.0000	0.5383	0.0000	0.5383
	NB Beta	0	6883	8188	1	0.0413	0.0000	0.5433	0.0000	0.5433
	RF Base	0	183	14888	1	0.5512	0.0000	0.9879	0.0000	0.9878
	RF Beta	0	168	14903	1	0.5552	0.0000	0.9889	0.0000	0.9888
	XGB Base	0	692	14379	1	0.4770	0.0000	0.9541	0.0000	0.9540
	XGB Beta	0	711	14360	1	0.4764	0.0000	0.9528	0.0000	0.9528

* KNN is k-nearest neighbor, NB is naive bayes, RF is randomforest and XGB is XGBoost.

5.3 Too-Big-To-Fail distortions

One issue raised in the literature comes from the notion of ‘Too-Big-To-Fail’ banks. These banks will be supported by the government to avoid the widespread consequences in case they would otherwise fail. This is an ambiguous problem, because the public safety net provided by government support for banks covers losses for debt holders, which does not affect the shareholders. However, the effect of this support for the shareholders correlates to the risk taking by banks. Afonso, Santos and Traina (2014) found that banks with higher support ratings would have higher impaired loan ratios. Because higher risk would also bring higher potential returns, this government support could introduce a moral hazard problem. Another effect this support could have is that risk assessments could be distorted, therefore impairing the correct assessment of financial stability of a financial institution. As an additional test, this paper will add a factor variable representing the levels of outside support. The data for this comes from the Fitch bank support rating[7]. Models predicting a bank failure will be evaluated on whether adding the support rating as an independent variable has changed the results from earlier testing. These results can be seen in table 14 in appendix B. Variable importance for random forest models is shown in table 15 in appendix B. These results show no clear distortions from the support rating variable for predicting bank distress. Table 15 also shows the support rating variable is the least important over all three models, indicated by the highest sum of ranks. This result confirms the conclusion by Miller et al. (2015), that Too-Big-To-Fail distortions have no effect on the distance to default testing. Earlier comments about the ambiguity of this issue should again be taken into account when considering these results. The effects shown in these results, do not guarantee these distortions have no effect. It should be noted that the data for this test was not adequate to provide robust conclusions. Appendix F shows the ROC graphs for models predicting Texas ratio distress with Fitch bank support rating as added variable.

5.4 Resampling robustness test

To provide some robustness to the results, additional testing was performed that changed parts of the basis of the methodology. One robustness test is to use multiple resampling techniques, and assess how much the results change for one model. Each time a new training sample is created, new observations are created for the minority class and random observations discarded from the majority class. The three best performing resampling techniques from table 3 (SMOTE, oversampling, and both over- and undersampling) are used to test the robustness of the results shown in table 6. These results are also compared to results when no resampling is applied. The model used for this has $Texas\,ratio_{q+4} > 100\%$ as the dependent variable. Table 13 shows the performance metrics for different resamples of the training data used as input. These results reconcile with the earlier results from SMOTE set 1, also shown in table 13. The majority of improvements shown by adding the DD variable are still marginal for predicting one year ahead. No resampling results in true positive predictions, except for NB which also has a lot of false positives, which indicates resampling improves the predictions of the minority class. This minority class is the class of interest because it contains the actual banks in distress. No resampling shows better performance metrics than oversampling and both over- and undersampling. SMOTE, however, still shows the best results for predicting the minority class. Also applying a different SMOTE resampled training set does not change the results significantly. XGBoost still shows the most improvement, followed by the RF algorithm when adding the DD indicator. The ROC graphs for these test are shown in appendix D.

[7] Fitch bank support ratings indicate the probability a bank will receive external support and is used as a proxy for ‘Too-Big-To-Fail’ indication (Schaeck, Zhou and Molyneux, 2010). The rating ranges from 1 (high likelihood of support) to 5 (low likelihood of support).

Table 13
Results of machine learning predictions shown for Texas ratio distress event in q+4 for different resampling methods applied to the training set.

Distress event	Resample method	Model*	TP	FP	TN	FN	AUC	Sensitivity	Specificity	Precision	Accuracy
$Texas\ ratio_{q+4} > 100\%$	SMOTE 1	KNN Base	3	796	12758	5	0.6864	0.3750	0.9413	0.0038	0.9409
	SMOTE 1	KNN DD	3	800	12754	5	0.7457	0.3750	0.9410	0.0037	0.9406
	SMOTE 1	NB Base	5	5891	7663	3	0.6223	0.6250	0.5654	0.0008	0.5654
	SMOTE 1	NB DD	5	5831	7723	3	0.6343	0.6250	0.5698	0.0009	0.5698
	SMOTE 1	RF Base	0	49	13505	8	0.7729	0.0000	0.9964	0.0000	0.9958
	SMOTE 1	RF DD	0	47	13507	8	0.8051	0.0000	0.9965	0.0000	0.9959
	SMOTE 1	XGB Base	1	134	13420	7	0.5576	0.1250	0.9901	0.0074	0.9896
	SMOTE 1	XGB DD	1	110	13444	7	0.5584	0.1250	0.9919	0.0090	0.9914
$Texas\ ratio_{q+4} > 100\%$	SMOTE 2	KNN Base	3	1076	12478	5	0.6672	0.3750	0.9206	0.0028	0.9203
	SMOTE 2	KNN DD	4	1081	12473	4	0.6706	0.5000	0.9202	0.0037	0.9200
	SMOTE 2	NB Base	5	5504	8050	3	0.6233	0.6250	0.5939	0.0009	0.5939
	SMOTE 2	NB DD	5	5392	8162	3	0.6357	0.6250	0.6022	0.0009	0.6022
	SMOTE 2	RF Base	0	31	13523	8	0.8189	0.0000	0.9977	0.0000	0.9971
	SMOTE 2	RF DD	0	29	13525	8	0.8300	0.0000	0.9979	0.0000	0.9973
	SMOTE 2	XGB Base	1	112	13442	7	0.5584	0.1250	0.9917	0.0088	0.9912
	SMOTE 2	XGB DD	1	93	13461	7	0.5591	0.1250	0.9931	0.0106	0.9926
$Texas\ ratio_{q+4} > 100\%$	Oversampling	KNN Base	1	118	13436	7	0.5526	0.1250	0.9913	0.0084	0.9908
	Oversampling	KNN DD	1	125	13429	7	0.5513	0.1250	0.9908	0.0079	0.9903
	Oversampling	NB Base	7	8243	5311	1	0.6601	0.8750	0.3918	0.0008	0.3921
	Oversampling	NB DD	7	8069	5485	1	0.6746	0.8750	0.4047	0.0009	0.4050
	Oversampling	RF Base	0	7	13547	8	0.6970	0.0000	0.9995	0.0000	0.9989
	Oversampling	RF DD	0	4	13550	8	0.6928	0.0000	0.9997	0.0000	0.9991
	Oversampling	XGB Base	0	15	13539	8	0.4994	0.0000	0.9989	0.0000	0.9983
	Oversampling	XGB DD	0	12	13542	8	0.4996	0.0000	0.9991	0.0000	0.9985
$Texas\ ratio_{q+4} > 100\%$	Both over and under	KNN Base	2	266	13288	6	0.7264	0.2500	0.9804	0.0075	0.9799
	Both over and under	KNN DD	2	292	13262	6	0.7853	0.2500	0.9785	0.0068	0.9780
	Both over and under	NB Base	7	8072	5482	1	0.6725	0.8750	0.4045	0.0009	0.4047
	Both over and under	NB DD	7	7920	5634	1	0.6856	0.8750	0.4157	0.0009	0.4159
	Both over and under	RF Base	0	10	13544	8	0.6757	0.0000	0.9993	0.0000	0.9987
	Both over and under	RF DD	0	8	13546	8	0.6881	0.0000	0.9994	0.0000	0.9988
	Both over and under	XGB Base	0	49	13505	8	0.4982	0.0000	0.9964	0.0000	0.9958
	Both over and under	XGB DD	1	65	13489	7	0.5601	0.1250	0.9952	0.0152	0.9947
$Texas\ ratio_{q+4} > 100\%$	No resampling	KNN Base	0	7	13547	8	0.5508	0.0000	0.9995	0.0000	0.9989
	No resampling	KNN DD	0	7	13547	8	0.5503	0.0000	0.9995	0.0000	0.9989
	No resampling	NB Base	7	7580	5974	1	0.6570	0.8750	0.4408	0.0009	0.4410
	No resampling	NB DD	7	7484	6070	1	0.6711	0.8750	0.4478	0.0009	0.4481
	No resampling	RF Base	0	0	13554	8	0.7467	0.0000	1.0000		0.9994
	No resampling	RF DD	0	0	13554	8	0.8514	0.0000	1.0000		0.9994
	No resampling	XGB Base	0	0	13554	8	0.5584	0.1250	0.9917	0.0088	0.9912
	No resampling	XGB DD	0	1	13553	8	0.5000	0.0000	0.9999	0.0000	0.9993

* KNN is k-nearest neighbor, NB is naive bayes, RF is randomforest and XGB is XGBoost.

5.5 Imputation robustness tests

Another possible cause of variation in the results is the formation of new data sets by applying multiple iterations of the MICE algorithm to the original sample with missing data. This was again tested on the models with $Texas\ ratio_{q+4} > 100\%$ as the dependent variable. These results are shown in table 16 in appendix C. The largest difference in true predicted positives was from two true positives in data set 1 to four true positives in data set 3 for the KNN algorithm. These results in table 16 show more variation than the results shown in table 13. This indicates multiple iterations of the SMOTE algorithm on the same sample data gives more robust results than when different MICE imputed datasets are used. For consistency, this research based testing on only one MICE imputed dataset. The ROC graphs for these tests are shown in appendix E.

As discussed earlier, mean/median imputation is another form of dealing with missing data. Therefore, additional robustness tests are performed to compare results between a imputed dataset with the MICE algorithm and a dataset where missing values are replaced with the median per variable. The results of this test are shown in table 17 in appendix C. Table 17 shows the results from using median imputation are similar to MICE imputation for the testing applied in this research. One notable difference is seen when predicting one year ahead with the XGBoost algorithm. Here MICE imputation shows an improvement of just 0.08% in the AUC when adding the DD variable, but median imputation shows an improvement of 12.54% in the AUC when adding the DD variable. Furthermore, table 17 shows that on average the performance metrics of models using MICE imputation do not differ from the models using median imputation. This gives no indication that MICE is a superior form of imputation. However, this adds to the robustness of the results, because the form of imputation does not significantly alter the results. The ROC graphs for models with median imputation are shown in appendix F. To complete this robustness test, the variable importance is calculated for random forest models using median imputation. These results can be seen in table 18 in appendix C. The importance of the DD variable shows a different progression when predicting bank distress over a longer horizon, compared to results using MICE imputation. The variable importance of DD under median imputation shows a continuing increase as the prediction horizon becomes larger.

6. Conclusions

The research question of this paper is as follows: Can the addition of a market indicator to established risk indicators improve predictions of bank distress? To help answer the research question, the following hypothesis was formulated: H1a: Market indicators improve machine learning models based on established risk indicators when predicting bank distress.

The failure to correctly predict the financial crisis of 2008 underscored the need for more robust prediction methods. The CAMELS rating system currently used by financial regulators provides a solid basis for predicting bank distress. The potential improvement of the distance to default indicator to models based on CAMELS indicators has been tested using advanced machine learning algorithms. The findings give an indication that the addition of a market variable improves models when predicting bank distress, so they support the hypothesis. It might be expected that variables containing information not yet covered by other variables would improve the predictive accuracy of the outcome variable. However, it is questionable whether the adding of the DD indicator has economic effect. This is because in most cases, where improvement is realized with the addition of the DD indicator, the improvement is only marginal. The finding that a market indicator becomes more important for prediction distress over a longer horizon, could have implications for models predicting an event at different times in the future. When the event predicted is close to the moment of the prediction, the accounting variables should provide ample information. When the event is further into the future however, it could be more important to include a market indicator as a predictor for bank distress, because of the forward-looking nature of most market indicators. Finally, the type of market variable has important implications for prediction models and perceived improvement of predictions of bank distress. The implied cost of capital was more important for predicting bank distress in the short-term, where the distance to default did not show this for the same prediction horizon. This could implicate different market variables could be used for predicting bank distress over different horizons. Testing the beta as market variable did not significantly alter the results, therefore could provide a simple alternative as added variable for predicting bank distress.

This paper adds to previous literature by testing the additional predictive power of a market indicator for predicting bank distress with advanced machine learning techniques. The result of previous research that a market indicator provides little additional predictive value over existing risk indicators, is now also tested with a research design that is based on more advanced statistical tests. The result of this testing implies that practitioners (regulators, investors) should not view a market indicator as the holy grail of bank distress prediction, just as they should not discard the value of a market indicator altogether. A distance to default market indicator could still have additional value, especially over a longer prediction horizon. Another consideration of the results stems from the market indicators used in the research. These indicators are an approximation or proxy for the assessment made by the market of the underlying risk of a bank. If this approximation could be improved, we would also see a greater effect on the prediction of bank distress. This is shown by the different assumptions underlying the different market indicators that could yield different results. This difference between market indicators should be taken into account when evaluating the added value for predicting bank distress.

This research and the produced results have some limitations that affect the conclusions. The sample taken is only tested on US banks, so the results cannot be generalized for all banks. Further research could be done that includes banks from other regions. This research also includes smaller banks into the sample for which all needed data was available. This could be a limitation as results could differ when taking only one size or specific bank holding companies into account. This is controlled with a logarithm of total assets to account for size, however other characteristic differences between a holding bank and its subsidiaries could impact the findings. One other limitation lies in the data needed to make all calculations. Because this research has a market indicator as the focal point, only publicly traded banks are included in the dataset, which reduces the number of observations by a substantial amount. Further research could also include other machine learning techniques. As this field of research is constantly evolving, new or improved methods are applied. The finding of this research, that the variable importance of the DD indicator became larger as the models predicted events further into the future, could also form the basis for further research. Further research could test when the improvement of importance of a market indicator for predicting bank distress holds and when it fails. Lastly, further research could done to test other market variables than the ones employed in this research.

7. Reference list

Acharya, V., Anginer, D., and Warburton, J. (2016), The End of Market Discipline? Investor Expectations of Implicit Government Guarantees, MPRA Paper, University Library of Munich.

Afonso, G., Santos, J., Traina, J. (2014). Do 'Too-Big-To-Fail' Banks Take on More Risk? Economic Policy Review, 20 (2), p. 41-58.

Auvray, T., and Brossard, O. (2012). Too dispersed to monitor? Ownership dispersion, monitoring, and the prediction of bank distress. Journal of Money, Credit and Banking, 44(4), 685-714.

Barboza, F., Kimura, H., and Altman, E. (2017). Machine learning models and bankruptcy prediction. Expert Systems with Applications, 83, 405-417.

Berger, A., Davies, S., and Flannery, M. (2000). Comparing market and supervisory assessments of bank performance: who knows what when? J. Money Credit Bank. 32, 641–667.

Bertomeu, J., Cheynel, E., Floyd, E., Pan, W. (2020). Using machine learning to detect misstatements. Review of Accounting studies, forthcoming.

Bliss, R. (2001). Market Discipline and Subordinated Debt: A Review of some Salient Issues. Federal Reserve Bank of Chicago Economic Review, first quarter, 24-45.

Calders T., Jaroszewicz S. (2007) Efficient AUC Optimization for Classification. In: Kok J.N., Koronacki J., Lopez de Mantaras R., Matwin S., Mladenič D., Skowron A. (eds) Knowledge Discovery in Databases: PKDD 2007. PKDD 2007. Lecture Notes in Computer Science, vol 4702. Springer, Berlin, Heidelberg.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, p. 321–357.

Claus, J., and Thomas, J. (2001). Equity premia as low as three percent? Evidence from analysts' earnings forecasts for domestic and international stock markets. The Journal of Finance, 56 (5), p. 1629–1666.

Cole, R., White, L., (2012). Déjà vu all over again: the causes of U.S. commercial bank failures this time around. Journal of Financial Services Research, 42, 5–29

Crosbie, P., Bohn, J. (2003). Modelling Default Risk. Moody's KMV.

Curry, T., Fissel, G., Elmer, P. (2003). Using Market Information to Help Identify Distressed Institutions: A Regulatory Perspective. FDIC Banking Review, 15, no. 3.

Curry, T., Elmer, P. and Fissel, G. (2007). Equity Market Data, Bank Failures and Market Efficiency. Journal of Economics and Business, 59, 536-559.

Danenas, P., & Garsva, G. (2015). Selection of support vector machines based classifiers for credit risk domain. Expert Systems with Applications, 42(6), 3194–3204.

Daske, H., Gebhardt, G. & Klein, S. (2006). Estimating the Expected Cost of Equity Capital Using Analysts' Consensus Forecasts. Schmalenbach Bus Rev 58, 2–36.

Dewatripont, M., and Tirole, J. (1993). The Prudential Regulation of Banks. Cambridge MA: MIT Press.

Diamond, D. (1984). Financial Intermediation and Delegated Monitoring. Review of Economic Studies , 51, 393-415.

Distinguin, I., Rous, P. & Tarazi, A. (2006) Market Discipline and the Use of Stock Market Data to Predict Bank Financial Distress. Journal of Financial Services Research , 30, 151-176.

Dumitrescu, E., Hué, S., Hurlin, C., Tokpavi, S. (2021). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, ISSN 0377-2217.

Easton, P. (2004). PE Ratios, PEG Ratios, and Estimating the Implied Expected Rate of Return on Equity Capital. The Accounting Review, 79 (1), p. 73-95.

Efron, B., & Hastie, T. (2016). Computer Age Statistical Inference

Elton, E., Gruber, M., Agrawal, D. and Mann, C. (2001). Explaining the Rate Spread on Corporate Bonds. Journal of Finance 56, 247-277.

Freixas, Xavier, and Jean-Charles Rochet. (1999) Microeconomics of Banking. Cambridge, MA: MIT Press.

Gebhardt, W., Lee, C., Swaminathan, B. (2002). Toward and Implied Cost of Capital. Journal of Accounting Research, 39 (1), p.135-176.

Gode, D. and Mohanram, P. (2003). Inferring the Cost of Capital Using the Ohlson-Juettner Model. Review of Accounting Studies, 8, p. 399–431.

Gordon, M.J., and Shapiro, E. (1956) Capital Equipment Analysis: The Required Rate of Profit. Management Science, 3 (1).

Gropp, Reint, and Jukka Vesala (2004). Deposit Insurance, Moral Hazard and Market Monitoring. Review of Finance 8, 571-602.

Gropp, R., Vesala, J., Vulpes, G., (2006). Equity and bond market signals as leading indicators of bank fragility. J. Money Credit Bank. 38, 399–428.

Gu, Z. and Wu, J. (2003). Earnings skewness and analyst forecast bias. Journal of Accounting and Economics, 35 (1), p. 5-29.

Gunther, J., Levonian, M., Moore, R., (2001). Can the stock markettell bank supervisors anything they don’t already know? Federal Reserve Bank Dallas Econ. Financ. Rev. (Second Quarter), 2–9.

Hanley, J. & McNeil, B. (1982). The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. Radiology. 143, p. 29-36.

Hancock, D. and Kwast, M. (2001). Using Subordinated Debt to Monitor Bank Holding Companies: Is It Feasible? Journal of Financial Services Research, 20, 147-187.

Heo, J. and Yang, J. (2014). AdaBoost based bankruptcy forecasting of Korean construction companies, Applied Soft Computing, Volume 24, Pages 494-499.

Huang, J., Huang, M. (2012). How Much of the Corporate-Treasury Yield Spread Is Due to Credit Risk? Review of Asset Pricing Studies, vol. 2 (2), 153-202.

Jin, J., Kanagaretnam, K., Lobo, G., (2011). Ability of accounting and audit quality variables to predict bank failures during the financial crisis. J. Bank. Finance 35, 2811–2819.

Kerstein, J., Kozberg, A. (2013). Using Accounting Proxies of Proprietary FDIC Ratings to Predict Bank Failures and Enforcement Actions During the Recent Financial Crisis. Journal of Accounting, Auditing and Finance, 28 (2), p. 128-151.

Kim, M., Kang, D., and Kim, H. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. Expert Systems with Applications, Volume 42, Issue 3, Pages 1074-1082.

Levonian, M. (2001). Subordinated Debt and the Quality of Market Discipline in Banking. Federal Reserve Bank of San Fransisco, May.

Lee, C., So, E., Wang, C. (2011). Evaluating Implied Cost of Capital Estimates. SSRN Electronic Journal. 6. 10.2139/ssrn.1653940.

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

Merton, R. (1974). An Analytical Derivation of the Cost of Deposit Insurance and Loan Guarantees. Journal of Banking and Finance, 1, pp. 3-11.

Milne, A., (2014). Distance to default and the financial crisis. Journal of Financial Stability, 12, 26–36.

Miller, S., Olson, E., Yeager, T. (2015). The relative contributions of equity and subordinated debt signals as predictors of bank distress during the financial crisis. Journal of Financial Stability, 16, p. 118-137.

Nier, E., and Baumann, U. (2006). Market Discipline, Disclosure and Moral Hazard in Banking. Journal of Financial Intermediation, 15 (3), 332-361.

Odom, M. D., & Sharda, R. (1990, June). A neural network model for bankruptcy prediction. In 1990 IJCNN International Joint Conference on neural networks (pp. 163-168). IEEE.

Ohlson, J., Jeuttner-Nauroth, B. (2005). Expected EPS and EPS Growth as Determinants of Value. Review of Accounting Studies, 10, p. 349–365.

Park, S., and Peristiani, S. (2007) Are Bank Shareholders Enemies of Regulators or a Potential Source of Market Discipline? Journal of Banking and Finance, 31 (8), 2493-2515.

Saunders, A. (2001). Comments on Evanoff and Wall / Hancock and Kwast. Journal of Fiancial Services Research, 20, 189-194.

Schaeck, K., Zhou, T. and Molyneux, P. 2010. 'Too-Big-To-Fail' and its Impact on Safety Net Subsidies and Systemic Risk. CAREFIN Research Paper No. 09.

Shah, A., Singh, M. and Aggarwal, N. (2013). Distance to default: Implementation in R. Finance routines developed by the IGIDR Finance Research Group.

Tam, K., (1991). Neural Network Models and the Prediction of Bank Bankruptcy. Omega, 19, 429, 445.

Tam, K., and Kiang, M., (1992). Managerial Applications of the Neural Networks: The Case of Bank Failure Predictions. Management Science, 38, 416-430.

Thomson, J. B. (1991). Predicting bank failures in the 1980s. Federal Reserve Bank of Cleveland Economic Review, 27(1), 9-20.

Van Buuren, S., Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45 (3).

8. Appendix A

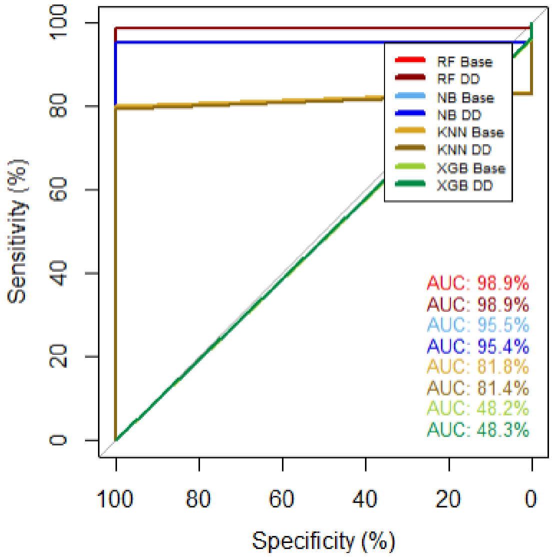

Figure 4: ROC curves for models with Texas ratio over one hundred as distress event in q+8. The corresponding area under curve (AUC) is also printed per model.

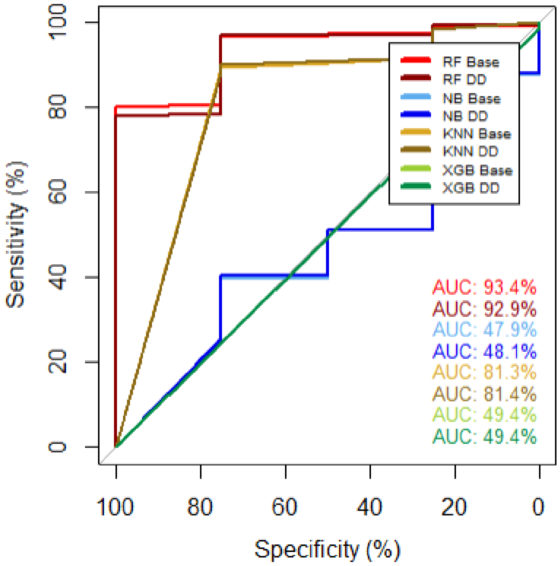

Figure 5: ROC curves for models with a bank failure as distress event in q+4. The corresponding area under curve (AUC) is also printed per model.

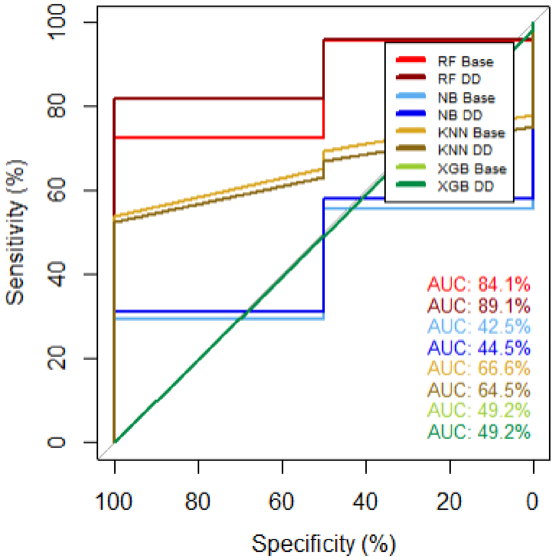

Figure 6: ROC curves for models with a bank failure as distress event in q+8. The corresponding area under curve (AUC) is also printed per model.

9. Appendix B

Table 14
Results of machine learning predictions shown by distress event and model. For all these models the Fitch support rating is added as independent variable.

Distress event	**Model***	**TP**	**FP**	**TN**	**FN**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
$Bank\ failure_{q+1}$	KNN Base	1	72	14039	6	0.6370	0.1429	0.9949	0.0137	0.9945
	KNN DD	1	42	14069	6	0.7083	0.1429	0.9970	0.0233	0.9966
	NB Base	0	13	14098	7	0.9415	0.0000	0.9991	0.0000	0.9986
	NB DD	0	13	14098	7	0.8876	0.0000	0.9991	0.0000	0.9986
	RF Base	0	9	14102	7	0.9462	0.0000	0.9994	0.0000	0.9989
	RF DD	0	7	14104	7	0.9626	0.0000	0.9995	0.0000	0.9990
	XGB Base	0	48	14063	7	0.4983	0.0000	0.9966	0.0000	0.9961
	XGB DD	0	48	14063	7	0.4983	0.0000	0.9966	0.0000	0.9961
$Bank\ failure_{q+4}$	KNN Base	1	393	13721	3	0.8140	0.2500	0.9722	0.0025	0.9720
	KNN DD	1	254	13860	3	0.8184	0.2500	0.9820	0.0039	0.9818
	NB Base	0	23	14091	4	0.4850	0.0000	0.9984	0.0000	0.9981
	NB DD	0	41	14073	4	0.4886	0.0000	0.9971	0.0000	0.9968
	RF Base	0	32	14082	4	0.8675	0.0000	0.9977	0.0000	0.9975
	RF DD	0	34	14080	4	0.8835	0.0000	0.9976	0.0000	0.9973
	XGB Base	0	131	13983	4	0.4954	0.0000	0.9907	0.0000	0.9904
	XGB DD	0	131	13983	4	0.4954	0.0000	0.9907	0.0000	0.9904
$Bank\ failure_{q+8}$	KNN Base	1	1008	13108	1	0.6420	0.5000	0.9286	0.0010	0.9285
	KNN DD	0	372	13744	2	0.6274	0.0000	0.9736	0.0000	0.9735
	NB Base	0	505	13611	2	0.4048	0.0000	0.9642	0.0000	0.9641
	NB DD	0	569	13547	2	0.4230	0.0000	0.9597	0.0000	0.9596
	RF Base	0	97	14019	2	0.7823	0.0000	0.9931	0.0000	0.9930
	RF DD	0	100	14016	2	0.7999	0.0000	0.9929	0.0000	0.9928
	XGB Base	0	332	13784	2	0.4882	0.0000	0.9765	0.0000	0.9763
	XGB DD	0	332	13784	2	0.4882	0.0000	0.9765	0.0000	0.9763

* KNN is k-nearest neighbor, NB is naive bayes, RF is randomforest and XGB is XGBoost.

Table 15

Variable importance for randomforest models, calculated with the Mean Decrease in Gini index, with Fitch support rating as added independent variable. Per variable and model the mean decrease in gini index is shown. The higher numbers mean a larger decrease, indicating these independent variables are more important in predicting the dependent variable for that model. The ranks are shown to indicate how important the different variables are in relation to the other variables.

	$Bank\ failure_{q+1}$		$Bank\ failure_{q+4}$		$Bank\ failure_{q+8}$		
Variable	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Sum of ranks
DD	0.8872	12	0.4481	14	0.6026	11	37
C1	24.0336	2	0.4945	13	0.2649	16	31
C2	9.8250	4	0.0437	21	0.3865	15	40
C3	0.1901	17	0.0187	22	7.3820	2	41
A1	1.2293	11	6.5777	1	2.3665	6	18
A2	9.7167	5	4.4144	4	1.8495	10	19
A3	0.3152	15	2.4136	7	0.1516	19	41
A4	2.5825	10	3.1751	6	0.0837	21	37
M1	2.9659	9	0.1145	18	0.1996	18	45
E1	31.6814	1	3.3110	5	0.5900	12	18
E2	9.6174	6	1.3515	9	7.6056	1	16
E3	3.5645	8	0.2796	17	0.5882	13	38
E4	5.2277	7	1.0575	11	3.4352	3	21
E5	20.6945	3	5.3742	2	0.5158	14	19
E6	0.1289	19	1.1122	10	2.9923	4	33
L1	0.3205	14	0.3509	15	1.9339	8	37
L2	0.0000	22	0.3095	16	0.0333	22	60
L3	0.2748	16	0.0810	20	0.2001	17	53
S1	0.1136	20	0.5090	12	2.1362	7	39
X1	0.0734	21	1.7552	8	2.6433	5	34
X2	0.3777	13	0.0970	19	0.1014	20	52
X3	0.1512	18	4.6818	3	1.8789	9	30
Support rating	0.0000	23	0.0000	23	0.0000	23	69

10. Appendix C

Table 16

Results of machine learning predictions shown for Texas ratio distress event in q+4 per iteration of the MICE imputed dataset. The MICE imputation algorithm is applied five different times.

Distress event	MICE imputation	**Model***	**TP**	**FP**	**TN**	**FN**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
$Texas\ ratio_{q+4} > 100\%$	1	KNN Base	3	796	12758	5	0.6864	0.3750	0.9413	0.0038	0.9409
	1	KNN DD	3	800	12754	5	0.7457	0.3750	0.9410	0.0037	0.9406
	1	NB Base	5	5891	7663	3	0.6223	0.6250	0.5654	0.0008	0.5654
	1	NB DD	5	5831	7723	3	0.6343	0.6250	0.5698	0.0009	0.5698
	1	RF Base	0	49	13505	8	0.7729	0.0000	0.9964	0.0000	0.9958
	1	RF DD	0	47	13507	8	0.8051	0.0000	0.9965	0.0000	0.9959
	1	XGB Base	1	134	13420	7	0.5576	0.1250	0.9901	0.0074	0.9896
	1	XGB DD	1	110	13444	7	0.5584	0.1250	0.9919	0.0090	0.9914
$Texas\ ratio_{q+4} > 100\%$	2	KNN Base	5	947	12607	3	0.8132	0.6250	0.9301	0.0053	0.9300
	2	KNN DD	5	945	12609	3	0.8184	0.6250	0.9303	0.0053	0.9301
	2	NB Base	6	5941	7613	2	0.7054	0.7500	0.5617	0.0010	0.5618
	2	NB DD	6	5912	7642	2	0.7114	0.7500	0.5638	0.0010	0.5639
	2	RF Base	0	46	13508	8	0.8598	0.0000	0.9966	0.0000	0.9960
	2	RF DD	0	42	13512	8	0.8619	0.0000	0.9969	0.0000	0.9963
	2	XGB Base	2	138	13416	6	0.6199	0.2500	0.9898	0.0143	0.9894
	2	XGB DD	2	125	13429	6	0.6204	0.2500	0.9908	0.0157	0.9903
$Texas\ ratio_{q+4} > 100\%$	3	KNN Base	2	1210	12344	6	0.7024	0.2500	0.9107	0.0017	0.9103
	3	KNN DD	2	1264	12290	6	0.7006	0.2500	0.9067	0.0016	0.9064
	3	NB Base	5	5427	8127	3	0.6436	0.6250	0.5996	0.0009	0.5996
	3	NB DD	5	5336	8218	3	0.6554	0.6250	0.6063	0.0009	0.6063
	3	RF Base	0	37	13517	8	0.8481	0.0000	0.9973	0.0000	0.9967
	3	RF DD	0	38	13516	8	0.8473	0.0000	0.9972	0.0000	0.9966
	3	XGB Base	1	134	13420	7	0.5576	0.1250	0.9901	0.0074	0.9896
	3	XGB DD	0	139	13415	8	0.4949	0.0000	0.9897	0.0000	0.9892
$Texas\ ratio_{q+4} > 100\%$	4	KNN Base	4	799	12755	4	0.8159	0.5000	0.9411	0.0050	0.9408
	4	KNN DD	5	876	12678	3	0.8270	0.6250	0.9354	0.0057	0.9352
	4	NB Base	5	6850	6704	3	0.6026	0.6250	0.4946	0.0007	0.4947
	4	NB DD	5	6781	6773	3	0.6147	0.6250	0.4997	0.0007	0.4998
	4	RF Base	1	33	13521	7	0.7708	0.1250	0.9976	0.0294	0.9971
	4	RF DD	1	32	13522	7	0.7864	0.1250	0.9976	0.0303	0.9971
	4	XGB Base	1	90	13464	7	0.5592	0.1250	0.9934	0.0110	0.9928
	4	XGB DD	1	105	13449	7	0.5586	0.1250	0.9923	0.0094	0.9917
$Texas\ ratio_{q+4} > 100\%$	5	KNN Base	4	955	12599	4	0.7301	0.5000	0.9295	0.0042	0.9293
	5	KNN DD	4	957	12597	4	0.7281	0.5000	0.9294	0.0042	0.9291
	5	NB Base	4	6189	7365	4	0.5696	0.5000	0.5434	0.0006	0.5434
	5	NB DD	4	6052	7502	4	0.5813	0.5000	0.5535	0.0007	0.5535
	5	RF Base	0	55	13499	8	0.8723	0.0000	0.9959	0.0000	0.9954
	5	RF DD	0	48	13506	8	0.8797	0.0000	0.9965	0.0000	0.9959
	5	XGB Base	2	122	13432	6	0.6205	0.2500	0.9910	0.0161	0.9906
	5	XGB DD	1	146	13408	7	0.5571	0.1250	0.9892	0.0068	0.9887

* KNN is k-nearest neighbor, NB is naive bayes, RF is randomforest and XGB is XGBoost.

Table 17

Results of machine learning predictions shown for Texas ratio distress event per imputed dataset. For median imputation, missing values are replaced with the median of that variable. The results for Texas ratio distress events from table 5 and 6 are included for comparison.

Distress event	Imputation	**Model***	**TP**	**FP**	**TN**	**FN**	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**Accuracy**
$Texas\ ratio_{q+1} > 100\%$	MICE	KNN Base	11	925	12619	7	0.8441	0.6111	0.9317	0.0118	0.9313
	MICE	KNN DD	11	967	12577	7	0.8425	0.6111	0.9286	0.0112	0.9282
	MICE	NB Base	14	6335	7209	4	0.8091	0.7778	0.5323	0.0022	0.5326
	MICE	NB DD	14	6332	7212	4	0.8083	0.7778	0.5325	0.0022	0.5328
	MICE	RF Base	4	60	13484	14	0.9503	0.2222	0.9956	0.0625	0.9945
	MICE	RF DD	5	61	13483	13	0.9591	0.2778	0.9955	0.0758	0.9945
	MICE	XGB Base	9	275	13269	9	0.7398	0.5000	0.9797	0.0317	0.9791
	MICE	XGB DD	7	221	13323	11	0.6863	0.3889	0.9837	0.0307	0.9829
$Texas\ ratio_{q+4} > 100\%$	MICE	KNN Base	3	796	12758	5	0.6864	0.3750	0.9413	0.0038	0.9409
	MICE	KNN DD	3	800	12754	5	0.7457	0.3750	0.9410	0.0037	0.9406
	MICE	NB Base	5	5891	7663	3	0.6223	0.6250	0.5654	0.0008	0.5654
	MICE	NB DD	5	5831	7723	3	0.6343	0.6250	0.5698	0.0009	0.5698
	MICE	RF Base	0	49	13505	8	0.7729	0.0000	0.9964	0.0000	0.9958
	MICE	RF DD	0	47	13507	8	0.8051	0.0000	0.9965	0.0000	0.9959
	MICE	XGB Base	1	134	13420	7	0.5576	0.1250	0.9901	0.0074	0.9896
	MICE	XGB DD	1	110	13444	7	0.5584	0.1250	0.9919	0.0090	0.9914
$Texas\ ratio_{q+8} > 100\%$	MICE	KNN Base	0	966	12595	1	0.8185	0.0000	0.9288	0.0000	0.9287
	MICE	KNN DD	0	989	12572	1	0.8139	0.0000	0.9271	0.0000	0.9270
	MICE	NB Base	0	5962	7599	1	0.0452	0.0000	0.5604	0.0000	0.5603
	MICE	NB DD	0	5897	7664	1	0.0464	0.0000	0.5652	0.0000	0.5651
	MICE	RF Base	0	57	13504	1	0.9889	0.0000	0.9958	0.0000	0.9957
	MICE	RF DD	0	60	13501	1	0.9890	0.0000	0.9956	0.0000	0.9955
	MICE	XGB Base	0	492	13069	1	0.4819	0.0000	0.9637	0.0000	0.9636
	MICE	XGB DD	0	452	13109	1	0.4833	0.0000	0.9667	0.0000	0.9666
$Texas\ ratio_{q+1} > 100\%$	Median	KNN Base	9	797	12747	9	0.8602	0.5000	0.9412	0.0112	0.9406
	Median	KNN DD	11	807	12737	7	0.8668	0.6111	0.9404	0.0134	0.9400
	Median	NB Base	14	7567	5977	4	0.6949	0.7778	0.4413	0.0018	0.4417
	Median	NB DD	14	7501	6043	4	0.6639	0.7778	0.4462	0.0019	0.4466
	Median	RF Base	3	81	13463	15	0.9340	0.1667	0.9940	0.0357	0.9929
	Median	RF DD	2	73	13471	16	0.9351	0.1111	0.9946	0.0267	0.9934
	Median	XGB Base	3	209	13335	15	0.5756	0.1667	0.9846	0.0142	0.9835
	Median	XGB DD	2	239	13305	16	0.5467	0.1111	0.9824	0.0083	0.9812
$Texas\ ratio_{q+4} > 100\%$	Median	KNN Base	4	841	12713	4	0.8045	0.5000	0.9380	0.0047	0.9377
	Median	KNN DD	4	885	12669	4	0.8009	0.5000	0.9347	0.0045	0.9344
	Median	NB Base	6	7445	6109	2	0.6831	0.7500	0.4507	0.0008	0.4509
	Median	NB DD	6	7307	6247	2	0.6906	0.7500	0.4609	0.0008	0.4611
	Median	RF Base	0	44	13510	8	0.9083	0.0000	0.9968	0.0000	0.9962
	Median	RF DD	0	48	13506	8	0.9197	0.0000	0.9965	0.0000	0.9959
	Median	XGB Base	0	119	13435	8	0.4956	0.0000	0.9912	0.0000	0.9906
	Median	XGB DD	2	108	13446	6	0.6210	0.2500	0.9920	0.0182	0.9916
$Texas\ ratio_{q+8} > 100\%$	Median	KNN Base	0	753	12808	1	0.4219	0.0000	0.9445	0.0000	0.9444
	Median	KNN DD	0	766	12795	1	0.4214	0.0000	0.9435	0.0000	0.9434
	Median	NB Base	0	5154	8407	1	0.0322	0.0000	0.6199	0.0000	0.6199
	Median	NB DD	0	5273	8288	1	0.0348	0.0000	0.6112	0.0000	0.6111
	Median	RF Base	0	98	13463	1	0.9883	0.0000	0.9928	0.0000	0.9927
	Median	RF DD	0	94	13467	1	0.9894	0.0000	0.9931	0.0000	0.9930
	Median	XGB Base	1	186	13375	0	0.9931	1.0000	0.9863	0.0053	0.9863
	Median	XGB DD	1	213	13348	0	0.9921	1.0000	0.9843	0.0047	0.9843

* KNN is k-nearest neighbor, NB is naive bayes, RF is randomforest and XGB is XGBoost.

Table 18

Variable importance for randomforest models, calculated with the Mean Decrease in Gini index, based on a median imputed dataset. The models indicated Texas ratio distress event. Per variable and model the mean decrease in gini index is shown. The higher numbers mean a larger decrease, indicating these independent variables are more important in predicting the dependent variable for that model. The ranks are shown to indicate how important the different variables are in relation to the other variables.

	$Texas\,ratio_{q+1} > 100\%$		$Texas\,ratio_{q+4} > 100\%$		$Texas\,ratio_{q+8} > 100\%$		
Variable	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Sum of ranks
DD	28.4241	16	40.1614	9	19.1390	6	31
C1	70.7551	8	30.2402	10	12.3105	8	26
C2	117.3476	4	19.5191	14	7.4090	12	30
C3	33.7410	11	17.0501	17	3.6049	21	49
A1	74.8924	7	100.8784	1	30.4298	3	11
A2	32.7948	13	24.0696	13	6.5117	13	39
A3	60.7847	9	19.0930	15	6.3580	14	38
A4	163.6012	3	86.3650	2	19.7818	4	9
M1	33.4781	12	13.7480	19	11.2545	9	40
E1	210.4071	1	50.8719	6	3.8147	20	27
E2	98.8395	5	41.5997	8	9.0633	11	24
E3	18.4413	19	48.0235	7	3.9198	19	45
E4	83.4001	6	16.2473	18	4.0186	17	41
E5	193.8032	2	29.9728	11	12.4976	7	20
E6	32.5769	14	53.6482	4	19.3286	5	23
L1	18.5544	18	27.2462	12	4.0063	18	48
L2	0.8251	22	0.7299	22	6.0509	15	59
L3	38.9121	10	52.8863	5	34.6365	2	17
S1	5.3567	21	3.6471	21	3.2425	22	64
X1	19.2445	17	60.3052	3	9.2542	10	30
X2	5.6728	20	7.5115	20	4.1161	16	56
X3	29.7472	15	17.9442	16	43.1485	1	32

11. Appendix D

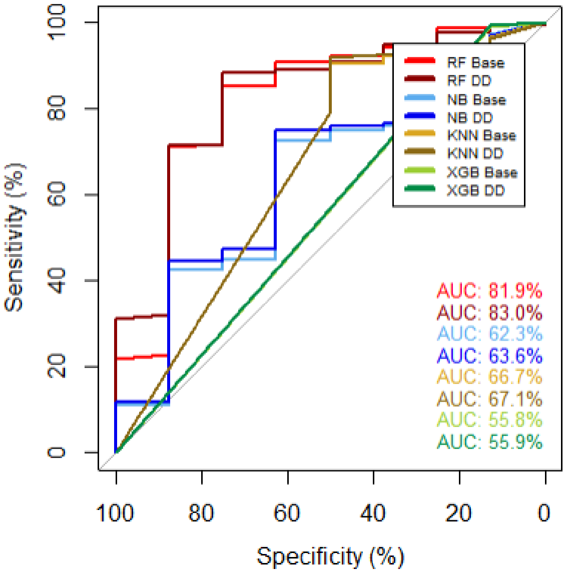

Figure 7: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for a different SMOTE training set than the one shown in figure 4. The corresponding area under curve (AUC) is also printed per model.

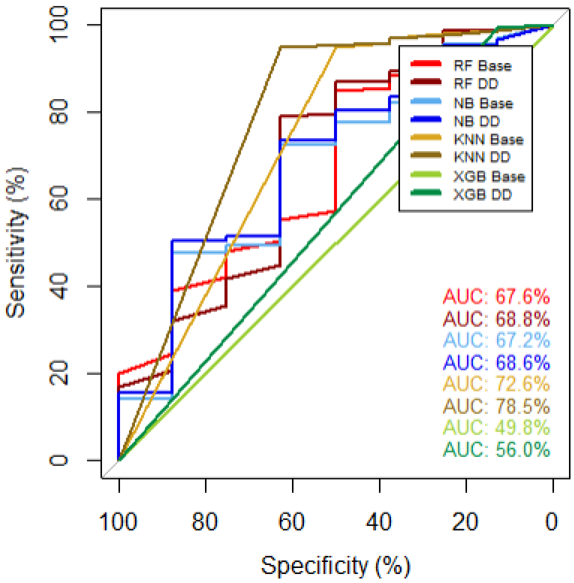

Figure 8: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for both over- and undersampled training set. The corresponding area under curve (AUC) is also printed per model.

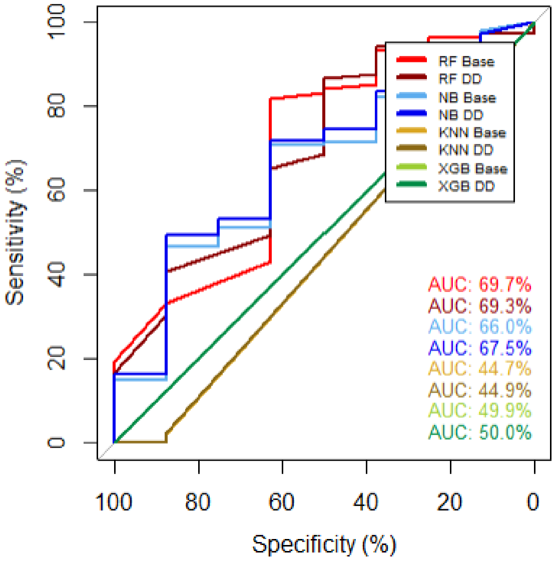

Figure 9: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for an oversampled training set. The corresponding area under curve (AUC) is also printed per model.

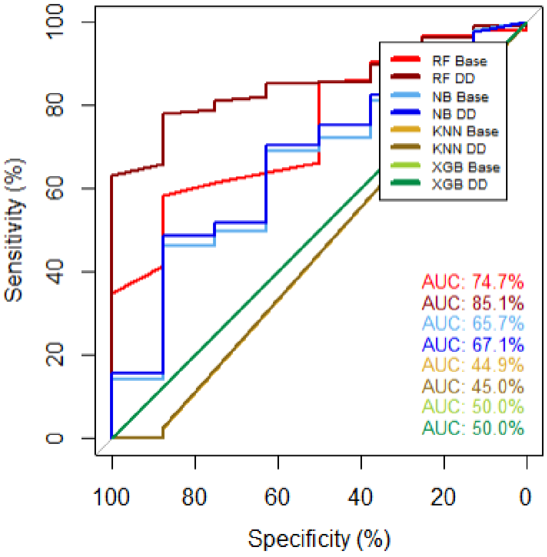

Figure 10: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for a training set that is not resampled. The corresponding area under curve (AUC) is also printed per model.

12. Appendix E

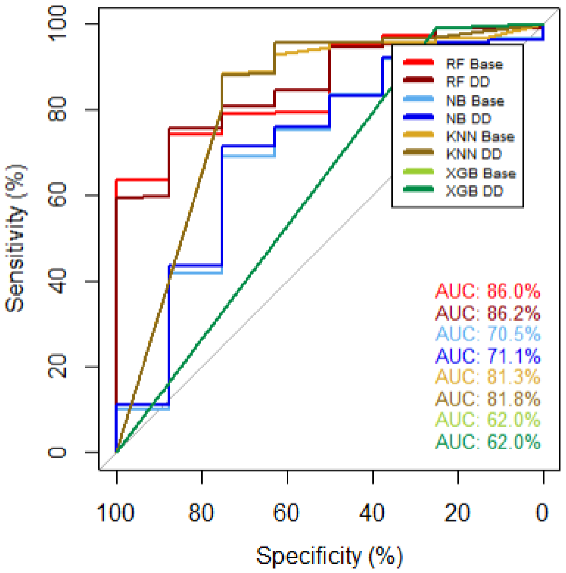

Figure 11: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for MICE imputed dataset 2. The corresponding area under curve (AUC) is also printed per model.

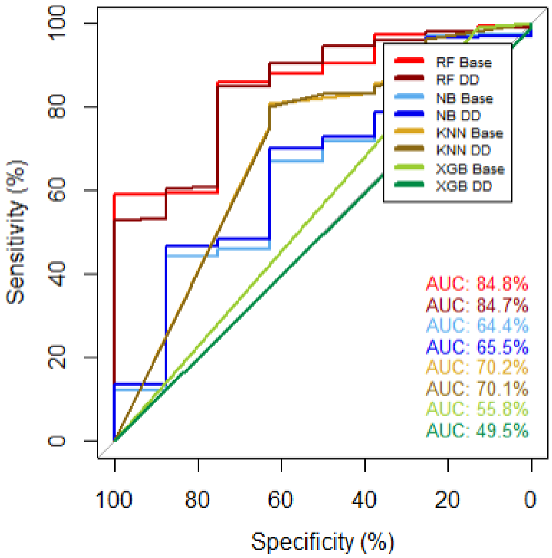

Figure 12: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for MICE imputed dataset 3. The corresponding area under curve (AUC) is also printed per model.

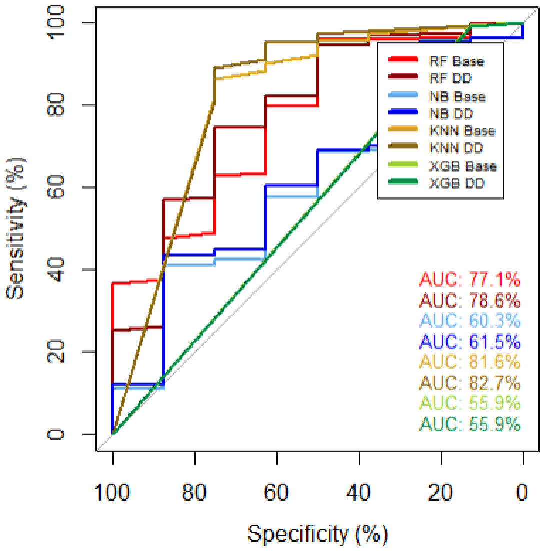

Figure 13: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for MICE imputed dataset 4. The corresponding area under curve (AUC) is also printed per model.

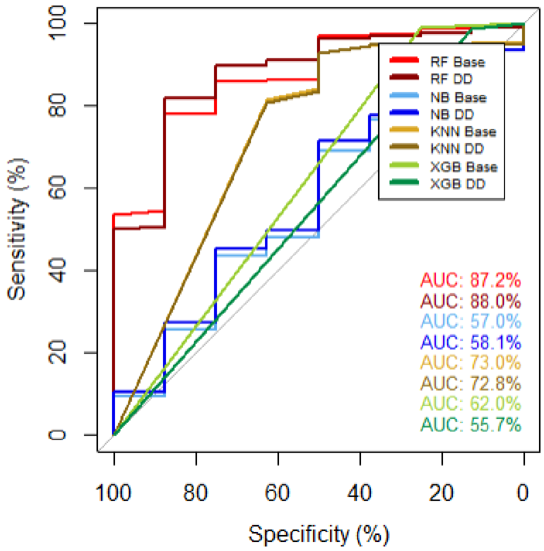

Figure 14: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for MICE imputed dataset 5. The corresponding area under curve (AUC) is also printed per model.

13. Appendix F

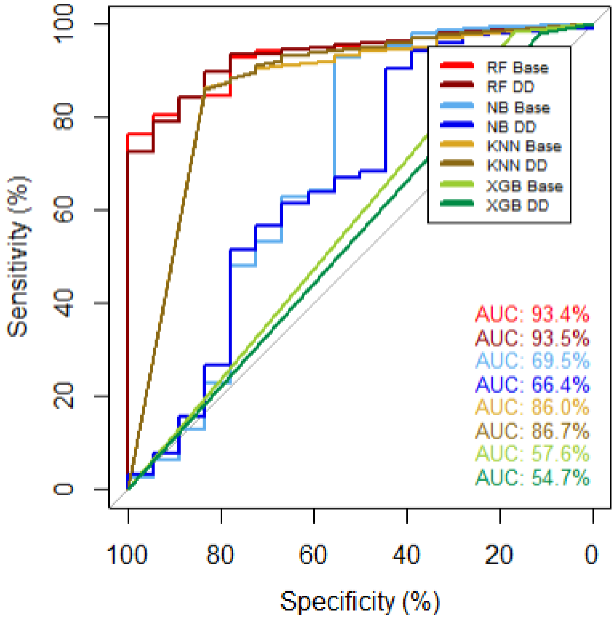

Figure 15: ROC curves for models with Texas ratio over one hundred as distress event in q+1, for a dataset with median imputation. The corresponding area under curve (AUC) is also printed per model.

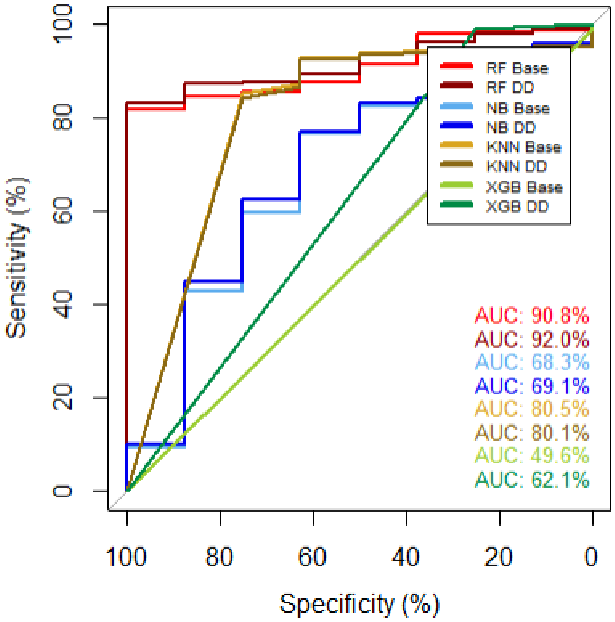

Figure 16: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for a dataset with median imputation. The corresponding area under curve (AUC) is also printed per model.

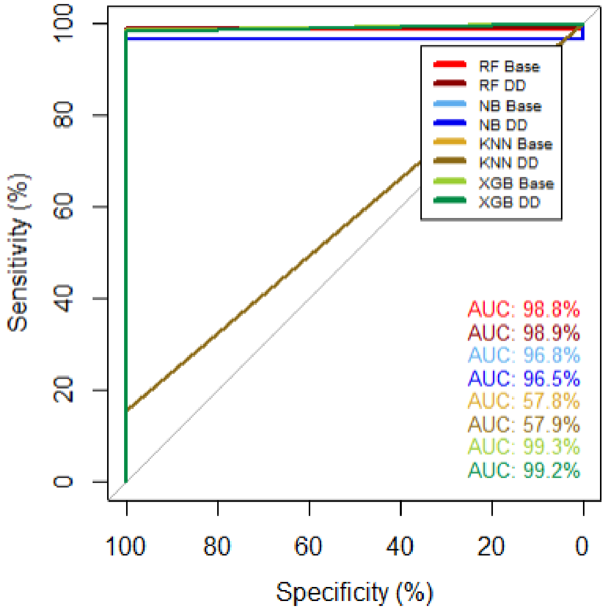

Figure 17: ROC curves for models with Texas ratio over one hundred as distress event in q+8, for a dataset with median imputation. The corresponding area under curve (AUC) is also printed per model.

14. Appendix G

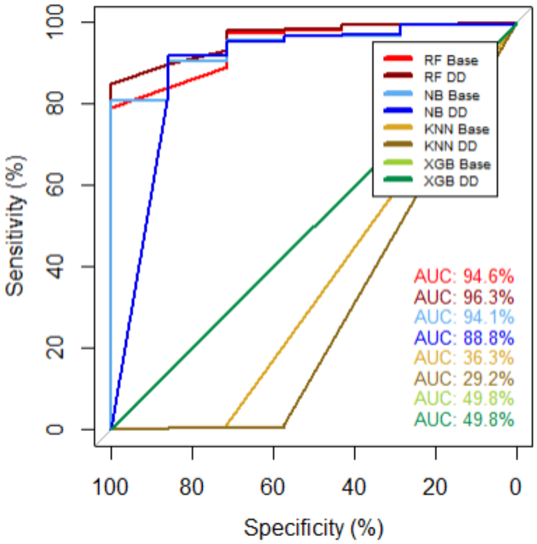

Figure 18: ROC curves for models with bank failure as distress event in q+1, with support rating as added independent variable. The corresponding area under curve (AUC) is also printed per model.

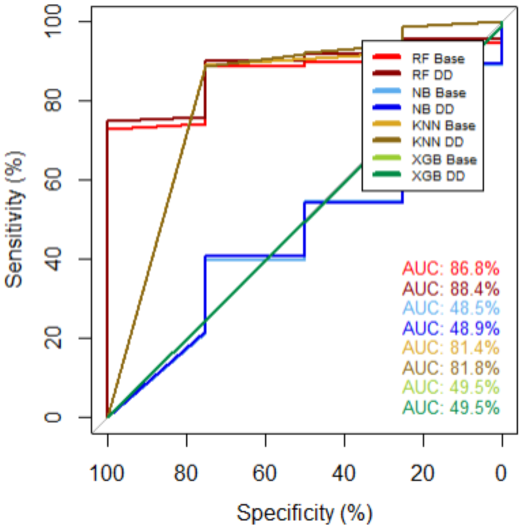

Figure 19: ROC curves for models with bank failure as distress event in q+4, with support rating as added independent variable. The corresponding area under curve (AUC) is also printed per model.

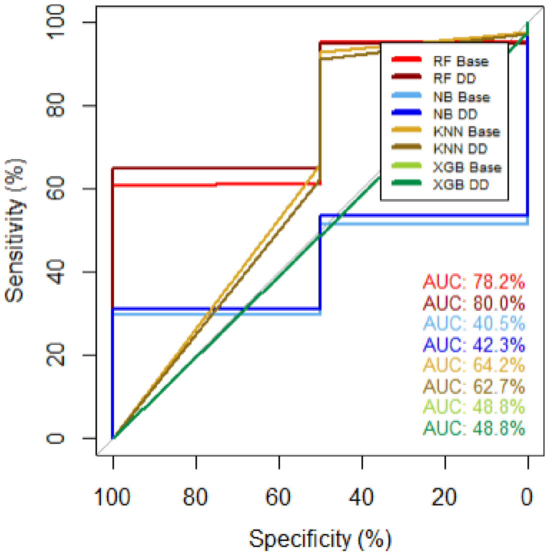

Figure 20: ROC curves for models with bank failure as distress event in q+8, with support rating as added independent variable. The corresponding area under curve (AUC) is also printed per model.

15. Appendix H

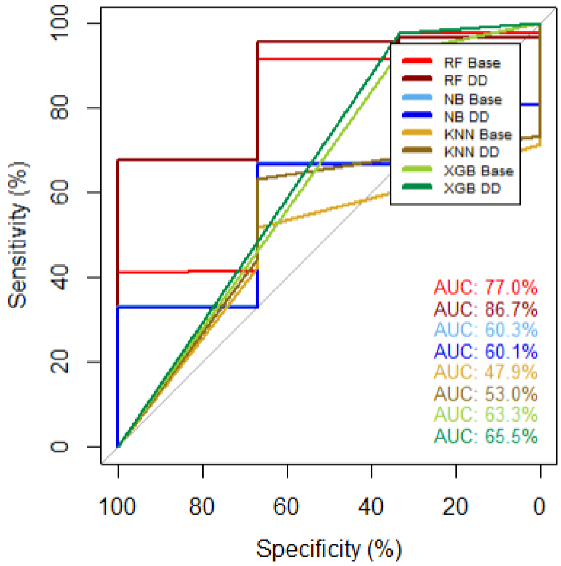

Figure 21: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for average implied cost of capital as added market indicator. The corresponding area under curve (AUC) is also printed per model.

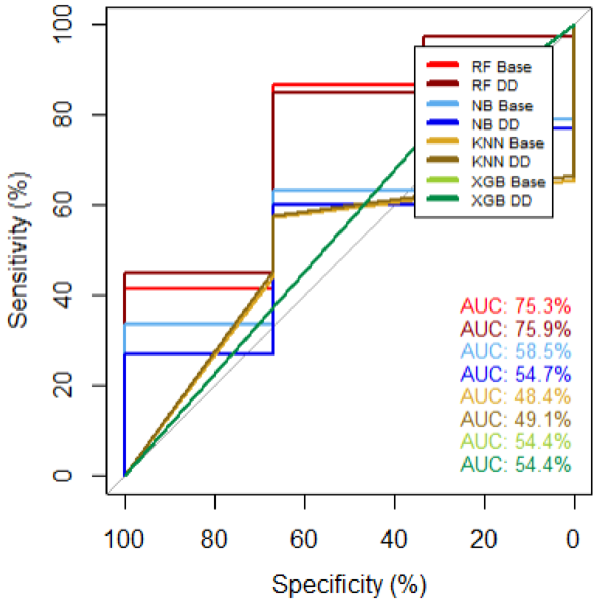

Figure 22: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for Gordon growth model implied cost of capital as added market indicator. The corresponding area under curve (AUC) is also printed per model.

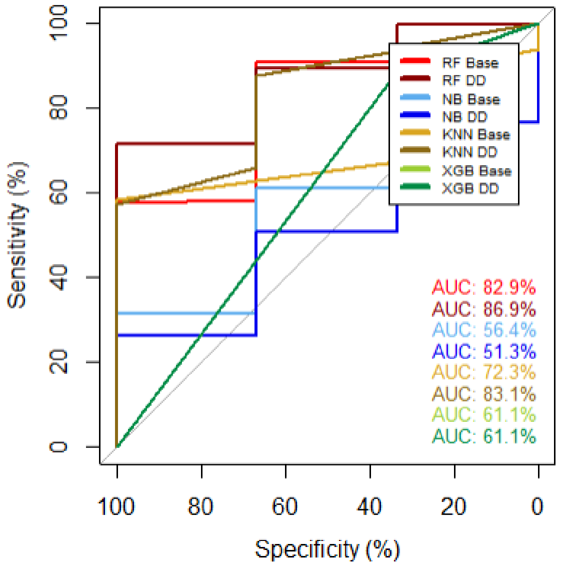

Figure 23: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for price earnings growth model implied cost of capital as added market indicator. The corresponding area under curve (AUC) is also printed per model.

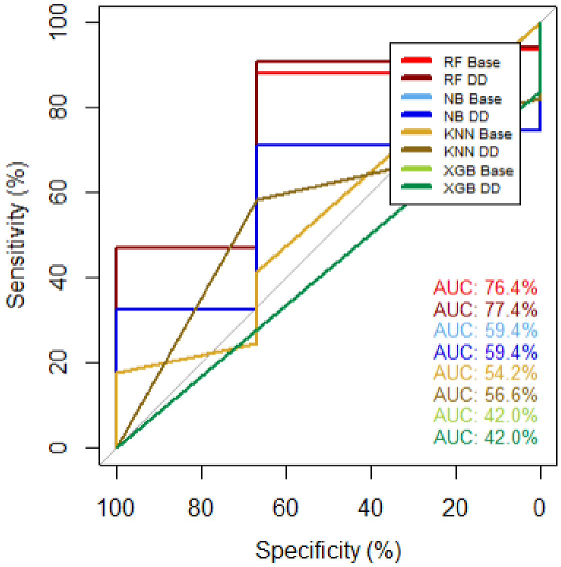

Figure 24: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for Ohlson-Juettner-Nauroth model implied cost of capital as added market indicator. The corresponding area under curve (AUC) is also printed per model.

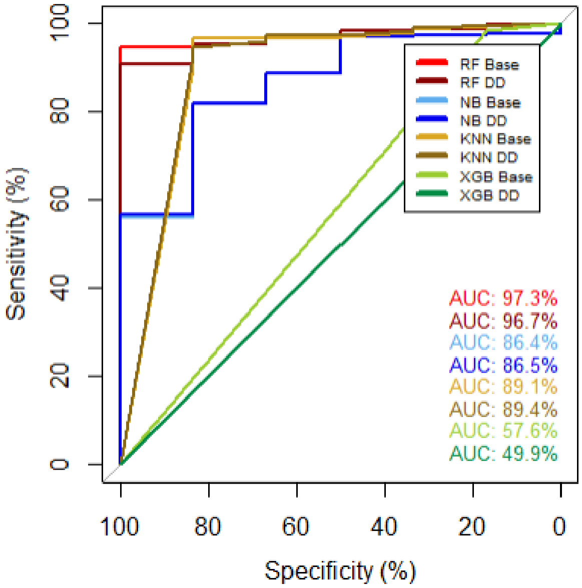

Figure 25: ROC curves for models with Texas ratio over one hundred as distress event in q+1, for average implied cost of capital as added market indicator. The corresponding area under curve (AUC) is also printed per model.

16. Appendix I

Table 19

Results of machine learning predictions shown for Texas ratio distress event in q+1 per method of ICC calculation.

Distress event	ICC*	Model**	TP	FP	TN	FN	AUC	Sensitivity	Specificity	Precision	Accuracy
$Texas\ ratio_{q+1} > 100\%$	Average	KNN Base	3	142	4579	3	0.8910	0.5000	0.9699	0.0207	0.9693
	Average	KNN ICC	3	120	4601	3	0.8935	0.5000	0.9746	0.0244	0.9740
	Average	NB Base	3	482	4239	3	0.8644	0.5000	0.8979	0.0062	0.8974
	Average	NB ICC	3	424	4297	3	0.8651	0.5000	0.9102	0.0070	0.9097
	Average	RF Base	1	8	4713	5	0.9729	0.1667	0.9983	0.1111	0.9972
	Average	RF ICC	0	5	4716	6	0.9673	0.0000	0.9989	0.0000	0.9977
	Average	XGB Base	1	68	4653	5	0.5761	0.1667	0.9856	0.0145	0.9846
	Average	XGB ICC	0	7	4714	6	0.4993	0.0000	0.9985	0.0000	0.9972
$Texas\ ratio_{q+1} > 100\%$	GGM	KNN Base	2	135	4586	4	0.8091	0.3333	0.9714	0.0146	0.9706
	GGM	KNN ICC	3	90	4631	3	0.8173	0.5000	0.9809	0.0323	0.9803
	GGM	NB Base	4	608	4113	2	0.8889	0.6667	0.8712	0.0065	0.8710
	GGM	NB ICC	1	51	4670	5	0.8562	0.1667	0.9892	0.0192	0.9882
	GGM	RF Base	1	4	4717	5	0.9685	0.1667	0.9992	0.2000	0.9981
	GGM	RF ICC	1	3	4718	5	0.9698	0.1667	0.9994	0.2500	0.9983
	GGM	XGB Base	2	23	4698	4	0.6642	0.3333	0.9951	0.0800	0.9943
	GGM	XGB ICC	3	24	4697	3	0.7475	0.5000	0.9949	0.1111	0.9943
$Texas\ ratio_{q+1} > 100\%$	OJM	KNN Base	4	135	4586	2	0.8954	0.6667	0.9714	0.0288	0.9710
	OJM	KNN ICC	5	145	4576	1	0.8941	0.8333	0.9693	0.0333	0.9691
	OJM	NB Base	3	515	4206	3	0.8668	0.5000	0.8909	0.0058	0.8904
	OJM	NB ICC	3	536	4185	3	0.8668	0.5000	0.8865	0.0056	0.8860
	OJM	RF Base	2	12	4709	4	0.9775	0.3333	0.9975	0.1429	0.9966
	OJM	RF ICC	2	10	4711	4	0.9782	0.3333	0.9979	0.1667	0.9970
	OJM	XGB Base	2	30	4691	4	0.6635	0.3333	0.9936	0.0625	0.9928
	OJM	XGB ICC	2	51	4670	4	0.6613	0.3333	0.9892	0.0377	0.9884
$Texas\ ratio_{q+1} > 100\%$	PEG	KNN Base	3	124	4597	3	0.8918	0.5000	0.9737	0.0236	0.9731
	PEG	KNN ICC	4	105	4616	2	0.8983	0.6667	0.9778	0.0367	0.9774
	PEG	NB Base	4	587	4134	2	0.8794	0.6667	0.8757	0.0068	0.8754
	PEG	NB ICC	4	533	4188	2	0.8839	0.6667	0.8871	0.0074	0.8868
	PEG	RF Base	1	11	4710	5	0.9616	0.1667	0.9977	0.0833	0.9966
	PEG	RF ICC	1	10	4711	5	0.9720	0.1667	0.9979	0.0909	0.9968
	PEG	XGB Base	3	25	4696	3	0.7474	0.5000	0.9947	0.1071	0.9941
	PEG	XGB ICC	3	25	4696	3	0.7474	0.5000	0.9947	0.1071	0.9941

* Implied Cost of Capital (ICC) calculation models; Average is the average of positive ICC values per bank quarter observation (5005 total observations), OJM is the Ohlson-Jeuttner-Nauroth Model (4872 total observations), PEG is the Price Earnings Growth Model (4391 total observations), and GGM is the Gordon Growth Model (4436 total observations).

** KNN is k-nearest neighbor, NB is naive bayes, RF is randomforest and XGB is XGBoost.

Table 20

Variable importance for randomforest models, calculated with the Mean Decrease in Gini index, for different implied cost of capital models as market variable. Per variable and model the mean decrease in gini index is shown. The higher numbers mean a larger decrease, indicating these independent variables are more important in predicting the dependent variable for that model. The ranks are shown to indicate how important the different variables are in relation to the other variables.

	$Texas\,ratio_{q+1}$		$Texas\,ratio_{q+1}$		$Texas\,ratio_{q+1}$		$Texas\,ratio_{q+1}$		
	$> 100\%$		$> 100\%$		$> 100\%$		$> 100\%$		
	Average ICC		**GGM***		**OJM***		**PEG***		
Variable	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Sum of ranks
ICC	43.1891	1	6.1961	11	3.8709	9	16.0259	5	26
C1	8.4252	7	6.5394	10	2.8634	14	3.0669	15	46
C2	2.3486	14	2.0497	17	1.5462	19	2.5725	18	68
C3	0.7819	22	0.4648	22	0.4870	21	0.6213	22	87
A1	7.4910	9	11.3189	7	12.5998	6	6.7995	8	30
A2	2.1092	17	1.9624	18	2.5129	15	1.7291	19	69
A3	3.1248	13	2.3217	16	3.3825	12	3.9197	11	52
A4	14.1411	6	20.8598	4	15.6631	5	13.4311	6	21
M1	2.2120	16	2.9088	13	1.9989	17	3.9174	12	58
E1	17.8156	5	14.5666	5	17.2414	4	31.5666	3	17
E2	4.5810	10	2.6677	14	2.4589	16	3.2225	14	54
E3	2.0792	19	1.2625	20	1.8124	18	2.6173	17	74
E4	0.8022	21	0.6681	21	0.4193	22	0.6592	21	85
E5	29.5873	2	38.0350	2	44.4550	2	36.7579	2	8
E6	2.0920	18	2.6129	15	3.2956	13	2.9554	16	62
L1	3.2719	12	7.9429	8	8.3337	7	4.7809	9	36
L2	2.3110	15	6.6841	9	3.7124	11	3.9058	13	48
L3	8.2333	8	14.0032	6	7.3412	8	10.7250	7	29
S1	23.0686	4	41.5592	1	47.1205	1	21.0713	4	10
X1	1.6116	20	1.4118	19	0.8863	20	1.3642	20	79
X2	4.3212	11	4.6372	12	3.7420	10	4.0077	10	43
X3	29.2876	3	22.2741	3	27.2043	3	37.2302	1	10

* GGM is the Gordon Growth Model, OJM is the Ohlson-Jeuttner-Nauroth Mode, and PEG is the Price Earnings Growth Model.

17. Appendix J

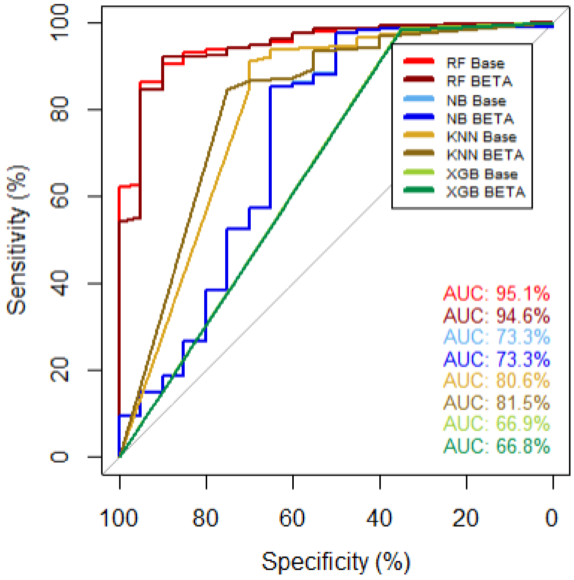

Figure 26: ROC curves for models with Texas ratio over one hundred as distress event in q+1, for beta as added market indicator. The corresponding area under curve (AUC) is also printed per model.

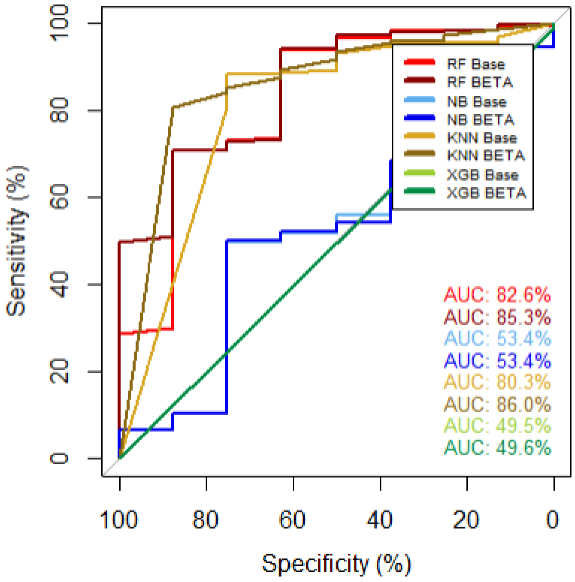

Figure 27: ROC curves for models with Texas ratio over one hundred as distress event in q+4, for beta as added market indicator. The corresponding area under curve (AUC) is also printed per model.

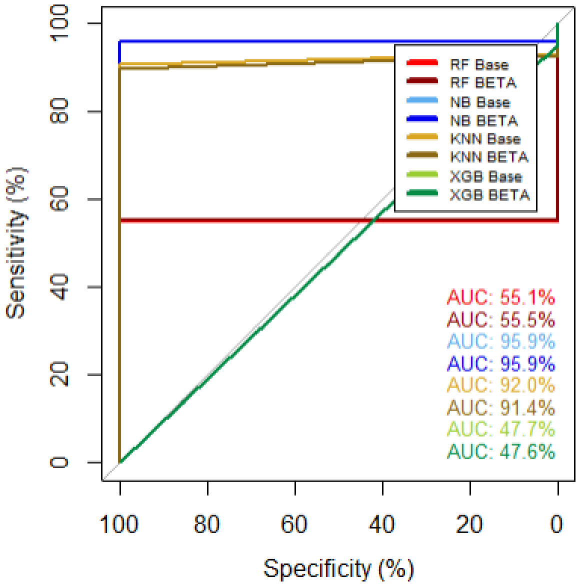

Figure 28: ROC curves for models with Texas ratio over one hundred as distress event in q+8, for beta as added market indicator. The corresponding area under curve (AUC) is also printed per model.

Table 21

Variable importance for randomforest models, calculated with the Mean Decrease in Gini index, for equity beta as market variable. Per variable and model the mean decrease in gini index is shown. The higher numbers mean a larger decrease, indicating these independent variables are more important in predicting the dependent variable for that model. The ranks are shown to indicate how important the different variables are in relation to the other variables.

	$Texas\,ratio_{q+1} > 100\%$		$Texas\,ratio_{q+4} > 100\%$		$Texas\,ratio_{q+8} > 100\%$		
Variable	Gini decrease	Rank	Gini decrease	Rank	Gini decrease	Rank	Sum of ranks
Beta	18.8452	18	21.9419	14	2.2300	22	54
C1	75.6489	8	27.7337	11	6.0610	13	32
C2	163.9155	4	15.0328	19	6.7884	9	32
C3	45.9354	12	24.7056	13	3.1858	19	44
A1	47.4494	11	96.7004	1	5.6020	14	26
A2	17.8116	20	21.2378	16	6.7728	10	46
A3	57.4694	10	21.2873	15	3.8448	18	43
A4	169.3061	3	85.8439	2	60.8860	1	6
M1	35.4020	15	13.5714	21	15.3874	5	41
E1	177.1434	2	75.2600	3	2.6041	21	26
E2	128.6075	5	28.2586	10	4.6607	16	31
E3	18.8284	19	42.8057	7	2.8082	20	46
E4	110.5124	6	19.1238	17	4.2897	17	40
E5	237.1339	1	41.4014	9	11.7561	7	17
E6	33.5243	16	65.5341	4	13.7061	6	26
L1	17.3033	21	42.1360	8	6.6236	11	40
L2	2.1177	22	2.3390	22	5.3423	15	59
L3	27.2565	17	61.9170	5	16.1308	4	26
S1	85.1134	7	57.5763	6	20.6121	3	16
X1	42.3713	13	26.3237	12	57.3935	2	27
X2	67.0432	9	13.9573	20	6.0754	12	41
X3	36.8362	14	18.0819	18	11.1493	8	40