Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Quantitative Finance

On the Value of Adding a Baseline Hazard Rate using Macro-Variables to Survival models for Single-Family Real Estate Mortgage Loans

Name Student: Koen Beijersbergen

Student ID Number: 415870

Supervisor: dr. Erik HJWG Kole

Second Assessor: Karel de Wit

Date Final Version: September 29, 2021

Abstract

Forecast models for single-family mortgage loans are important to financial institutions for risk estimation. It has been shown that both loan defaults and firm bankruptcy show clustering effects. Survival models are capable of modelling these clustering effects for firm bankruptcy. Here we investigate the performance of a survival model to forecast delinquency and default for mortgage loans. We used the Fannie Mae single-family fixed rate mortgage loan data set for training and validation. We find that survival models struggle to correctly predict the delinquency and defaults of mortgage loans. We studied the inclusion of a baseline hazard rate reflecting the changes in macroeconomic environment on forecasting performance in survival models forecasting mortgage loan delinquency and default. We show that a baseline provides benefit to the forecasting accuracy for performing loans, increasing the AUC score from 0.500 to 0.6577. Despite this improvement, further refinement of survival models for single-family mortgage loans remains necessary for their practical use.

Acknowledgements

The completion of this thesis would not be possible without the help and expertise of Dr. Erik Kole, my thesis supervisor. His support and patience were indispensable. I would also like to thank Karel de Wit for his help in reading and reviewing my thesis. Lastly, I would like to thank my parents for their unwavering support and encouragement during my studies.

Contents

1 Introduction

Quantitative prediction models have received attention as early as the 1997 Asian economic crisis. Forecast models that precisely forecast delinquency and defaults of current loans are of great interest to academics, practitioners, and regulators. Regulators use these models to monitor the financial health of banks, lending agencies and other institutions. Practitioners use this information and models to price the risks of mortgage loans. And lastly, academics use the default forecast models to test hypotheses regarding performance of mortgage loans and to predict housing crises. Given the broad application of accurate forecasts, new forecasting tools are of major importance, also contributing to the scientific field of forecasting research.

Empirical research by Das et al.(2002) shows that bankruptcy probabilities for U.S. firms vary over time, and are positively correlated. These correlations between firms vary in time in a manner related to an economy-wide level of default risk. They further show that the joint bankruptcy risk increases as the bankruptcy risk in the economy increases. The positive correlation between firm bankruptcies is known as the clustering effect. This clustering effect is also present in mortgage loans as shown by Ma and Zhao (2018) and Neumann(2018). This clustering effect plays an important role in the risk estimation of both mortgage loans and firm bankruptcy.

Nam et al. (2008) demonstrate improvements in out-of-sample forecasting of firm bankruptcy by using discrete time survival models. These models incorporate the varying nature as well as the correlatedness of firm bankruptcy, the clustering effect. Traditional models fail to reflect the properties of panel data and the influence that stems from the varying macro economic conditions as those vary over time. The improvements in forecasting by these survival models indicate their ability to model clustering effects of firm bankruptcy. As the same clustering is also observed for mortgage loans, it is hypothesized that similar improvements for out-of-sample forecasting are possible when such survival models are applied to mortgage loans.

The financial crisis of 2007-2008 was a severe worldwide economical crisis. An important contributor was the collapse in value of mortgage-backed securities linked to real estate, and the mortgage loan defaults as consequence of predatory lending to low-income home buyers. This global financial crisis highlighted the need for models forecasting mortgage performance, especially those that account for clustering effects.

One of the main characteristics of single-family mortgage loan data used for forecasting is its class imbalance. Class imbalance occurs when the number of observations in one of the classes, i.e. default, is far outnumbered by other classes, i.e. performing. The majority class of loans that never exhibit any delinquency in payments or default vastly outnumbers the class of loans that do display delinquency or default. Researchers address the class imbalance problem through a combination of techniques, such as resampling or synthetic data creation. If models were able to properly deal with class imbalance present in mortgage loan data, that would be of great value.

Different performance measures evaluate various aspects of the forecasting accuracy of prediction models. Some performance measures focus on evaluating the accuracy of probabilistic predictions, such as the Brier Score. Other measures, such as the area under the curve score (AUC) of the receiver operating characteristic curve (ROC) focus on the evaluation of the minority class classification ability of the model. This evaluation of the minority class classification ability is of particular importance for models that use data with a class imbalance problem.

In this paper we investigate the value of using a longitudinal survival model for out-of-sample forecasting of delinquency and default probability for single-family mortgage loans. We use and adapt survival models developed in Shumway (2001) and Nam et al. (2008) and use data from the Fannie Mae single-family fixed rate mortgage loan data set. We test the survival model on single-family mortgage loans and tried to improve the survival model by adding a baseline hazard rate estimated using macro-variables to simulate variations in the macro-economic environment.

Our results find that the use of the Gross Domestic Product (GDP) Growth Rate and the Interest Rate are appropriate macro-economic variables to model a baseline hazard rate for the delinquency and default probabilities of the loans. Linear Regression of these two variables on the average delinquency or default rate gives significant values for their estimated parameters to model their relation to these transition rates.

Our findings give us a number of insights. Firstly, the results of the Brier Score display high values for survival models with and without the baseline in the forecasting of the transition of loans in the performing state. These scores decrease significantly for both models for their forecasting of loans in the one or two month delinquent state. The high

values of Brier Scores indicate to us extremely weak performance for both survival models for the out-of-sample forecasting power for loans in initial state performing. This is likely a result of the extreme class imbalance that is present in this initial state. The models appear to be largely unable to predict the exact loans that will turn delinquent that month as a large number of the loans are exposed to the same level of risk, but few loans transition. The lower values of the Brier Scores for forecasting of loans in initial states 1 or 2 months delinquent indicate improved probabilistic predictive power for forecasts of loans in those states.

The findings with the Brier Scores are supported by the findings of the receiver operating characteristic (ROC) curves of the forecasts of both models. The area under the curve (AUC) scores similarly indicate weak predictive power for both survival models as well. For predictions of loans in the performing state see values for the AUC close to 0.5 for the survival model without the baseline, and a value of 0.6577 for the survival model with the baseline. For the survival model without the baseline this increases to 0.5500 for loans starting in the 2 months delinquency state, where it decreases to 0.4287 for the model including the baseline. This similarly indicates overall poor performance of the survival models, but the varying numbers result from the AUC accounting differently than the Brier Score for the classification of the extreme minority class transitions.

Both these performance measures indicate weak predictive out-of-sample forecasting power for the survival models. This resulting from the scores of the Brier Score and the AUC scores that we find for both the model including and excluding the baseline. However, the model that includes the baseline shows significantly higher AUC scores for predicting loans with the initial state performing. These scores reach a maximum of 0.6577 for the AUC score as opposed to the value of 0.500 for the model excluding the baseline. This does show significant increase in performance for the predictions of loans that are in the performing state when including a baseline in the survival model. However, the predictive power in isolation is poor.

Our findings indicate that estimation of a baseline hazard rate is possible for a discrete time survival model through the use of a combination of macro-economic variables. The improvement in performance with the inclusion of this extension indicates to us that for further development of survival models this extension to the survival model should be included. However, both of the survival models show weakness in predictive performance and require further improvement.

The paper is organized as follows. Section 2 presents background and the discrete-time survival model with time-varying covariates and a baseline hazard function reflecting macro-economic condition effects and loan default correlations. In section 3 we present the methods by which we measure the performance of the models and the added value of the baseline. In section 4 we present the data used in the empirical analysis. In section 5 we empirically test the survival models and discuss the results. Section 6 contains the conclusions and discussion.

2 Theoretical Background

Discrete time survival models improve on the out-of-sample forecasting performance of static models for firm bankruptcy, as shown by Nam et al. (2008). We use the survival model as opposed to a static model. Static models include only one risk estimation across the lifetime of the loan, and a single binary outcome for their default or delinquency along that period. The use of a static model results in both biased as well as inconsistent probability estimates, whereas the survival model we use results in consistent and unbiased estimations, as shown by Shumway (2001). The survival model that forms the basis for our model outperforms the current benchmark models in terms of out-of-sample forecasts for firm bankruptcy over a long time period. Using the survival model, we develop a default or delinquency estimation model that uses idiosyncratic variables as well as macro-economic variables to identify the probability of transitioning for the mortgage loans. We test the survival model on the single-family mortgage loan data set to observe if we achieve similar out-of-sample forecasting performance.

In our model the survival time is denoted by T, the time where the loan reaches delinquency status. T is a continuous random variable that follows from a probability density function $f(t)$, and a cumulative density function $F(t)$. The survival function is represented by $S(t)$, which gives the survival probability over time span t. This survival function is defined as follows:

$$S(t) = Pr(T \geq t) = 1 - F(t) = \sum_{t}^{\infty} f(u)du$$

The hazard function h(t) is measured as the conditional probability of default at time t given survival to that time.

$$h(t) = lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

This describes the instantaneous risk of default, and is used in continuous time survival models. These survival models contain a hazard function. Widely used is the Cox(1972) semi-parametric proportional survival model:

$$h(t|x_i) = \exp\{x_i'\beta\}h_0(t)$$

In the formula the covariates for each loan are represented by x_i. The firm specific part is time-invariant and is represented by the first part of the equation. The second part is the baseline hazard function and is time-dependent. In Cox (1972), the covariates

vector proportionally influences the hazard function and as such is also time-dependent. The parameters for this equation are estimated by maximization of the accompanying partial likelihood function.

$$L(\beta) = \prod_{i=1}^{k} \frac{exp(x_i'\beta)}{\sum_{j \in R(t)} exp(x_j'\beta)}$$

The maximum likelihood estimator here has a beneficial covariance matrix as shown by Lawless (1982) and is asymptotically normally distributed. In recent years, with the need for time varying covariates, the hazard formula has been adjusted into the following form:

$$h(t|x_{i,t}) = exp\{x_{i,\tau}'\beta\}h_0(t)$$

With parameters $h(tx_{i,t})$ as the individual hazard rate of the loans, and the $x_{i,t}$s as the covariate vectors with the idiosyncratic variables of each loan. Shumway (2001) describes that using the old hazard formula with multi-period data would lead to bias and inconsistency as a result of the static nature of the model.

Beck et al. (1998) show that the continuous proportional hazard model, given by the above equation, can be estimated via multi-period logit models, they also show the discrete analogue of this continuous hazard rate. This is the discrete hazard rate we use in our discrete-time survival model, represented by the following formula.

$$P(y_{i,t} = 1|x_{i,t}) = h(t|x_{i,t}) = 1 - exp\{-exp(x_{i,t}'\beta + \kappa_t)\}$$

In this formula the k_t is a dummy variable that represents the length of non-failure that have occurred prior to the default. Beck et al. (1998) show how the following equation can be used when the probabilities of failure are sufficiently small as present in our data.

$$P(y_{i,t} = 1|x_{i,t}) = h(t|x_{i,t}) = \frac{1}{1 + exp\{-(x_{i,t}'\beta + \kappa_t)\}}$$

Shumway (2001) defined a multi-period logit model as 'a logit model that is estimated with data on each firm in each year of its existence as if each firm-year were an independent observation' and show that a multi-period logit model is equivalent to a discrete-time survival model because the likelihood functions of the two models are identical. As such we can estimate the discrete-time survival model with time-varying covariates with tools of analysis of binary dependent variables.

The discrete-time survival model still uses a discrete hazard rate, rather than a continuous rate. In our discrete time-model we use a single month as our discrete time period. Our discrete model uses the discrete hazard rate. The discrete hazard rate is defined as the probability of a negative transition in a one month time period. The discrete hazard rate transition probability is equal to the transition probability of a loan that is exposed to the continuous hazard rate for the entire month.

The discrete hazard rate formula we use contains a baseline hazard rate term, the k_t term. In the work of Beck et al. (1998) this term represents the length of the sequence of zeros of non-failure that occurred prior to the default. With that their baseline hazard rate implies that the individual hazard rate is is determined by a firm's survival period. In our paper we use a term for our k_t that is dependent on the macro-economic environment at that point in time. This baseline is represented in our model by globally varying macro-economic variables that simultaneously shift for all loans. The globally shifting macro-economic variables model the varying macro-economic environment and how this environment affects the average delinquency and default rate the loans are exposed to.

The base line hazard rate term allows for various forms of survival models to be used by varying the specification of this baseline. For example, a constant allows for duration-independent hazard rate. However, use of this formula differs from Shumway (2001) and Beck et al.(1998) as they use a dummy variable which marks the length of non-failure prior to failure as their baseline hazard term. This implies an individual hazard rate that is defined by each firm's survival period. These indirect measures however do not allow for capturing macro-economic effects as the historic survival data does not reflect the overall macro-dependencies and correlations of loan failure. Recent economic crises have brought the role of macro dependencies in loan failure to the forefront. To model the correlation between macro-economic conditions and the high tendency of clustering for loan failure, we propose using these same macro-economic variables as the determining factors for our baseline hazard rate. Hillegeist et al. (2001) take a similar approach to handling temporal dependencies by using two direct measures of the baseline hazard rate; the rate of recent defaults and changes in interest rates (CIR). In our analysis we examine other macro-economic factors as well. We do use the CIR as suggested by Hillegeist et al. (2001), but also observe the Real Estate Prices and the GDP Growth.

For our empirical analysis we formulate two models to show the two conditions that need to be satisfied for an efficient estimation. Firstly, idiosyncratic covariates must be

allowed to vary with time and allow for measurement of survival analysis. Secondly, the necessity of a baseline hazard function calculated directly with macro-economic variables to reflect the changes in the macroeconomic environment.

In the empirical analysis we make use of a survival model without baseline that uses covariates that fluctuate over the lifetime, allowing for observations across the lifetime of the loan and the estimation of the hazard across the lifetime. We also use the survival model with added time dependent baseline hazard rate. This allows for temporal changes in the covariates as well as in the baseline hazard rate at that time. By comparison between the performance of these two models we address the question of the value of the baseline hazard.

3 Data

For the empirical comparison of our study we use the Fannie Mae Single-Family Fixed Rate Mortgage Data Set. This data set contains monthly acquisition and performance data. From this data set we use the time period from 2000Q1 to 2018Q4. The data is split into an Acquisition file and a Performance file. The Acquisition file contains the data of the moment of inception for the loan, whereas the Performance file contains the monthly performance data for each individual mortgage loan. The threshold Fannie Mae set for loan default is three month payment delinquency. We also use this three month delinquency as the default status.

The data set is of considerable size consisting of around 200 million monthly data points for loans and loan performing status. One other defining feature of the data is that the fraction of loans that defaults is extremely low, with a mean of only 1.2 percent defaulting, showing clear class imbalance. Class imbalance impacts the forecasting performance of the models that are compared. However, as we evaluate the relative forecasting performance of the models, in the single-family mortgage loan space class imbalance presents a prevalent feature. The prevalence of class imbalance in single-family mortgage loan data makes superior forecasting performance of data with class imbalance highly beneficial. To tackle the class imbalance problem we use performance measures that focus on probabilistic predictive power and minority class classification power.

As a first step, we split the data set into three groups containing different loans in each group. Each group contains the loans that have a single initial state for their period in time. The groups are as follows: Performing, 1 Month Delinquent, 2 Months Delinquent. These three groups cover all three initial states that a loan can have at one point in time in our data set. We separate those groups for the purpose of estimating separate models for each of these three initial states as they represent different transitions. We define four classes for our loans, four separate states that the loans can be in. The performing state, where loans are up-to-date on all their payments. The one month delinquent state, where loans have entered delinquency in the last month. The two month delinquent state, where loans have been delinquent for the past two months and the defaulted state, which loans enter when they are 3 months delinquent or longer.

Our duration survival model including the baseline hazard makes use of idiosyncratic explanatory variables to calculate the individual risk that is present in each of the loans at each point in time. The Fannie Mae data set contains 108 variables for each loan. 26 of those variables pertain to information about those loans after they have been defaulted on. This information is not available for those loans as they are being managed in the portfolio, as such we dismiss those variables as possible idiosyncratic variables to use in our models, secondly a group of 10 variables pertain to information that is lender sided, such as the insurance firm the lender uses for this loan or the property valuation agency used. These variables do not affect the loans themselves and as such we dismiss those variables for use. Eight of the variables pertain to names of various aspects of the loan, containing no information for the risk of the loans as they are either identical for all loans or unique for each loan. We dismiss these variables for use as well. We find that for 29 variables, only 20 percent of the loans have data available. We choose to exclude these variables from the model because their missing data does not allows us to properly evaluate the effects of these variables. We have 10 groups of variables that contain equivalent information. For each of these groups we select one variable to represent the information while excluding the others. Seven variables all pertain to a variation of the age, origination, or time to maturity. These variables all contain equivalent information; we choose to move forward with the Loan Age variable and exclude the other variables from use. We start formal testing with eight numeric covariates and seven categorical covariates (Table 1).

Table 1: Features we consider before selection with Information Value

Numeric Variables	Categoric Variables
1: Original Interest Rate	9: Origination Channel
2: Original UPB	10: First time Home Buyer*
3: Original Loan Term	11: Loan Purpose
4: Original Loan-To-Value	12: Property Type
5: Number Of Borrowers	13: Occupancy Type
6: Original Debt-To-Income Ratio	14: Relocation Morgage Indicator
7: Borrower Credit Score at Origin	15: Loan Age**
8: Number of Units	

*A value of 1 for this dummy corresponds to a mortgage given to a first time home buyer. **The Loan Age dummy has a value of 0 for loans younger than 40 months old and a value of 1 for loans exceeding 40 months in age.

Following this first selection we make use of the backwards selection technique for further selection of risk drivers for the characteristics of the loan to be included in the models. The technique fits a default vector (a dummy with 1 for a defaulted loan and 0 for a non-defaulted loan) on all considered variables and uses stepwise elimination of the covariates that is least significant. These steps are repeated until all the covariates still included are significant in the logistic regression of the default vector.

Because of the size of the data set we include an additional variable selection method, the information value criterion. This criterion indicates the strength of influence of the covariates with respect to the default. We select the covariates with information values exceeding 0.1. Covariates that exceed this Information Value have a stronger than medium predictive power. The IV is computed as

$$\sum_{i=1}^{k} \left[(Distr_{1i} - Distr_{0i}) \cdot \log \left(\frac{Distr_{1i}}{Distr_{0i}} \right) \right] \tag{1}$$

With the following guideline of Siddiqi (2006), we can interpret the strength of a covariate as:

1. Less than 0.02, then the predictor is not useful for modeling (separating the defaulting loans from the non-defaulted loans)

2. From 0.02 to 0.1, then the predictor has only a weak relationship to the defaulted loan/non-defaulted loan odds ratio

3. From 0.1 to 0.3, then the predictor has a medium strength relationship to the defaulted loan/non-defaulted loan odds ratio

4. 0.3 or higher, then the predictor has a strong relationship to the defaulted loan/non-defaulted loan odds ratio.

These two selection methods result in five idiosyncratic variables for the survival model; Original Interest Rate, Original Loan-To-Value, Original Debt-To-Income, the First Time Home Buyer Indicator, and the Loan Age.

The data is split into two parts; a training set and a test set with a seventy-thirty split respectively. This split is applied after division of the loans into three separate groups on the basis of their initial state at that point in time; Performing, One Month Delinquent or Two Months Delinquent. This as all three initial states are trained separately and have separate estimated models. As such we have a total of three training data sets and three test data sets. We have 180.000.000 data points in the initial state "performing", 600.000 data points in the initial state "1 month delinquent", and 60.000 data points in the initial state "2 months delinquent.

For the estimation of the baseline hazard macro-variables from the Federal Reserve Bank of St. Louis are used. This macro-variable data is quarterly. As our loan data is monthly, the quarterly data needs to be transformed for use in our model. The Federal Reserve Bank provides quarterly data as an average over that period. We take their data as the three month average across that quarter. With this quarterly data being the three month average we take the monthly value of each of the three months in that quarter to be equal to that three-month average. Although, this is not a complete accurate representation of the monthly shifts in the variables it is the closest approximation we can use. It does manage to capture the long term movements of the variables and with a time span sufficiently large to model the effects of the variance in these variables.

Variable Differences

Before testing we compare the values of the idiosyncratic explanatory variables for the loans that transition into delinquency or default versus those that do not. Tables 2, 3, and 4 list these values.

Table 2: Values of Explanatory Variables for Loans with the Initial State Performing

Explanatory Variable	No Transition	Transition to 1 Month Delinquent
Loan Age (Dummy)	0.34 (0.06)	0.39 (0.06)
Original Interest Rate	3.87 (0.24)	4.03(0.24)
Original Loan-To-Value	78.4 (2.87)	78.98 (2.87)
Original Debt-To-Income	32.3 (0.87)	34.8 (0.87)
First Time Home Buyer Indicator (Dummy)	0.175 (0.07)	0.200 (0.07)

Comparison of values for explanatory variables for loans starting in the Performing state that transition to the 1 month Delinquent status, versus the loans that do not transition

Table 3: Values of Explanatory Variables for Loans with the Initial 1 Month Delinquent

Explanatory Variable	Transition to Performing	Transition to 2 Months Delinquent
Loan Age (Dummy)	0.405 (0.06)	0.490 (0.05)
Original Interest Rate	4.02 (0.27)	4.07 (0.26)
Original Loan-To-Value	78.8 (2.40)	79.77 (2.25)
Original Debt-To-Income	34.6 (0.74)	35.89 (0.46)
First Time Home Buyer Indicator (Dummy)	0.195 (0.05)	0.241 (0.05)

Comparison of values for explanatory variables for loans starting in the 1 month Delinquent state that transition to the Performing status, versus the loans that transition to the 2 months Delinquent state

Table 4: Values of Explanatory Variables for Loans with the Initial 2 Months Delinquent

Explanatory Variable	Transition to Performing	Transition to Default
Loan Age (Dummy)	0.508 (0.06)	0.513 (0.04)
Original Interest Rate	4.05 (0.25)	4.06 (0.26)
Original Loan-To-Value	79.4 (2.11)	80.0 (2.29)
Original Debt-To-Income	35.4 (0.25)	35.9 (0.28)
First Time Home Buyer Indicator (Dummy)	0.245 (0.04)	0.227 (0.05)

Comparison of values for explanatory variables for loans starting in the 2 month Delinquent state that transition to the Performing status, versus the loans that transition to the Defaulted State

Transition Rates

We note the initial transition rates and their variance over time. We observe the trend of the probability of a transition increasing as the loans become more delinquent (Figure 1). Identical spikes in transition rate for all three initial states near the tail end of the time period are observed. Interestingly, we observe low initial transition rates in the time period for initial state performing and initial state 1 month delinquent. In contrast we observe transition rates for the initial state 2 months delinquent close to the maximum value across the full time period.

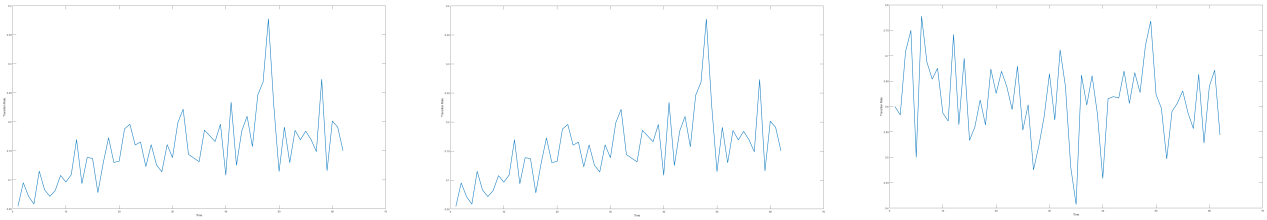

Figure 1: Transition Rate over Time for Loans with Initial State Performing, 1 Months Delinquent, and 2 Months Delinquent

4 Methods

We train and apply our models using empirical data. The training data is used to estimate the parameters and the test data is used to create forecasts and evaluate the out-of-sample forecasting performance. Consistent and robust performance measures are essential to asses the forecasting performance of a model. We use two different measures to compare the relative forecasting performance of the models and to evaluate their individual forecasting performance. The performance measures we use to evaluate individual and relative forecasting performance are the Brier Score and the area under the curve (AUC) score of the receiver operating characteristic (ROC) curve. We consider the AUC as most informative as it gauges the discriminatory ability of the models. We are particularly interested in the discriminatory ability of the models. As mentioned above the data associated with mortgage loan performance is subject to class imbalance. The class imbalance and the impact of the minority class transition makes minority class classification power particularly valuable for our model.

Area Under the ROC Curve

To create the ROC curve we classify the forecasts into four categories; True Positive, False Positive, True Negative, and False Negative. The True or False labels denote whether a forecast is True or False. The Positive or Negative labels denote whether the forecast predicted a Positive or Negative classification (performing versus delinquent, further delinquent, or defaulted). For the ROC curve we plot the True Positive classification rate ($\frac{TP}{TP+FN}$) against the False Positive classification rate ($\frac{FP}{TN+FP}$). AUC is defined as the Area Under the ROC curve and reflects the discriminatory power of the classifier. The AUC score can also be interpreted as the probability that a randomly chosen loan that defaulted, receives higher default probability than a randomly chosen loan that did not default as mentioned in Lessmannn et al. (2015). The minimum score for a correct model would be a score of 0.5000. A model the predicts exclusively positive transitions achieves this score of 0.5000.

Brier Score

The Brier Score(BS) is calculated as the mean-squared error of the probability estimate $\hat{p}_i$ and the binary dependent variable y_i as seen in Hernandez-Orallo et al. (2011), that is

$$\text{BS} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (\hat{p}_i - y_i)^2. \tag{2}$$

The Brier Score measures the accuracy of the probabilistic predictions. The Brier score allows us to evaluate the general out-of-sample forecasting performance of the models.

Macro Variables

We specify a number of potential factors for use as the explanatory variables of the baseline hazard rate for out survival models.

We specify the national Gross growth, the real estate price index growth rate, the change in interest rate (CIR), and the interest rate (IR) itself. We apply a t-test to test the explanatory power of the macro-economic factors against the baseline hazard rate one period into the future. Results of the t-test indicate that GDP Growth and the Interest Rate best model the baseline hazard rate. We use these two macro-economic variables as the specification for the baseline hazard rate.

Before we continue with the chosen explanatory variables, we normalize variables that benefit from normalization. The variables we normalize are both macro-variables, and the Original Debt-To-Income ratio. Both of the macro-variables, and the Original Debt-To-Income ratio exhibit the form of a normal distribution. To improve parameter estimates we normalize these variables by transforming the values of the parameter vectors into their z-scores with a mean of 0 and a standard deviation of 1.

Using the information of the explanatory variables and the macro-variables, we now apply the survival model excluding baseline, and the survival model including baseline to the training set. With this we obtain the estimation results for both models resulting in six total estimation results, three for either model covering all three initial states.

5 Results

Training the models using the empirical data results in the estimation results represented in tables 5, 6, and 7. These are the averaged results of the estimation done over 5 training sets. These parameters represent the effect the explanatory variables have on the probability of a transition into a state of delinquency, further delinquency or default. The probability of transition and consequently classification into the negative class increases with increasing values of the parameters. The parameters coefficients represent the rate of change in the logarithmic probability of negative transition as the explanatory variable changes.

Table 5: Parameter Estimation Results for initial state Performing

Explanatory Variable	survival model excluding baseline	survival model including baseline
Intercept	7.5 (0.58)	7.6 (0.66)
Loan Age (Dummy)	-0.15 (0.17)	-0.16 (0.31)
Original Interest Rate (Normalized)	-0.39 (0.04)	-0.38 (0.04)
Original Loan-To-Value (Normalized)	-0.01 (0.00)	-0.01 (0.00)
Original Debt-To-Income (Normalized)	-0.03 (0.00)	-0.03 (0.00)
First Time Home Buyer Indicator (Dummy)	-0.11 (0.08)	-0.11 (0.09)
GDP Growth Rate	-	0.00 (0.09)
Interest Rate	-	0.04 (0.03)

Table 6: Parameter Estimation Results for initial state 1 Month Delinquent

Explanatory Variable	survival model excluding baseline	survival model including baseline
Intercept	3.1 (0.42)	3.2 (0.51)
Loan Age (Dummy)	-0.2 (0.17)	-0.27 (0.33)
Original Interest Rate (Normalized)	-0.16 (0.03)	-0.15 (0.03)
Original Loan-To-Value (Normalized)	-0.01 (0.00)	-0.01 (0.00)
Original Debt-To-Income (Normalized)	-0.02 (0.00)	-0.02 (0.00)
First Time Home Buyer Indicator (Dummy)	-0.22 (0.04)	-0.21 (0.05)
GDP Growth Rate	-	0.06 (0.01)
Interest Rate	-	0.04 (0.05)

Table 7: Parameter Estimation Results for initial state 2 Months Delinquent

Explanatory Variable	survival model excluding baseline	survival model including baseline
Intercept	0.66 (0.25)	0.68(0.24)
Loan Age (Dummy)	0.09 (0.13)	-0.08 (0.14)
Original Interest Rate (Normalized)	-0.04 (0.01)	-0.03 (0.02)
Original Loan-To-Value (Normalized)	-0.01 (0.00)	-0.01 (0.00)
Original Debt-To-Income (Normalized)	-0.01 (0.00)	-0.01 (0.00)
First Time Home Buyer Indicator (Dummy)	0.1 (0.07)	0.11 (0.07)
GDP Growth Rate	-	0.13 (0.04)
Interest Rate	-	0.06 (0.06)

In these estimation results we observe the value of the intercept at a maximum value of 7.6 for loans in the initial state performing. The intercept decreases to a value of 3.2 for loans in the initial state 1 month delinquent, further decreasing to 0.66 for loans in the initial state 2 months delinquent. Together this indicates the general risk of transitioning into a negative state increases with loan delinquency status.

We also observe a large standard deviation for the parameter of the loan age explanatory variable. Parameter estimations for the explanatory variable range from a value of -0.37 to 0.17, indicating differences in the effect of the loan age between training sets. Loan Age increases the probability of negative transition for some training sets and decreases the probability of negative transition for others.

For both the initial state Performing class and the initial state 1 Month Delinquent class we observe negative values for all the estimated parameters for the idiosyncratic covariates. This indicates that all idiosyncratic covariates increase the probability of a negative transition. All estimated parameter values are also negative for the initial state 2 months delinquent with the exception of the estimated value for the parameter of the First Time Home Buyer indicator. We hypothesize this results from first time home buyers near the edge of defaulting were not as prepared or aware of the debt they were taking on and as such once they get into delinquency, their risk of default is higher.

We observe nearly identical estimated parameter values for the explanatory variables in both the survival model including baseline and survival model excluding baseline. This creates confidence in the estimated values for these parameters as well as their significance and explanatory power against the baseline hazard rate estimated by the macro-economic variables. We observe positive parameter values for all the estimated macro-variable parameters, in contrast we observe smaller values for estimated parameter of the First Time Home Buyer indicator than hypothesized. These values indicate that increases in

our macro-variables correspond to less delinquencies and defaults across the board for all loans, effectively decreasing the baseline delinquency and default rate with increases in both macro variables.

We use the estimated survival model without the baseline to create an in-sample forecast for a first estimate of the forecasting accuracy. These forecasts were used to calculate the 1 month ahead forecast of each of the data points available. Forecasting results indicate transitions from performing to 1 month delinquent or staying in performing; for 1 month delinquent back to performing or to 2 months delinquent and for 2 months delinquent back to performing or to fully defaulted. The prediction cut-off used here is a probability value of 0.5. The results, presented in table 8, represented the results of our first forecast in terms of true and false positives and negatives.

Table 8: In-Sample Forecasting Performance survival model excluding baseline

Simple Forecasting Performance	Forecasting Result			
Initial State	True Positive	True Negative	False Positive	False Negative
Performing	0.996	0.001	0.999	0.004
1 Month Delinquent	0.859	0.230	0.770	0.141
2 Months Delinquent	0.350	0.604	0.396	0.650

In table 8 we observe the difficulty encountered by the model in the forecasting of the transition of performing loans into delinquency. This is indicated by the low value for the True Negative. The in-sample forecast struggles to forecast the exact moment for loans to enter delinquency. The forecasting of the transition further into delinquency or into default is improved, indicated by the increase in the value of the true negative from 0.001 to 0.230 , and to 0.604 respectively. We observe a decrease in accuracy of forecasting the true positives for the loans that have an initial state of 1 month delinquent. The True Positives decrease from 0.996 for the initial state performing to 0.859 for the initial state 1 month delinquent. Further decrease in accuracy of the True positives for the initial state 2 months delinquent is represented by the decrease in the True Positives from 0.859 to 0.350. Initial in-forecast analysis shows us that the model struggles to combine accuracy for both positive and negative classifications.

With these initial results we now conduct the calculations and forecasting out-of-sample for both the including and excluding baseline Hazard Rate survival model to estimate and evaluate their predictive power. Using our Training set parameter estima-

tion we forecast using the initial states from the test data set. This gives us the True and False, Positive and Negatives rates for the Out-Of-Sample forecasting for the survival model excluding baseline presented in Table 9.

Table 9: Out-Of-Sample Forecasting Performance survival model excluding baseline

Out-Of-Sample Forecasting Performance	Forecasting Result			
Initial State	True Positive	True Negative	False Positive	False Negative
Performing	0.997	0.010	0.990	0.003
1 Month Delinquent	0.824	0.251	0.749	0.176
2 Months Delinquent	0.394	0.655	0.345	0.606

We observe similar results to the in-sample forecast for the out-of-sample forecast. The True Positive rate shows a high value for initial state performing (0.997), decreasing for initial state 1 month delinquent (0.824) and decreasing for the initial state 2 months delinquent (0.394). We do observe an increase in the true negative value for all three initial states, from 0.001, 0.230, and 0.604 to 0.010, 0.251 and 0.655. This indicates improved forecasting accuracy for the minority class predictions. Table 10 presents the out-of-sample forecasting results for the survival model including baseline.

Table 10: Out-Of-Sample Forecasting Performance Survival model including baseline

Out-Of-Sample Forecasting Performance	Forecasting Result			
Initial State	True Positive	True Negative	False Positive	False Negative
Performing	0.994	0.010	0.990	0.006
1 Month Delinquent	0.827	0.264	0.736	0.173
2 Months Delinquent	0.322	0.614	0.386	0.678

We also observe similar results to the out-of-sample forecast of the survival model excluding baseline for the out-of-sample forecast of the survival model including baseline. There is an increase in the value for the True Negative rate for the initial state 1 month delinquent versus the True Negative rate for that initial state in the model that does not include the baseline. The value of the True Negative rate increases from 0.251 to 0.264, indicating improved minority class predictive power for loans in initial state 1 month delinquent. We also observe a decrease in the value of both the True Positives and True Negatives for the initial state 2 months delinquent versus the results of the survival model that does not include the baseline. The values decrease from 0.394 and 0.655 to 0.322

and 0.614.

With these results from the first partition of the data set, we now repeat the model estimation for both models using different parts of the data set to evaluate the AUC and the Brier Score for the models under different out-of-sample forecast data sets. First, we obtain the results for the survival model excluding baseline, and followed by those same results for the survival model including baseline.

Table 11: Repeated Results for the survival model excluding baseline

Survival model excluding baseline	Data Set 1		Data Set 2		Data Set 3	
Initial State	AUC	Brier Score	AUC	Brier Score	AUC	Brier Score
Performing	0.5000	0.9909	0.5000	0.9867	0.5000	0.9886
1 Month Delinquent	0.5754	0.5431	0.5585	0.5062	0.5872	0.5183
2 Months Delinquent	0.5302	0.3086	0.5500	0.3384	0.5390	0.3454

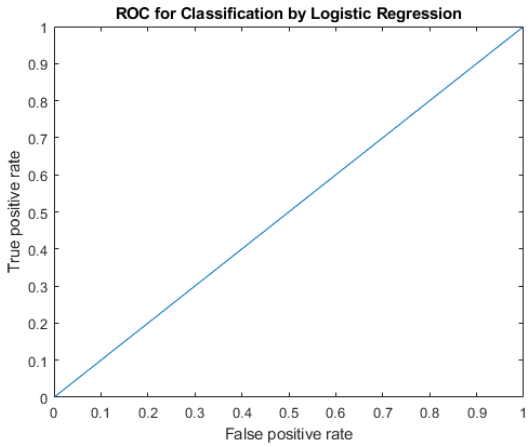

Figure 2: Area under the ROC Curve for survival model excluding baseline and initial state performing

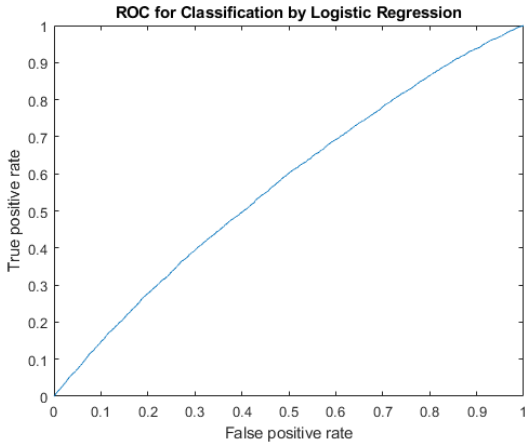

Figure 3: Area under the ROC Curve for survival model excluding baseline and initial state 1 month delinquent

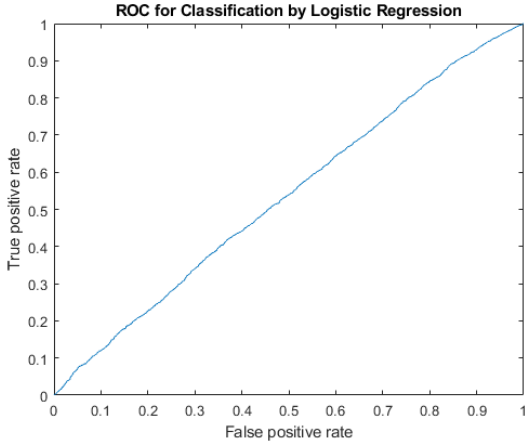

Figure 4: Area under the ROC Curve for survival model excluding baseline and initial state 2 month delinquent

Table 11 shows the results of the Performance Measures and figures 2, 3, and 4 display the ROC curves for the three initial states for the survival model excluding baseline. The values of the AUC are identical for all three of the results (Table 11). The value of the AUC is close to 0.5 indicating weak predictive power of the minority class. The AUC scores increase for the initial state 1 month delinquent show a small increase in minority class predictive power. The AUC scores for the initial state 2 months delinquent equals the AUC score for the initial state 1 month delinquent for one of the results and is lower for two of the results. This indicates a slight increase in minority class predictive power versus the initial state performing. These findings are reflected in the ROC curves. Figure 3 shows that the curve for initial state performing is a straight line from $(0, 0)$ to $(1, 1)$, indicative of low minority class predictive power. The curve for initial state 1 month

delinquent has a noticeable curve to it (Figure 3), indicating improvement over the curve for initial state performing. The curve for initial state 2 months delinquent shows a wavy curve(Figure 4), indicating performance in between those for the two other initial states. For the Brier Score we observe extremely low probabilistic predictive power for loans with initial state performing. The value of the Brier for initial performing is over 0.9867 for all three sets. This value is close to 1, the minimum score possible for this performance measure. For the initial state 1 month delinquent the Brier Scores decrease to values around 0.52, indicating improved probabilistic predictive power. The Brier Score further decreases to values around 0.33 for the initial state 2 months delinquent, indicating further increase in probabilistic predictive power. We now represent the results of the AUC and the Brier Score for the survival model including baseline in Table 12. Figures 5, 6, and 7 represent the ROC curves calculated for the survival model including baseline.

Table 12: Repeated Results for the survival model including baseline

Survival model excluding baseline	Data Set 1		Data Set 2		Data Set 3	
Initial State	AUC	Brier Score	AUC	Brier Score	AUC	Brier Score
Performing	0.6438	0.9909	0.6348	0.9867	0.6577	0.9886
1 Month Delinquent	0.6022	0.5442	0.5814	0.5063	0.5910	0.5190
2 Months Delinquent	0.4605	0.3111	0.4526	0.3404	0.4287	0.3507

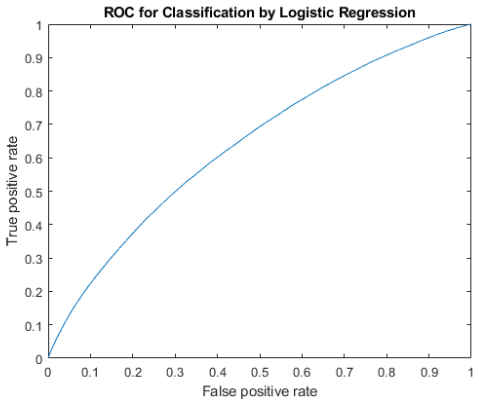

Figure 5: Area under the ROC Curve for survival model including baseline and initial state performing

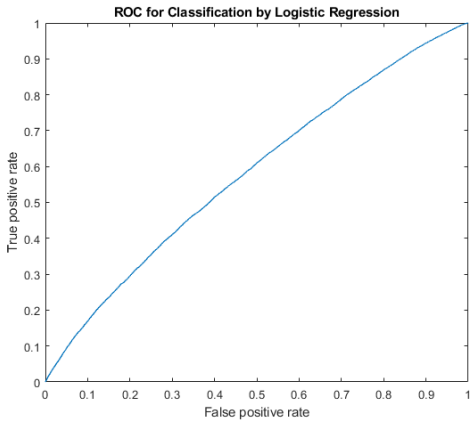

Figure 6: Area under the ROC Curve for survival model including baseline and initial state 1 month delinquent

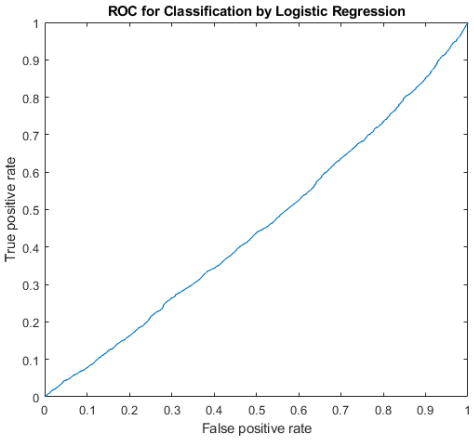

Figure 7: Area under the ROC Curve for survival model including baseline and initial state 2 month delinquent

In table 12 we note that the initial state performing for the survival model including baseline has the highest values for the AUC, 0.6577. The AUC values indicate an improvement for the minority class predictive power of the survival model when the baseline is added to the survival model. We observe a slight improvement in the AUC score of the initial state 1 month for the survival model including baseline compared to the AUC score of the survival model excluding baseline. The AUC score increases from around 0.57 to around 0.59. This indicates a corresponding increase in minority class predictive power. For the initial state 2 months delinquent we observe AUC scores of under 0.5, this indicates classification power that is outperformed by simple forecasting, i.e. forecasting the majority class to every loan. The ROC curve in figure 5 reflects the increase in AUC score in the prominent curve present. The ROC curve in figure 6 displays a slight curve,

similar to the ROC curve in figure 3. This mirrors the identical AUC scores for both survival models for the initial state 1 month delinquent. The ROC curve for initial state 2 months delinquent, represented in figure 7 reflects the AUC scores of under 0.5 in the convex curve we observe. The Brier Scores for the survival model including baseline are nearly identical to those for the survival model excluding baseline. They indicate identical low probabilistic predictive power for loans with initial state performing. they indicate increase in probabilistic predictive power for loans with initial state 1 month delinquent, versus initial state performing. They also indicate a further increase in probabilistic predictive power for loans with initial state 2 months delinquent.

6 Discussion and Conclusions

In our research we evaluated the use of survival models for the forecasting of single-family mortgage loan delinquency and default. We used and compared survival models including and excluding a baseline hazard rate. For the explanatory variables idiosyncratic covariates were selected through the use of the Information Criterion and the backwards selection technique. We studied the effect of adding a baseline, estimated by macro-variables, to the survival model.

To test the survival models we used an empirical data set from Fannie Mae to empirically estimate the parameter coefficients for the explanatory variables. We split the data into three groups of loans based on the initial at that point in time; performing, 1 month delinquent, or 2 months delinquent. All three initial states show class imbalance. The class imbalance is most severe for the initial state performing.

In the results we see that the survival models face significant hardship in their forecasting of the transitions of the loans. This is present in both survival models, including and excluding the baseline. This is reflected in the values of the AUC scores, none of the models exceed an AUC score of 0.6577, which shows that both models struggle with overall performance. This does show a weakness of the basic survival model we use in this research as the AUC reflects specifically on the ability of the loans to classify the minority class transitions, which are most important to us.

The lack of overall performance of both models is also supported by the value of the Brier Scores. The Brier Scores did show near identical performance for survival including and excluding baseline. Both models showed the poorest Brier Score performance in the prediction of loans in the initial state performing. The Brier Score showed improved scores for predictions of loans in states 1 and 2 months delinquent. One of our main interests deals with the data set of loans in the performing state. The Brier Score also indicates a poor probabilistic performance by the survival models for this group.

The AUC scores showed an increase in the performance for the survival models excluding baseline compared to the survival models including baseline for loans in initial state performing. The AUC scores for the survival model including baseline were higher than those of the survival model excluding baseline. However, the AUC scores for the initial state 2 months delinquent are decreased for the survival model including baseline com-

pared to the survival model excluding baseline. The performance of both models is poor in regards to loans in initial state performing as indicated by their Brier Scores. Based on the difference in AUC scores, we conclude there is value in the inclusion of a baseline hazard rate, estimated by macro-economic variables, in survival models for single-family mortgage loan data. Based on the Brier Scores and the AUC scores we also conclude that overall predictive power of survival models for single-family mortgage loans is poor. This poor performance is particularly present for the loans in the performing state, the largest group of loans and the group with the largest possible impact. The performance of the survival models for firm bankruptcy does not appear to extend to survival models for single-family mortgage loans.

Further adjusting and extending of the basic survival model we use in this research paper can be of value to practitioners and academics. The value of the baseline hazard rate, estimated using macro-economic variables, provides support for its inclusion in survival models. Although the inclusion of a baseline did improve the forecasting performance, it still requires further improvement. Possible improvements on the model are estimating the baseline with monthly macro-economic data as opposed to the quarterly average used in this paper, or covariate selection unique for each initial state of the loans.

References

Abella´n Joaqui´n; Mantas, C. J. Improving Experimental Studies About Ensembles of Classifiers for Bankruptcy Prediction and Credit Scoring. Expert Systems with Applications 2014, 41 (8), 3825–3830.

Akkoç Soner. An Empirical Comparison of Conventional Techniques, Neural Networks and the Three Stage Hybrid Adaptive Neuro Fuzzy Inference System (anfis) Model for Credit Scoring Analysis: The Case of Turkish Credit Card Data. European Journal of Operational Research 2012, 222 (1), 168–168.

Baesens, T Van Gestel, S Viaene, M Stepanova, J Suykens J Vanthienen (2003) Benchmarking state-of-the-art classification algorithms for credit scoring, Journal of the Operational Research Society, 54:6, 627-635.

Beck N, Katz JN, Tucker R. 1998. Taking time seriously: time-series-cross-section analysis with a binary dependent variable. American Journal of Political Science 42: 1260–1288.

Bellotti, Anthony Crook, Jonathan. (2009). Credit scoring with macroeconomic variables using survival analysis. Journal of the Operational Research Society. 60. 1699-1707.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140.

Brown, J., McGourty, B. and Schuermann, T., 2015. Model risk and the great financial crisis: The rise of modern model risk management.

Bruneau, C., De Bandt, O. and El Amri, W. (2012), 'Macroeconomic fluctuations and corporate financial fragility', Journal of Financial Stability 8(4), 219–235.

Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). Classification and regression trees. Wadsworth Int. Group, 37(15), 237-251.

Jonathan N. Crook, David B. Edelman, Lyn C. Thomas, Recent developments in consumer credit risk assessment, European Journal of Operational Research, Volume 183, Issue 3, 2007, 1447-1465.

Cox DR. 1972. Regression models and life-tables. Journal of the Royal Statistical Society, Series B 34: 187–220.

Das, S.R., Freed, L., Geng, G., Kapadia, N. (2002). Correlated Default Risk. Santa Clara University Leavey School of Business Research Paper Series.

Davis, E. P. and Zhu, H. (2011), 'Bank lending and commercial property cycles: Some cross-country evidence', Journal of International Money and Finance 30(1), 1 – 21.

European Systemic Risk Board (2015), Report on commercial real estate and financial stability in the EU, Technical report, European System Risk Board, Frankfurt am Main.

Finlay, Steven. Multiple classifier architectures and their application to credit risk assessment, European Journal of Operational Research, Volume 210, Issue 2, 2011, Pages 368-378.

Grice, J. S. and Dugan, M. T. (2001), 'The limitations of bankruptcy prediction models: Some cautions for the researcher', Review of Quantitative Finance and Accounting 17(2), 151–166.

Hillegeist SA, Keating EK, Cram DP, Lundstedt KG. 2001. Corporate bankruptcy: do debt covenant and disclosure quality measures provide information beyond options and other market variables? Working paper, Kellogg Graduate School of Management.

Karels, G.V. and Prakash, A.J. (1987), Multivariate Normality and Forecasting of Business Bankruptcy. Journal of Business Finance Accounting, 14: 573-593.

Kline, Douglas and Victor L. Berardi. "Revisiting squared-error and cross-entropy functions for training neural network classifiers." Neural Computing Applications 14 (2005): 310-318.

Kumar, P. R.; Ravi, V. Bankruptcy Prediction in Banks and Firms Via Statistical and Intelligent Techniques - a Review. European Journal of Operational Research 2007, 180 (1), 1–1.

Lessmann, Stefan Baesens, Bart Seow, Hsin-Vonn Thomas, Lyn. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research.

Ma, Chao and Hongbiao Zhao. "Correlation in Mortgage Defaults." (2018).

A.I. Marqués, V. García, J.S. Sánchez, Exploring the behaviour of base classifiers in credit scoring ensembles, Expert Systems with Applications, Volume 39, Issue 11, 2012, Pages 10244-10250.

Mokas and Nijskens. (2019). Credit risk in commercial real estate bank loans: the role of idiosyncratic versus macro-economic factors. DNB Working Paper No.653.

Møller, M. F. (1990). A scaled conjugate gradient algorithm for fast supervised learning. Aarhus University, *Computer Science Department.*

Nam, Chae Kim, Tong Park, Nam Lee, Hoe. (2008). Bankruptcy prediction using a discrete-time survival model incorporating temporal and macroeconomic dependencies. Journal of Forecasting. 27. 493-506. 10.1002/for.985.

Neumann, Tobias, 2018. "Mortgages: estimating default correlation and forecasting default risk," Bank of England working papers 708, Bank of England.

Reichert A.K., Cho C-C, Wagner G.M., (1983) An examination of the conceptual issues involved in developing credit scoring models, J. Business and Economic Statistics 1, 101-114.

Sayari, N. and Mugan, C. S. (2017), 'Industry specific financial distress modeling', BRQ Business Research Quarterly 20(1), 45–62.

Schapire, R. (1989). On the strength of weak learnability. In Proceedings of the 30th IEEE Symposium on the Foundations of Computer Science (pp. 28-33).

Shumway T. (2001). Forecasting bankruptcy more accurately: a simple survival model. The Journal of Business 74: 101–124.

Thomas, L. C. (2000). A survey of credit and behavioral scoring: Forecasting financial risks of lending to customers. International Journal of Forecasting, 16(2), 149–172.

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

Tong, E. N. C.; Mues, C.; Thomas, L. C. Mixture Cure Models in Credit Scoring: If and When Borrowers Default. European Journal of Operational Research 2012, 218 (1), 132–132.

Wang, G. et al. (2011) "A Comparative Assessment of Ensemble Learning for Credit Scoring," 38(1), pp. 223–230.

West, D. (2000). Neural network credit scoring models. *Computers Operations Research*, 27(11-12), 1131-1152.

Yang, Y. Adaptive Credit Scoring with Kernel Learning Methods. European Journal of Operational Research 2007, 183 (3), 1521–1521.