



Time to default in credit scoring using survival analysis: a benchmark study

Lore Dirick^{1,2*}, Gerda Claeskens^{1,2} and Bart Baesens^{2,3,4}

¹ORSTAT, Faculty of Economics and Business, KU Leuven, Leuven, Belgium; ²Leuven Statistics Research Center (LSTAT), KU Leuven, Leuven, Belgium; ³LIRIS, Faculty of Economics and Business, KU Leuven, Leuven, Belgium; and ⁴School of Management, University of Southampton, Southampton, UK

We investigate the performance of various survival analysis techniques applied to ten actual credit data sets from Belgian and UK financial institutions. In the comparison we consider classical survival analysis techniques, namely the accelerated failure time models and Cox proportional hazards regression models, as well as Cox proportional hazards regression models with splines in the hazard function. Mixture cure models for single and multiple events were more recently introduced in the credit risk context. The performance of these models is evaluated using both a statistical evaluation and an economic approach through the use of annuity theory. It is found that spline-based methods and the single event mixture cure model perform well in the credit risk context.

Journal of the Operational Research Society (2017) **68**(6), 652–665. doi:10.1057/s41274-016-0128-9;

published online 1 December 2016

Keywords: benchmarking; credit risk modeling; competing risks; mixture cure model; survival analysis

1. Introduction

With the introduction of compliance guidelines such as Basel II and Basel III, and the resulting higher need for more accurate credit risk calculations, survival analysis gained more importance over the recent years. Historically, survival analysis is mainly used in the medical context as well as in engineering, where the time duration until an event is analyzed, for example the time until death or machine failure (see Kalbfleisch and Prentice, 2002; Collett, 2003; Cox and Oakes, 1984).

As an alternative to logistic regression, Narain (1992) first introduced the idea of using survival analysis in the credit risk context. The advantage of using survival analysis in this context is that the time to default can be modeled, and not just whether an applicant will default or not (Thomas *et al.*, 2002). Many authors followed the example of Narain (1992) and started to use more advanced methods as compared to the parametric accelerated failure time (AFT) survival methods used in this first work. An overview is given in Table 1. With its flexible nonparametric baseline hazard, the Cox proportional hazards (Cox PH) model was an obvious first alternative to the AFT model (Banasik *et al.*, 1999), and subsequent contributions extended both Cox PH and AFT models by using, among others, coarse classification (Stepanova and Thomas, 2002) and time-varying covariates (Bellotti and

Crook, 2009). In recent research some authors have experimented with mixture cure models. These models allow to model a “cured” fraction, a part of the population that will not go into default, in survival models.

In the existing literature, some questions remain. Firstly, except for Zhang and Thomas (2012), there has been no attempt to compare a wide range of the available methods in one paper. Secondly, in each of the papers listed in Table 1, only one data set was analyzed, not allowing to draw conclusions on which of the survival methods to use. Finally, in the majority of the papers, the evaluation remains largely focused on classification and the area under the receiver operating characteristics curve (AUC). In this paper, we contribute to the existing literature by analyzing ten different data sets from five banks, using all classes of models listed in Table 1, and using both statistical (AUC and default time predictions) and economic evaluation measures (by predicting the future value of the loan), applicable to all model types considered, the “plain” survival models as well as the mixture cure models.

Other interesting modeling approaches exist, though are not included in this comparison. These include discrete time hazard models such as used in Bellotti and Crook (2013, 2014) and Leow and Crook (2016). For further research it would be interesting to compare continuous time with discrete time models.

The remainder of this paper is organized as follows. Section 2 gives an overview of the survival analysis techniques used. In Section 3 the data and the experimental setup are discussed in more detail. The evaluation measures are

*Correspondence: Lore Dirick, Leuven Statistics Research Center (LSTAT), KU Leuven, Leuven, Belgium.
E-mail: Lore.Dirick@kuleuven.be

Table 1 Overview of the existing literature on the use of survival analysis in credit risk modeling

Paper	Parametric/ AFT	Cox PH	AFT/Cox PH + extensions	Nonparametric	Mixture cure	Multi- event mixture cure	Sample size	Number of inputs	Evaluation measure
Narain (1992)	X						1242	7	None
Banasik et al (1999)	X		X				50000	>7	Classification
Stepanova and Thomas (2001)			X X				11500	16	Classification, AUC, profit measure
Stepanova and Thomas (2002)			X X				50000	16	Classification, AUC
Bellotti and Crook (2009)			X				200000	>11	Cost of a bad case
Cao et al (2009)	X		X	X			25000	1	AUC
Tong et al (2012)			X		X		27527	14	AUC, H-measure, Kolmogorov–Smirnov
Zhang and Thomas (2012)	X		X X				27000	21	Error in default time prediction
Dirick et al (2015)					X	X	7521	8	AUC

The listed number of inputs is before variable selection (if applicable)

covered in Section 4, followed by the results and discussion in Sections 5 and 6.

2. Survival analysis methods

In survival analysis, one is interested in the timing, T , of a certain event. The survival function can be expressed as the probability of not having experienced the event of interest by some stated time t , hence $S(t) = P(T > t)$. In the context of credit risk, the event of interest is default (together with early repayment and maturity for the mixture cure model with multiple events, see Section 2.5). Given the survival function, the probability density function $f(u)$ is given by $f(u) = -\frac{d}{du}S(u)$. Additionally, the hazard function

$$h(t) = \lim_{\tau \rightarrow 0} \frac{P(t \leq T < t + \tau | T > t)}{\tau}$$

models the instantaneous risk. This function can also be expressed in terms of the survival function and the probability density function

$$h(t) = \frac{f(t)}{S(t)}.$$

In survival analysis, a certain proportion of the cases is censored, which means that for these cases, the event of interest has not yet been observed at the moment of data gathering. In this paper, we use two different definitions for censoring.

1. In the first definition, censored cases are the loans that did not reach their predefined end date at the moment of data gathering (called “mature” cases), and did not experience default nor early repayment by this time.
2. According to the second definition, a censored case corresponds to a loan that did not experience default by the moment of data gathering. Early repayment and mature cases are marked censored. This kind of censoring is used in models where default is the only event of interest.

One can interpret these two types of censoring as follows. The first definition defines censoring as we observe it when obtaining the data: some loan applicants defaulted, repaid early or some loans matured (which is, fully repaid at the end of the loan term). The remaining loans, where none of these events have (yet) been observed, are censored. According to the second definition, however, only two possible states are considered (instead of four): default and censoring. Here, all the cases that are labeled mature or early repayment according to the first definition get the label “censored.”

Hence, when applying survival analysis to model the time to default, the second definition is used (models in Sections 2.1–2.4). Only for the multiple event mixture cure models in Section 2.5, where competing event types are taken into account, the first definition is used. The censoring indicator for the i th case is denoted by δ_i , which is equal to 1 for an uncensored observation and is zero when censored.

When using survival models as regression models, a covariate vector and corresponding parameter vector are present. In all models in Section 2, the covariate vector is denoted by x , and the parameter vector by β .

2.1. Accelerated failure time models

Accelerated failure time (AFT) models are fully parametric survival models where explanatory variables act as acceleration factors to speed up or slow down the survival process as compared to the baseline survival function. Formally, this is denoted by

$$S(t | \mathbf{x}) = S_0(t \cdot \exp(-\beta' \mathbf{x}))$$

where the event rate is slowed down when $0 < \exp(-\beta' \mathbf{x}) < 1$ and speeded up when $\exp(-\beta' \mathbf{x}) > 1$. The hazard function is given by

$$h(t | \mathbf{x}) = h_0(t \cdot \exp(-\beta' \mathbf{x})) \exp(-\beta' \mathbf{x}).$$

In the general form, the accelerated failure time model can be expressed as a log-linear model for the timing of the event of interest $\log(T) = \beta' \mathbf{x} + \sigma \epsilon$, with ϵ a random error following some distribution and σ an additional parameter that rescales ϵ . As many classical survival distributions such as the Weibull distribution, exponential distribution and log-logistic distribution have event times that are log-linear, AFT models are often used as a starting point in order to parametrize these distributions. The three models mentioned above are used in the benchmark study and covered in Sections 2.1.1–2.1.3. For a full overview on AFT models and more technical details we refer to Collett (2003) and Kleinbaum and Klein (2011). AFT models are used in the credit risk context by Narain (1992) (who used an exponential distribution), Banasik *et al* (1999) (who used exponential and Weibull distributions) and Zhang and Thomas (2012) (who used Weibull, log-logistic and gamma distributions).

2.1.1. Weibull AFT model The Weibull model in its classical form can be expressed by the following survival and hazard function with scale λ and shape p

$$S(t) = \exp(-\lambda t^p), \quad h(t) = \lambda p t^{p-1}.$$

Using the relationship $\sigma = \frac{1}{p}$, it can be shown that a Weibull-distributed random event time $T_i = \exp(\beta' \mathbf{x}_i + \sigma \epsilon_i)$ corresponds to a survival function

$$S_i(t | \mathbf{x}) = \exp\left(-\lambda_i t^{1/\sigma}\right),$$

where $\lambda_i = \exp\left(-\frac{\beta' \mathbf{x}_i}{\sigma}\right)$ is the reparametrization used to incorporate the explanatory variables.

2.1.2. Exponential AFT model The exponential distribution is a special case of the Weibull distribution, with $p = 1$. This leads to a survival and hazard function

$$S(t) = \exp(-\lambda t), \quad h(t) = \lambda.$$

In the exponential distribution the strong assumption of a constant hazard rate λ is made, and for each case $\lambda_i = \exp(-\beta' \mathbf{x}_i)$. Note that ϵ for the exponential is not rescaled ($\sigma = \frac{1}{p} = 1$).

2.1.3. Log-logistic AFT model The log-logistic distribution with parameters θ and κ has a survival and hazard function

$$S(t) = \frac{1}{1 + \exp(\theta)t^\kappa}, \quad h(t) = \frac{\exp(\theta)\kappa t^{\kappa-1}}{1 + \exp(\theta)t^\kappa}.$$

Using the AFT reparametrization, the relationship $\sigma = \frac{1}{\kappa}$ and the log-logistically distributed event time T_i has a survival function

$$S_i(t | \mathbf{x}_i) = \frac{1}{1 + \exp(\theta_i)t^{1/\sigma}}$$

$$\text{where } \theta_i = -\frac{\beta' \mathbf{x}_i}{\sigma}.$$

2.2. Cox proportional hazards model

Another method which is commonly used in survival analysis is the Cox proportional hazards model (Cox, 1972). This method is more flexible than any AFT model as it contains a nonparametric baseline hazard function, $h_0(t)$, along with a parametric part. In this model, the hazard function is given by

$$h(t | \mathbf{x}) = h_0(t) \exp(\beta' \mathbf{x}), \quad (1)$$

and the survival function is

$$\begin{aligned} S(t | \mathbf{x}) &= \exp\left(-\exp(\beta' \mathbf{x}) \int_0^t h_0(u) du\right) \\ &= \exp(-\exp(\beta' \mathbf{x}) H_0(t)), \end{aligned}$$

with $H_0(t)$ the cumulative baseline hazard function. In this paper, Breslow's method is used to estimate the cumulative baseline hazard rate, given by

$$\widehat{H}_0(t) = \sum_{t_i \leq t} \frac{1}{\sum_{r \in R(t_i)} \exp(\beta' \mathbf{x}_r)},$$

where $R(t_i)$ denotes the group of individuals at risk at time t_i (which are, in the credit risk context, the ones that have not yet defaulted by time t_i). For more information for the Breslow and other estimators for the Cox PH model, we refer to Klein and Moeschberger (2003). The Cox PH model was first used in the credit context by Banasik *et al* (1999).

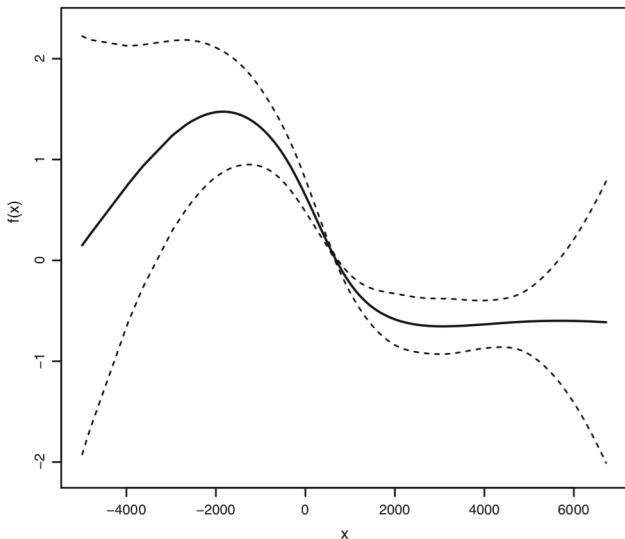


Figure 1 Functional form for one of the covariates x , describing the relationship between x and spline approximation $f(x)$ using penalized splines. x is a variable in one of the ten data sets (more details are not disclosed due to confidentiality reasons). The pointwise 95% confidence bands are given by the dotted lines.

2.3. Cox proportional hazards model with splines

The hazard function in the Cox PH model, see formula (1), assumes a proportional hazards structure with a log-linear model for the covariates. As a result, for any continuous variable, e.g., age, the default hazard ratio between a 25- and a 30-year-olds is the same as the hazard ratio between an 70- and 75-year-olds. As it is likely that this assumption does not hold, one has been looking for other functional forms of covariates; for an overview, see Therneau and Grambsch (2000). One of the most popular methods to deal with this is by using splines. Splines are flexible functions defined by piecewise polynomials that are joined in points called “knots.” Some constraints are imposed to ensure that the overall curve is smooth. Any continuous variable can be represented by a spline, hence where in formula (1) the linear predictor is denoted by

$$\beta' \mathbf{x} = \sum_{j=1}^m \beta_j x_j.$$

Splines can be introduced modeling some or all, say these are $m-l$, continuous covariates by a spline approximation $f_j(\mathbf{x}_j)$,

$$\beta' \mathbf{x} = \sum_{j=1}^l \beta_j x_j + \sum_{j=l+1}^m f_j(\mathbf{x}_j).$$

For an example, see Figure 1. To get a smooth function, a basis of functions with continuous first derivatives is often used to construct a spline function. A popular spline basis is the basis of cubic spline functions

$$1, x, x^2, x^3, (x - \kappa_1)_+^3, \dots, (x - \kappa_q)_+^3$$

with q knots $\kappa_1, \dots, \kappa_q$. A spline model is formed by taking a linear combination of the spline basis functions. The disadvantage of power bases, however, lies in the fact that they can become numerically unstable when a large number of knots are included. For this reason, an equivalent basis with more stable numerical properties, the B-spline basis (de Boor, 2001), is nowadays widely used. Both spline models in this study use a cubic B-spline basis in the Cox PH model. For an overview on splines in a general framework, we refer to Ruppert et al (2003).

2.3.1. Natural splines A commonly used modification of the cubic B-spline basis is the natural cubic spline basis. Natural cubic splines satisfy the additional constraint that they are linear in their tails beyond the boundary knots, which are taken to be the endpoints of the data.

2.3.2. Penalized splines As the number of knots in a spline becomes relatively large, a fitted spline function will show more variation than justified by the data. To limit overfitting, O’Sullivan (1986) introduced a smoothness penalty by integrating the square of the second derivative of the fitted spline function. Later, Eilers et al (1996) showed that this penalty could also be based on higher-order finite differences of adjacent B-splines. Penalized splines or “P-splines” use the latter method to estimate spline functions.

2.4. Mixture cure model

In the medical context, mixture cure models were motivated by the existence of a subgroup of long-term survivors, or a “cured” fraction (see Sy and Taylor, 2000; Peng and Dear, 2000). In contrast to non-mixture survival models, where the event of interest is assumed to take place in the long run, these types of models are typically used in contexts where a certain proportion of the population will not experience the event. Looking at the credit data, it becomes indeed apparent that a very large proportion of the population will not go into default. “Cure” (or “non-susceptibility”) can here be interpreted as a situation where an individual is not expected to experience default. The mixture cure model is then in fact a mixture distribution where, on one hand, a logistic regression model provides a mixing proportion of the “non-susceptible” cases. On the other hand, a survival model describes the “survival behavior” of the cases susceptible to the event of interest.

This type of models is of particular interest in credit risk modeling as the event of interest here, default, will not occur for a very high proportion of the cases. This idea was introduced in the credit risk context for the first time by Tong et al (2012). In Dirick et al (2015), a model selection criterion adapted to these models was introduced and applied to credit

risk data. For the mixture cure model, the unconditional survival function is given by:

$$S(t | \mathbf{x}) = \pi(\mathbf{x})S(t | Y = 1, \mathbf{x}) + 1 - \pi(\mathbf{x}), \quad (2)$$

where Y is the susceptibility indicator ($Y = 1$ if an account is susceptible, and $Y = 0$ if not). Note that a new covariate vector \mathbf{x} is introduced, which is the covariate vector of the logistic regression model, in this case the binomial logit,

$$\pi(\mathbf{x}) = P(Y = 1 | \mathbf{x}) = \frac{\exp(b' \mathbf{x})}{1 + \exp(b' \mathbf{x})},$$

with corresponding parameter vector b . In this paper, the conditional survival function modeling the cases that are susceptible is given by a Cox proportional hazards model,

$$S(t | Y = 1, \mathbf{x}) = \exp\left(-\exp(b' \mathbf{x}) \int_0^t h_0(u | Y = 1) du\right).$$

As in the Cox proportional hazards model in a non-mixture context, the Breslow-type estimator is used for estimation of the cumulative baseline hazard. Figure 2 shows the difference between the survival curves for plain survival functions (such as non-mixture Cox PH and AFT functions) compared to the unconditional survival functions of the mixture cure model. Whereas plain survival curves go to zero as the time goes to infinity, the unconditional survival curves for the mixture cure model “plateau” at a positive value ($1 - \pi(\mathbf{x})$).

The mixture cure model is computationally more intensive than plain survival models, as the use of an iterative procedure,

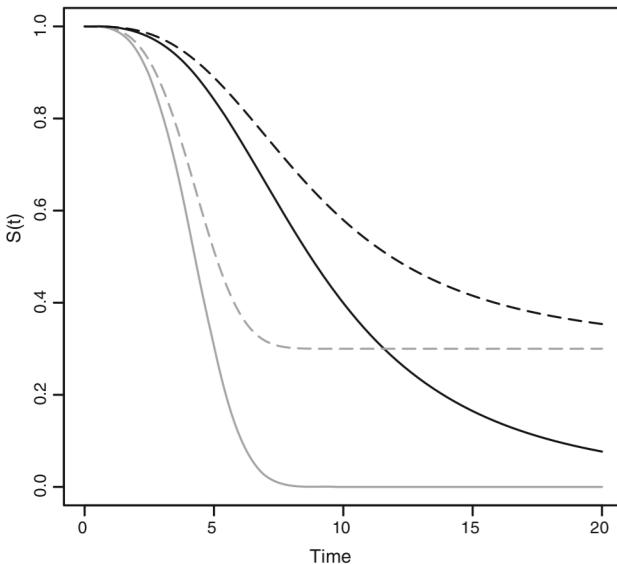


Figure 2 A graphical example pointing out the difference between two plain survival curves and the unconditional survival curve in a mixture cure model. *Full lines* are plain survival curves (modeled using a Weibull AFT model for the *gray curve*, and a log-logistic AFT model for the *black curve*), and *dotted lines* represent their corresponding unconditional survival curves in a mixture cure model when assuming a cure rate of 30%.

the expectation maximization (EM)-algorithm, is needed in order to overcome incomplete information on Y . For more information on mixture cure models, we refer to Farewell (1982), Tong *et al* (2012) and Dirick *et al* (2015).

2.5. Mixture cure model with multiple events

In the medical context it is unusual to ever truly observe cure. In cancer research, for example, a subject might pass away from the specific cancer under research immediately after the observation period, even though having a high probability of being cured. Observed cure does exist in the credit risk context, since as a loan reaches maturity, it is known that default can not occur anymore. As the censoring indicator in the mixture cure model only provides information on whether default took place or not, information on maturity is not used in the model. Another shortcoming is the fact that it does not account for an important “competing risk,” early repayment, where a lender repays the loan before the predetermined end date.

Watkins *et al* (2014) recently proposed a method that provides simultaneous modeling of multiple events, along with a mature group. Dirick *et al* (2015) extended this model by allowing for the semi-parametric Cox proportional hazards to model the survival times, instead of the parametric survival models proposed by the former authors. Applied to the credit risk example, three indicators are introduced:

1. Y_m , indicating that the loan is considered to be mature, so repaid at the indicated end date of the loan;
2. Y_d , indicating that default takes place;
3. Y_e , indicating that early repayment takes place.

Note that an important limitation here is that Y_m is only defined for fixed term loans. As a result, the multiple event mixture cure model in this form is not usable for applications on revolving credit data sets. For the fixed end term data sets used in this paper, the set of (Y_m, Y_d, Y_e) is exhaustive and mutually exclusive. However, when an observation is censored (according to the first definition in Section 2), it is not known which event type will occur. In analogy to Equation (2), the unconditional survival function can be written as

$$S(t | \mathbf{x}) = \pi_e(\mathbf{x})S_e(t | Y_e = 1, \mathbf{x}) + \pi_d(\mathbf{x})S_d(t | Y_d = 1, \mathbf{x}) + (1 - \pi_e(\mathbf{x}) - \pi_d(\mathbf{x})),$$

with $S_e(t | Y_e = 1, \mathbf{x})$ and $S_d(t | Y_d = 1, \mathbf{x})$ denoting the conditional survival functions for, respectively, early repayment and default, which are modeled using a Cox proportional hazards model, as in Equation (2). The $\pi_j(\mathbf{x})$'s with $j \in \{e, d\}$ are modeled using a multinomial logit model, hence:

$$\pi_d(\mathbf{x}) = P(Y_d = 1 | \mathbf{x}) = \frac{\exp(b_d' \mathbf{x})}{1 + \exp(b_d' \mathbf{x}) + \exp(b_e' \mathbf{x})}, \quad (3)$$

$\pi_e(\mathbf{x})$ is found analogously.

Table 2 Bank data sets

Bank	Type of loans	Country
Bank A	Personal loan data	Belgium
Bank B	Personal loan data	UK
Bank C	Personal loan data and SME	Belgium
Bank D	Personal loan data and SME	Belgium
Bank E	Personal loan data	Belgium

3. The data and experimental setup

3.1. Data preprocessing and missing inputs

We received data sets from five financial institutions in the UK and Belgium, consisting of mainly personal loans and loans of small enterprises, with varying loan terms (for details, see Table 2). Note that this resulted in data sets with either only personal loans or data sets with a mix of personal and small enterprise loans, for banks C and D. As the SMEs (small and medium-sized enterprises) in our data sets were all sole proprietorships, their properties were nearly identical to those of personal loans. More information on the use of survival models in SMEs in the broader sense can be found in, among others (Fantazzini and Figini, 2009; Gupta et al., 2015; Holmes et al., 2010).

For the banks with data covering several loan terms, the data were split in order to get only one loan term per data set, resulting in ten data sets. Table 3 lists these data sets which were used to evaluate the different survival techniques listed in Section 2. Except for bank C, where default is defined as missing two consecutive payments, all banks defined default as missing three consecutive months of payments.

As survival analysis techniques are unable to cope with missing data, and with several data sets having a considerable amount of missing inputs, some preprocessing mechanism to cope with missing data is needed. We want to stress that there are many ways of doing this. As this benchmark paper aims to focus on different models, however, rather than data preprocessing (which is typical for benchmarking studies (see Baesens et al., 2003; Dejaeger et al., 2012; Loterman et al., 2012, among others)), we chose to employ the rule of thumb

also used in the benchmarking paper by Dejaeger et al (2012). Therefore, for continuous inputs, median imputation is used when $\leq 25\%$ of the values are missing, and the inputs are removed if more than 25% is missing. For categorical inputs, a missing value category is created if more than 15% of the values is missing, otherwise the observations associated with the missing values are removed from the data set.

The number of input variables in the resulting data sets varies from 6 to 31, and the number of observations from 7521 to 80,641. For each observation, an indicator for default, early repayment and maturity is included, taking the value of 1 for the respective event of interest that took place, and 0 for the others (note that only one event type can occur for each observation). Percentages of occurrences of these three event types per data set are given in Table 3. For censored observations according to the first censoring definition, all indicators are 0. According to the second censoring definition, only defaults are considered uncensored. In terms of our data sets, this means that censoring rates are ranging from around 20 to 85% according to the first definition (used for the multiple event mixture cure model), whereas censoring percentages are not lower than 94.56% up to 98.16% according to the second definition.

Additionally, a time variable is included for each observation, representing the respective month of the event, which takes an integer value. Note that the time variable for a mature event is always equal to the length of the loan term (e.g., a matured loan for data set 5 has value 24), and the time variable for a censored event is given by the last observed month in which a repayment was observed to take place.

3.2. Experimental setup

Each data set was randomly split up in a training set and a test set consisting of 2/3 and 1/3 of the observations, respectively. The models are estimated on the training sets, and the corresponding test sets are used for evaluation.

For all models, the software R is used. AFT and Cox proportional hazards modeling is possible through the use of the R-package *survival* (Therneau, 2015), with additional use of

Table 3 Data set specifications

Data set	Bank	Loan term (months)	Data size	Inputs (number)	Cat. (number)	Cont. (number)	Default (%)	Early (%)	Mature (%)
Data set 1	Bank A	36	42903	31	13	18	4.03	26.80	49.72
Data set 2	Bank A	48	46970	31	13	18	4.02	28.61	40.49
Data set 3	Bank A	60	80641	31	13	18	5.44	32.74	24.80
Data set 4	Bank B	12	10027	13	7	6	2.73	53.80	24.43
Data set 5	Bank B	24	9979	13	7	6	4.74	38.46	28.88
Data set 6	Bank B	36	7521	13	7	6	5.00	39.78	3.58
Data set 7	Bank C	48	9980	6	5	1	1.84	9.20	19.80
Data set 8	Bank C	60	17378	6	5	1	1.84	8.95	4.13
Data set 9	Bank D	37	35856	11	8	3	3.56	19.27	46.83
Data set 10	Bank E	60	9785	8	4	4	1.62	10.09	17.85

functions `ns` and `pspline` for inclusion of natural splines and penalized splines in the covariates, respectively. An ad hoc method was used to decide on which of the continuous variables a spline function should be introduced. Using the `pspline`-function on each continuous variable in the model separately, the resulting spline curves were inspected to track some possible nonlinear relationships, with knots determined by the adapted AIC method (Eilers *et al.*, 1996, included in the package). The resulting Cox proportional hazards `pspline` models consisted of all the P-splines where nonlinear relationships were observed. As the `ns`-function does not have a built-in function to optimize the number of knots, the same continuous variables and number of knots were chosen as in the `pspline` models. For some of the data sets, the number of splines or knots using the natural splines was altered in comparison with the `pspline` models, in order to get a feasible fit.

For the mixture cure model, the R-package `smcure` by Cai *et al.* (2012a, b) is used. An extended code based on this package, as in Dirick *et al.* (2015), is used for the multiple event mixture cure model (code available upon request).

4. Performance criteria/evaluation metrics

Three main performance criteria were used.

4.1. AUC in ROC curves

In the credit risk context, an ubiquitous method to evaluate binary classifiers is by means of the receiver operating characteristics curve. This curve illustrates the performance of a binary classifier for each possible threshold value, by plotting the true positive rate against the false positive rate. The specific measure of interest is the area under the curve (AUC), which can also be computed in the context of survival analysis. In this context, evaluation is possible at any timepoint of the survival curve (see Heagerty and Saha, 2000). For each data set and each model, the AUC for the test sets at 1/3 and 2/3 of the time to maturity and at the maturity time itself (which is equal to the loan term) is listed in Table 4.

Despite the fact that AUC and other classification-based evaluation methods are most common in the literature (see Table 1), this way of evaluating survival analysis in credit risk does not fully highlight the benefits of using survival analysis in this context. First of all, the AUC does not fully summarize the time aspects of survival analysis (the AUC is calculated at one specific timepoint), and secondly, the financial aspect is neglected. The next two sections focus on the timing aspects and economic/financial evaluation, respectively.

4.2. Evaluation through default time prediction

When evaluating through default time prediction, we look at how we are able to predict the default times of the defaults in

the test set. A survival curve does not give one time estimate, but a distribution of time estimates. With a high amount of censoring, mean values of these survival analyses do not give good predictors. Zhang and Thomas (2012) compute a predictor for the recovery rate in survival analysis by looking at each percentile of the training set and calculate the squared and absolute deviations from the predictions to the observed values of the default cases. Next, the percentiles resulting in the lowest deviations are withheld and used to compute the deviations in the test set.

We use the same method as Zhang and Thomas (2012), but consider the default time instead of recovery values, and look at each permille. For each data set, the permilles that result in smallest deviations for the training sets are withheld and used to compute default time predictions in the test sets. The results are listed in Table 5, where the MSE columns list the mean of the squared differences between the predicted and observed default times, and the MAE columns list the mean of the absolute differences between the predicted and observed default times.

Note that the part of the data set which is evaluated is considerably smaller here compared to the AUC method. A schematic representation is given in Figure 3. Each letter represents an observation in the entire data set, where four possible end states are possible: early repayment ("E"), default ("D"), maturity ("M") and censored ("C"). The green circle encompasses the test set elements, which are all evaluated when computing the AUC. The default time prediction method, however, only evaluates the default times of the "actual" defaults in the test set, which are in the red circle. As the evaluation set differs from one method to another, the sample size for each of the resulting test sets is included in result (Tables 4, 5, 6, 7).

4.3. Evaluation using annuity theory

When banks grant a loan to a customer, they are particularly interested in the expected future value at the end of the loan term. One can use the principles of annuity theory (for an overview, see Kellison and Irwin 1991) to compute this value, though these basic principles do not incorporate risk; hence, these formulas start from the assumption that loans will be repaid with a 100% certainty. Including this risk aspect is exactly what can be done using survival analysis, as it provides us with an accurate estimate for the probability that a customer is still repaying his loan at every time instant of the survival curve.

In this study, we computed the true future value of the uncensored test set loans (given by the observations in the blue circle in Figure 3), taking into account their true end-state (default, early repayment or maturity) and compare them to their estimated values using each of the survival models. In order to make the results comparable, some assumptions are made and applied when evaluating the models for all data sets:

Table 4 Test set “areas under the curve” (AUC) for the different methods applied to the ten data sets when evaluating at several timepoints, corresponding to 1/3, 2/3 and the full loan term, which depends on the data set

Method/AUC	Sample size	Data set 1			Data set 2			Data set 3			Data set 4			Data set 5		
		36 months			48 months			60 months			12 months			24 months		
		14,301			15,656			26,880			3342			3326		
AFT Weibull	0.829	0.826	0.847	0.831	0.835	0.827	0.812	0.809	0.816	0.846	0.804	0.779	0.704	0.711	0.716	
AFT exponential	0.829	0.826	0.847	0.831	0.834	0.826	0.812	0.809	0.815	0.845	0.804	0.779	0.703	0.711	0.716	
AFT log-logistic	0.829	0.826	0.847	0.831	0.835	0.827	0.812	0.809	0.816	0.846	0.804	0.779	0.704	0.711	0.716	
Cox PH	0.828	0.824	0.846	0.830	0.834	0.826	0.812	0.809	0.815	0.846	0.804	0.780	0.704	0.711	0.716	
Cox PH nat. splines	0.852	0.837	0.854	0.851	0.835	0.829	0.821	0.810	0.775	0.775	0.676	0.695	0.698			
Cox PH penal. splines	0.832	0.827	0.849	0.838	0.838	0.830	0.820	0.817	0.821	0.834	0.820	0.789	0.684	0.691	0.698	
Mixt. cure	0.829	0.828	0.847	0.827	0.833	0.825	0.817	0.814	0.823	0.875	0.833	0.787	0.702	0.705	0.699	
Multi-event mixt. cure	0.829	0.824	0.844	0.832	0.835	0.825	0.811	0.806	0.816	0.821	0.801	0.779	0.715	0.716	0.715	
Method/AUC	Sample size	Data set 6			Data set 7			Data set 8			Data set 9			Data set 10		
		36 months			48 months			60 months			37 months			60 months		
		2507			3326			5792			11,952			3261		
AFT Weibull	0.736	0.706	0.664	0.749	0.715	0.654	0.596	0.668	0.662	0.852	0.850	0.849	0.711	0.751	0.766	
AFT exponential	0.736	0.706	0.664	0.745	0.714	0.653	0.598	0.667	0.659	0.852	0.849	0.849	0.710	0.750	0.764	
AFT log-logistic	0.736	0.706	0.663	0.751	0.715	0.654	0.599	0.669	0.663	0.852	0.850	0.849	0.712	0.752	0.766	
Cox PH	0.736	0.705	0.663	0.733	0.710	0.652	0.596	0.668	0.663	0.852	0.850	0.849	0.712	0.750	0.765	
Cox PH nat. splines	0.698	0.657	0.731	0.688	0.610	0.582	0.664	0.649	0.859	0.856	0.855	0.739	0.778	0.797		
Cox PH penal. splines	0.719	0.711	0.649	0.732	0.710	0.651	0.602	0.670	0.664	0.859	0.858	0.857	0.723	0.782	0.791	
Mixt. cure	0.727	0.702	0.674	0.654	0.657	0.623	0.630	0.682	0.643	0.854	0.852	0.850	0.699	0.739	0.748	
Multi-event mixt. cure	0.729	0.704	0.665	0.723	0.706	0.654	0.630	0.678	0.640	0.850	0.849	0.849	0.695	0.750	0.760	

The three best values are underlined. AUCs at the three timepoints are comparable within one data set, so columnwise

Table 5 Deviation measures when predicting the default times for observed defaults in the test set of the ten data sets, using different methods

	Data set 1		Data set 2		Data set 3		Data set 4		Data set 5	
Loan term	36 months		48 months		60 months		12 months		24 months	
sample size	567		626		1507		95		166	
Method/deviation measure	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
AFT Weibull	333.43	13.51	662.40	18.39	612.89	18.16	121.34	4.29	79.48	6.75
AFT exponential	434.42	15.94	762.30	20.67	677.47	19.40	4233.42	20.17	127.89	8.46
AFT log-logistic	344.53	13.55	678.44	18.36	613.16	18.22	123.08	4.32	79.52	6.75
Cox PH	229.08	11.91	412.77	15.62	510.87	16.93	12.59	2.97	63.78	6.34
Cox PH with natural splines	235.98	12.03	415.69	15.71	532.30	17.41	17.02	3.25	71.60	6.83
Cox PH with penalized splines	235.61	12.14	412.00	15.69	512.63	17.06	14.44	3.12	70.90	6.80
Mixture cure	233.33	11.99	412.21	15.68	527.01	17.22	12.55	2.86	68.54	6.63
Multiple event mixture cure	262.81	12.65	475.08	16.72	564.77	17.70	13.47	3.11	68.11	6.54
	Data set 6		Data set 7		Data set 8		Data set 9		Data set 10	
Loan term	36 months		48 months		60 months		37 months		60 months	
sample size	116		57		104		428		59	
Method/deviation measure	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
AFT Weibull	88.29	6.87	281.89	12.50	434.65	16.52	385.94	13.18	490.57	18.37
AFT exponential	136.93	8.14	444.42	17.81	768.92	20.76	487.06	16.32	837.56	24.31
AFT log-logistic	90.79	6.89	263.62	12.53	436.83	16.55	382.82	13.13	494.64	18.44
Cox PH	89.09	7.15	245.18	11.25	414.76	16.34	185.27	10.37	505.24	18.31
Cox PH with natural splines	111.78	7.54	285.44	12.75	415.78	16.62	188.13	10.40	521.93	18.20
Cox PH with penalized splines	92.15	7.13	237.89	11.32	413.55	16.25	189.38	10.42	491.15	17.56
Mixture cure	92.28	7.24	359.54	15.23	410.42	16.37	186.75	10.62	511.81	18.00
Multiple event mixture cure	90.17	7.74	278.32	12.19	418.88	16.48	210.38	11.06	524.63	18.86

Top performances for each test set are underlined. Performances that are significantly different at a 5% level from the top performance with respect to a one-sided Mann-Whitney test are denoted in boldface

1. Loans are repaid at the end of each month, with a fixed sum;
2. The (yearly) interest rate i_y used is 5%;
3. The loans are treated as if they all started at the same point in time, in order to make them comparable.

Let us introduce:

- (a) L_s the initial amount of the loan, or the debt of subject s ;
- (b) R_s the constant monthly repayment for subject s ;
- (c) n the number of periods;
- (d) i the monthly interest rate ($i = (1 + i_y)^{1/12} - 1$);
- (e) (E)FV the (expected) future value of a loan.

A bank can reinvest the repayment sums R_s as soon as they are paid by the client. Assume that the same interest rate applies. If there is no risk for default nor early repayment, the future value can be given by

$$\begin{aligned} FV_s &= R_s \left((1+i)^{n-1} + (1+i)^{n-2} + \cdots + (1+i)^0 \right) \\ &= R_s \frac{(1+i)^n - 1}{i}. \end{aligned} \quad (4)$$

For the uncensored test set loans, we wish to estimate the future loan values. In Section 4.3.1, we describe how we compute the true future values when knowing the eventual state ("D," "M" or "E"), and in Sections 4.3.2–4.3.3, the expected future loan value is estimated when using the model predictions. Table 6 lists the mean absolute differences between the observed future values and the expected future loan values using the model estimations. In Table 7, we consider the mean expected future values per loan and compare them with the mean observed future value.

4.3.1. The true future loan values The true future loan value depends on the eventual loan outcome or state. For mature loans, Equation (4) can be used with n the total number of periods or the loan term. Hence,

$$FV_{s \in \text{mature}} = R_s \frac{(1+i)^n - 1}{i}.$$

For the future value of a loan with early repayment, the resulting amount of the debt in any time period k is given by

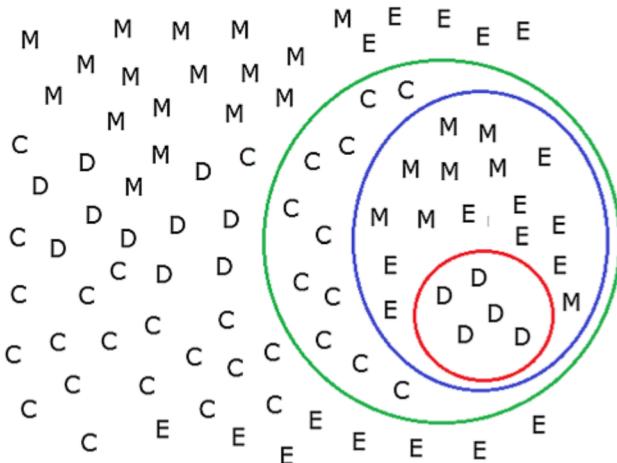


Figure 3 Schematic representation of the data set. Each letter represents an observation in the data set. The data set elements that are in the test set are in the largest (green) circle. All test set elements are evaluated using the AUC evaluation method. The uncensored test set elements (according to the first definition of censoring, see Section 2) that are in the middle (blue) circle are evaluated through the economic evaluation method using annuity theory. Default time prediction evaluation can only be performed on the defaulted elements of the test set, encompassed by the smallest (red) circle (Color figure online).

$$L_{s,k} = \left(1 - \frac{(1+i)^k - 1}{(1+i)^n - 1}\right) L_s.$$

When an early repayment takes place in period k , we assume that the loan is repaid as usual until this period k and that the sum L_k is fully being repaid in this period. This sum can still be reinvested for $n - k - 1$ periods. Note that early repayment always yields a smaller revenue compared to a matured loan,

$$FV_{s\text{early}} = R_s \left(\sum_{j=1}^k (1+i)^{n-j} \right) + L_{s,k} (1+i)^{n-k-1}.$$

Table 6 Analyzing model performance using financial metrics

MAD from FV	DS 1	DS 2	DS 3	DS 4	DS 5	DS 6	DS 7	DS 8	DS 9	DS 10
Loan term (months)	36	48	60	12	24	36	48	60	37	60
Sample size	11517	11453	16901	2705	2424	1249	985	824	8339	947
AFT Weibull	334.1	748.8	1695.7	29.3	133.2	384.0	669.0	1522.4	420.1	887.6
AFT exponential	350.4	769.2	1709.6	32.6	139.1	386.7	687.2	1524.3	446.7	917.1
AFT log-logistic	334.5	750.6	1699.7	29.4	133.4	384.9	670.1	1524.2	424.7	889.3
Cox PH	332.0	744.4	1693.1	30.2	135.9	383.2	667.1	1519.7	428.0	887.5
Cox PH nat. splines	335.2	756.9	1716.1	30.0	136.2	388.5	661.5	1556.7	409.1	885.4
Cox PH penal. splines	332.0	740.3	1673.7	29.9	134.5	387.6	659.3	1519.7	409.1	888.6
Mixt. cure	330.8	746.2	1692.8	29.5	135.7	382.0	672.4	1526.4	411.6	906.3
Multi-event mixt. cure	409.9	940.9	2109.5	31.5	148.9	435.3	744.5	1691.1	467.4	1054.6

Mean absolute deviations from the observed future loan values for the uncensored cases (first definition) of the test set. Top performances for each test set are underlined. Performances that are significantly different at a 5% level from the top performance with respect to a one-sided Mann–Whitney test are denoted in boldface

The future value for a loan where default takes place after k months is equal to

$$FV_{s\in\text{default}} = R_s \left(\sum_{j=1}^k (1+i)^{n-j} \right),$$

hence, we assume that when default takes place, nothing of the remaining sum L_k is recovered.

4.3.2. The expected future loan values using non-mixture survival models In each of the survival models in Sections 2.1–2.3, the model provides us with a survival probability estimate at each point in time. We denote $\widehat{S}(t)_{s,m}^d$ the estimated probability that subject s has not defaulted by time t , using model m . Then we can calculate the expected terminal value of a loan according to a certain model m as follows:

$$EFV_{s,m} = R_s \left(\sum_{j=1}^n \widehat{S}(j)_{s,m}^d (1+i)^{n-j} \right).$$

4.3.3. The expected future loan values using mixture cure models When computing the expected future loan values by means of the mixture cure models, we need to take the results of the binomial (single event) and multinomial logit (multiple event) part of the model into account. For the mixture cure model in Section 2.4, we have probabilities of being susceptible to default or not (PD and 1 – PD) for every subject. We define

$$PD_s = \hat{\pi}(\mathbf{x}_s) = \frac{\exp(\hat{b}'\mathbf{x}_s)}{1 + \exp(\hat{b}'\mathbf{x}_s)}, \quad (5)$$

then we have

$$\begin{aligned} EFV_{s,m} &= PD_s \cdot R_s \left(\sum_{j=1}^n \widehat{S}(j)_{s,m}^d (1+i)^{n-j} \right) \\ &\quad + (1 - PD_s) \cdot R_s \frac{(1+i)^n - 1}{i}, \end{aligned} \quad (6)$$

Table 7 Analyzing model performance using financial metrics

Mean EFV per loan	<i>DS 1</i>	<i>DS 2</i>	<i>DS 3</i>	<i>DS 4</i>	<i>DS 5</i>	<i>DS 6</i>	<i>DS 7</i>	<i>DS 8</i>	<i>DS 9</i>	<i>DS 10</i>
Loan term (months)	36	48	60	12	24	36	48	60	37	60
Sample size	11517	11453	16901	2705	2424	1249	985	824	8339	947
AFT Weibull	8180.1	14,213.5	19,585.4	1020.8	2119.1	4067.2	15,793.6	21,021.2	9686.8	21,972.0
AFT exponential	8159.9	14,186.5	19,566.7	1016.9	2111.9	4063.7	15,772.6	21,017.0	9655.7	21,935.4
AFT log-logistic	8179.5	14,210.9	19,580.4	1020.7	2118.8	4065.9	15,792.2	21,018.5	9681.2	21,969.9
Cox PH	8183.9	14,222.2	19,596.5	1019.8	2115.7	4068.6	15,796.6	21,025.3	9677.8	21,971.9
Cox PH nat. splines	8175.6	14,182.3	19,515.8	1019.7	2115.7	4062.8	15,805.7	20,990.2	9695.9	21,967.7
Cox PH penal. splines	8182.2	14,214.9	19,590.6	1019.8	2117.7	4062.5	15,804.7	21,025.5	9697.0	21,966.1
Mixt. cure	8099.1	14,004.2	19,120.2	1018.2	2101.3	4007.3	15,708.1	20,803.8	9633.0	21,779.4
Multi-event mixt. cure	8185.0	14,217.4	19,588.4	1020.3	2115.8	4070.3	15,794.5	21,011.6	9693.7	21,951.2
Mean FV per loan	8164.3	14,173.2	19,339.2	1012.2	2096.5	3966.0	15,590.8	20,137.1	9649.7	21,563.7

Mean expected future loan values of the uncensored cases (first definition) of the test set. The three best values are underlined

where $\widehat{S}(t)_{s,m}^d = \widehat{S}(t | Y = 1)_{s,m}$, denoting the conditional aspect of the survival estimates in the mixture cure context as in (2).

For the multiple event mixture cure model in Section 2.5, the multinomial logit (expression 3) leads in a similar way as (5) to probabilities of early repayment PE and probabilities of maturity PM (which is, in fact, $1 - PD - PE$). Here, $\widehat{S}(t)_{s,m}^e = \widehat{S}(t | Y_e = 1)_{s,m}$ and $\widehat{S}(t)_{s,m}^d = \widehat{S}(t | Y_d = 1)_{s,m}$ are again conditional probabilities (given Y_d and Y_e) that subject s has not repaid early or defaulted, respectively, by time t . The expected future value is given by

$$\begin{aligned} EFV_{s,m} = & PD_s \cdot \mathbf{R}_s \left(\sum_{j=1}^n \widehat{S}(j)_{s,m}^d (1+i)^{n-j} \right) \\ & + PM_s \cdot \mathbf{R}_s \frac{(1+i)^n - 1}{i} \\ & + PE_s \cdot \left(\mathbf{R}_s \left(\sum_{j=1}^n \widehat{S}(j)_{s,m}^e (1+i)^{n-j} \right) \right. \\ & \left. + \sum_{j=1}^{n-1} \left((\widehat{S}(j-1)_{s,m}^e - \widehat{S}(j)_{s,m}^e) \mathbf{L}_{sj} (1+i)^{n-j-1} \right) \right). \end{aligned} \quad (7)$$

The first two lines of (7) are completely identical to (6), where $(1 - PD)$ is replaced by PM (or, in other words, $1 - PD - PE$, as early repayment is also considered here). The second part of the expression is dominated by the event of early repayment. Early repayment works in a similar way as default, in the sense that repayment of the fixed sum \mathbf{R}_s occurs each month with a probability $\widehat{S}(j)_{s,m}^e$, which explains the first term in the second line of (7). The main difference with default, however, is that when early repayment occurs at timepoint j (this happens with a probability $\widehat{S}(j-1)_{s,m}^e - \widehat{S}(j)_{s,m}^e$), the bank receives \mathbf{L}_{sj} , the resulting amount of the outstanding debt at timepoint j . This idea is displayed in the last term of (7).

Note that this expression assumes that the penalty term for early repayment is equal to zero, where in reality usually a

fixed fee needs to be payed [see Ma *et al* (2010), where the fee is 2 months of interest on the outstanding debt]. The reason for this assumption is twofold. First of all, with data from different sources and no information on the extent of early repayment fees, it seems that taking a fee of zero is the more fair decision. On the other hand, where including a fixed fee will increase both the observed and expected future value, it does not seem that the fee will affect the relative performance of the methods.

4.3.4. Evaluating the expected future value with respect to the observed future value For each of the uncensored test set cases, the observed future value can be computed giving the eventual outcome and be compared with the expected future values using the models. Table 6 lists the mean of the absolute differences between the expected and the observed values per case. In Table 7, the mean expected future values of all uncensored test set loans are listed and can be compared with the mean of the true future loan value at the bottom of the table.

5. Results

The results in Tables 4, 5, 6 and 7 are grouped per evaluation measure. For Tables 5 and 6, we used a notational convention where the best test result (each time the smallest value) per data set is underlined and denoted in boldface. Performances that are significantly different at a 5% level from the top performance with respect to a one-sided Mann-Whitney test are denoted in boldface (a Bonferroni correction was used due to multiple testing). As the AUC values in Table 4 are point estimates and do not represent samples, here simply the three highest values are underlined for each evaluation time and data set. In Table 7 the three values that lie closest to the mean future value per loan are underlined. Table 8 summarizes the results of all preceding tables by giving the average ranks of the models for all evaluation methods.

In Table 4 we note that the sample size is of real importance to get better receiver operating characteristics curves, as AUC values are generally larger for data sets with more

Table 8 Average ranking of the methods used depending on the evaluation method

	<i>AUC 1/3</i>	<i>AUC 2/3</i>	<i>AUC 3/3</i>	<i>MSE</i>	<i>MAE</i>	<i>MAD from FV</i>	<i>EFV versus FV</i>
AFT Weibull	<u>4.20</u>	<u>4.00</u>	<u>4.20</u>	5.15	5.00	<u>3.10</u>	5.80
AFT exponential	<u>4.70</u>	<u>5.30</u>	<u>5.50</u>	8.00	8.00	<u>6.60</u>	<u>2.20</u>
AFT log-logistic	<u>4.00</u>	<u>3.80</u>	<u>3.50</u>	5.65	5.80	4.20	4.70
Cox PH	<u>4.90</u>	<u>5.40</u>	<u>4.80</u>	<u>1.80</u>	<u>2.00</u>	<u>3.30</u>	5.90
Cox PH nat. splines	4.30	4.40	4.90	4.70	4.80	4.85	<u>3.55</u>
Cox PH penal. splines	<u>3.80</u>	<u>2.80</u>	<u>3.20</u>	<u>2.90</u>	<u>2.80</u>	<u>2.55</u>	<u>5.65</u>
Mixt. cure	<u>4.80</u>	<u>4.70</u>	<u>4.80</u>	<u>3.20</u>	<u>3.20</u>	<u>3.50</u>	<u>2.70</u>
Multi-event mixt. cure	5.30	5.60	5.00	4.60	4.40	7.90	<u>5.50</u>

The three best values are underlined

observations. Another factor that seems important is the length of the loan duration. Comparing the AUC results of data set 3 with data set 2, and data set 8 with data set 7, AUC values seem to go down when moving from a shorter loan term to a longer one, though data come from the same bank and has bigger sample size. This might be expected as it is known that making predictions becomes harder when moving to longer time frames. Looking at the overall result in Table 4, however, it is hard to draw conclusions regarding the preferred survival method when looking at the AUC alone, as the values are very close to each other (we note that ties in Table 4 are due to rounding). This can also be seen in Table 8, as average rankings regarding AIC range from 2.8 to 5.6. A Cox PH method with penalized splines shows to be the preferred method each time. A log-logistic AFT model seems to be a good alternative when considering the average ranking, although only appearing 10 out of 30 times among the top three in Table 4.

Next we consider mean squared differences (MSE) and mean absolute differences (MAE) from the observed default time (see Table 5). Although many performance measures are not significantly different from the top performance at the 5% level, a general trend for these evaluation measures is that the non-AFT models clearly outperform the AFT models. An interesting observation occurs when looking at the respective sample sizes of the evaluated sets. As depicted in Figure 3, only the actual defaults are evaluated in Table 5. Where most data sets here are quite small (166 cases and less), four sets are still considerable in size: those for data sets 1, 2, 3 and 8. For these data sets, more models can be excluded as their results (for MSE) are significantly worse (in bold). Additionally, we observe for these data sets that the Cox PH is very dominant here, being the best model in seven out of eight cases (considering both MSE and MAE for all four data sets). Considering all ten data sets, especially the exponential AFT model seems to have default time predictions that are significantly far off the true default times. With average rankings having a bigger range compared to ROC (from 1.8 to 8), it seems that the default time prediction measure clearly favors the plain Cox PH model when the sample size is considerable. When less cases were evaluated, the Cox PH

with natural splines and the mixture cure model seem to be good alternatives.

Table 6 lists the mean of the absolute differences between the model expected future loan value estimates and the true values. Note that these differences are bigger for loans with a longer loan term, which makes sense, as here the loan amounts are larger too. Consulting Table 6 along with 8, it becomes clear that the Cox PH model with penalized splines is again outperforming the other methods (although insignificantly), followed by the Weibull AFT and the plain Cox PH model. The table lists two clearly inferior methods, which are the exponential AFT model and the multiple event mixture cure model.

Regarding the financial metrics in Table 6, the mean absolute differences can get to a substantial size (e.g., in data set 3), but considering Table 7 we note that the mean expected values per loan are close to the mean observed value of the loans for all methods. The results in this table clearly highlight the abilities of survival analysis in the credit risk context. It should be noted that all estimates are very close to the observed mean future value per loan. Where Table 8 highlights the exponential AFT model with an average ranking of 2.2, the mixture cure model performs better than all other methods in five out of ten data sets (data sets 5–8 and 10), whereas exponential AFT is ranked best in three out of ten. Additionally, for Table 7, the mixture cure model tends to outperform on smaller sample sizes, whereas the exponential AFT performs better on bigger sample sizes.

Drawing a general conclusion from Table 8, the shortcomings of AUC again become apparent. Having become a major metric in the financial world to evaluate classification models, this metric is currently often applied to survival analysis models, but there are several issues. Firstly, it does not seem to be able to discriminate one survival model from another one, given small ranges of average rankings, and secondly, it evaluates the models by looking at predictions on individual case levels, in contrast with the default time or expected loan value predictions. Considering these latter evaluation methods, Cox PH models with and without splines and single event mixture cure models seem to be consistently good performers. The main advantage of the mixture cure model lies in the fact

that one basic assumption of survival models, namely the fact that a survival curve should go toward zero when time goes to infinity, and which is often violated for the loan data, is not needed at all when using the mixture cure model. All non-mixture cure models (here wrongly) assume this condition on the survival curve to hold.

The multiple event mixture cure model does not seem to live up to the expectations. It is important to note, though, that for a fair evaluation, one would have had to consider the results of other methods when using these to predict early repayment as well, not only default. Modeling default and early repayment in one model, as the multiple event mixture cure model does, as opposed to using two different survival models will likely lead to a better overall result, but additional research needs to be done to verify this.

6. Discussion

In this paper, we studied the performance of several survival analysis techniques in the credit scoring context. Ten real-life data sets were used, and we used three main evaluation measures to assess model performance: AUC, default time prediction differences and future loan value estimation. It is shown that Cox PH-based models all work particularly well, especially a Cox PH model in combination with penalized splines for the continuous covariates. The Cox PH model usually outperforms the multiple event mixture cure model, but the mixture cure model does not perform significantly different in most of the cases, and is among the top models using economic evaluation. This model has the advantage of not requiring the survival function to go to zero when time goes to infinity, which often is most appropriate for credit scoring data.

Starting from these findings, it would be interesting to further extend the mixture cure model and study the performance of the resulting model in comparison with a Cox PH model with penalized splines. This could be done by allowing for splines in the continuous covariates or time-dependent covariates for these models. Additionally, it would be interesting to run all the models again over data that have been coarse-classified, and compare its results with the results in this study. In particular, it would be interesting to compare the results of coarse classification to the spline-based methods in this study, which can be seen as an alternative for handling nonlinearity in the data. This study also points out that finding an appropriate evaluation measure to compare survival analysis remains an interesting challenge, as the AUC does not seem to have the right properties to really distinguish one method from another.

Acknowledgments—We acknowledge the support of the Fund for Scientific Research Flanders, KU Leuven Grant GOA/12/14, and of the IAP Research Network P7/06 of the Belgian Science Policy. The authors thank the referees for their constructive remarks.

References

- Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J and Vanthienen J (2003) Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* **54**(6): 627–635.
- Banasik J, Crook J and Thomas L (1999). Not if but when will borrowers default. *The Journal of the Operational Research Society* **50**(12): 1185–1190.
- Bellotti T and Crook J (2009). Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society* **60**(12): 1699–1707.
- Bellotti T and Crook J (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting* **29**(4): 563 – 574.
- Bellotti T and Crook J (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society* **65**(3): 340–350.
- Cai C, Zou Y, Peng Y, Zhang J (2012a). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine* **108**(3): 1255–1260.
- Cai C, Zou Y, Peng Y, Zhang J (2012b). *smcure: Fit Semiparametric Mixture Cure Models*. R package version 2.0. <https://CRAN.R-project.org/package=smcure>.
- Cao R, Vilar JM, Devia A (2009). Modelling consumer credit risk via survival analysis. *SORT* **33**(1): 3–30.
- Collett D (2003). *Modelling Survival Data in Medical Research, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press: Boca Raton, Florida.
- Cox D, Oakes D (1984). *Analysis of Survival Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Boca Raton, Florida.
- Cox DR (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2): 187–220.
- de Boor C (2001). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer: New York.
- Dejaeger K, Verbeke W, Martens D and Baesens B (2012). Data mining techniques for software effort estimation: A comparative study. *IEEE Transactions on Software Engineering* **38**(2): 375–397.
- Dirick L, Claeskens G, Baesens B (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research* **241**(2): 449–457.
- Eilers PHC, Rijnmond DM and Marx BD (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**(2): 89–121.
- Fantazzini D and Figini S (2009). Random survival forests models for SME credit risk measurement. *Methodology and Computing in Applied Probability* **11**(1): 29–45.
- Farewell VT (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**(4): 1041–1046.
- Gupta J, Gregoriou A and Ebrahimi T (2015). *Using Hazard Models Correctly: A Comparison Employing Different Definitions of SMEs Financial Distress*, pp 1–50. Available at SSRN:<http://ssrn.com/abstract=2457917>.
- Heagerty P and Saha P (2000). SurvivalROC: Time-dependent ROC curve estimation from censored survival data. *Biometrics* **56**(2): 337–344.
- Holmes P, Hunt A and Stone I (2010). An analysis of new firm survival using a hazard function. *Applied Economics* **42**(2): 185–195.
- Kalbfleisch J and Prentice R (2002). *The Statistical Analysis of Failure Time Data, 2nd edition*. Wiley: Hoboken.

- Kellison SG and Irwin RD (1991). *The Theory of Interest*, vol. 2. Irwin: Homewood.
- Klein J and Moeschberger M (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer: Berlin.
- Kleinbaum D and Klein M (2011). *Survival Analysis: A Self-Learning Text, Third Edition*. Statistics for Biology and Health. Springer.
- Leow M and Crook J (2016). The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. *European Journal of Operational Research* **249**(2): 457–464.
- Loterman G, Brown I, Martens D, Mues C and Baesens B (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting* **28**(1): 161–170.
- Ma P, Crook J and Ansell J (2010). Modelling take-up and profitability. *Journal of the Operational Research Society* **61**(3): 430–442.
- Narain B (1992). Survival analysis and the credit granting decision. In: Thomas LC, Crook JN and Edelman DB, editors, *Credit Scoring and Credit Control*, pp. 109–121. Clarendon Press: Oxford.
- O'Sullivan F, et al. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science* **1**(4): 502–518.
- Peng Y and Dear K (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56**(1): 227–236.
- Ruppert D, Wand MP and Carroll RJ (2003). *Semiparametric Regression*. Cambridge university Press: Cambridge.
- Stepanova M and Thomas L (2001). PHAB scores—Proportional hazards analysis behavioural scores. *The Journal of the Operational Research Society* **41**(9): 1007–1016.
- Stepanova M. and Thomas L (2002). Survival analysis methods for personal loan data. *Operations Research Quarterly* **50**(2): 277–289.
- Sy J and Taylor J (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56**(1): 227–236.
- Therneau T (2015). *A Package for Survival Analysis in S*. R package version 2.37-7. <http://CRAN.R-project.org/package=survival>.
- Therneau TM and Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model*. Springer: New York.
- Thomas L, Edelman D and Crook J (2002). *Credit Scoring and Its Applications*. Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics: Philadelphia.
- Tong ENC, Mues C and Thomas LC (2012). Mixture cure models in credit scoring: if and when borrowers default. *European Journal of Operational Research* **218**(1): 132–139.
- Watkins JGT, Vasnev AL and Gerlach R (2014). Multiple event incidence and duration analysis for credit data incorporating non-stochastic loan maturity. *Journal of Applied Econometrics* **29**(4): 627–648.
- Zhang J and Thomas L (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting* **18**(2): 204–215.

*Received 6 February 2015;
accepted 13 September 2016*