European Journal of Operational Research 218 (2012) 132–139

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Mixture cure models in credit scoring: If and when borrowers default

Edward N.C. Tong *, Christophe Mues, Lyn C. Thomas

Southampton Management School, University of Southampton, Southampton SO17 1BJ, United Kingdom

ARTICLE INFO

Article history:
Received 8 December 2010
Accepted 7 October 2011
Available online 18 October 2011

Keywords:
Credit scoring
Survival analysis
Mixture cure models
Regression
Risk analysis

ABSTRACT

Mixture cure models were originally proposed in medical statistics to model long-term survival of cancer patients in terms of two distinct subpopulations – those that are cured of the event of interest and will never relapse, along with those that are uncured and are susceptible to the event. In the present paper, we introduce mixture cure models to the area of credit scoring, where, similarly to the medical setting, a large proportion of the dataset may not experience the event of interest during the loan term, i.e. default. We estimate a mixture cure model predicting (time to) default on a UK personal loan portfolio, and compare its performance to the Cox proportional hazards method and standard logistic regression. Results for credit scoring at an account level and prediction of the number of defaults at a portfolio level are presented; model performance is evaluated through cross validation on discrimination and calibration measures. Discrimination performance for all three approaches was found to be high and competitive. Calibration performance for the survival approaches was found to be superior to logistic regression for intermediate time intervals and useful for fixed 12 month time horizon estimates, reinforcing the flexibility of survival analysis as both a risk ranking tool and for providing robust estimates of probability of default over time. Furthermore, the mixture cure model's ability to distinguish between two subpopulations can offer additional insights by estimating the parameters that determine susceptibility to default in addition to parameters that influence time to default of a borrower.

1. Introduction

Survival analysis has been well established in the medical sciences and engineering as a method for modelling time-to-event data (Hosmer et al., 2008). The method was first introduced to credit scoring by Narain (1992). Its use in this context was further developed by Banasik et al. (1999), Stepanova and Thomas (2002) and Hand and Kelly (2001). Banasik et al. (1999) compared the performance of parametric and semi-parametric hazard models to logistic regression and found they were competitive and in some cases more predictive than the industry standard logistic regression. Stepanova and Thomas (2002) showed that survival analysis could also be used for predicting time to early loan repayment in addition to conducting behavioural scoring, which allowed expected profit calculations. Subsequently, Baesens et al. (2005) further extended the use of survival analysis for credit scoring to the non-linear domain by incorporating neural network methods.

With survival analysis, one can predict not only *whether* people will default on their loan repayments but also *when* they are likely to default; i.e., survival analysis methods allow one to estimate the probability of default over any time horizon of choice. Particularly the Cox proportional hazards (PH) model (Cox, 1972) seems to have become increasingly popular in the consumer credit risk literature over the past decade.

In addition to its use as a dynamic time-to-default prediction method in credit scoring, survival analysis also enables the investigation of seasoning effects where default intensity varies with time since loan origination (Brown and Larson, 2007). Such methods could also be useful for building consumer credit risk models that are compliant with the Basel II Accord (Basel Committee on Banking Supervision, 2005a).

The Basel II Accord has had a major impact on how banks in most countries determine their regulatory capital. With these new regulations, banks are allowed to use their own credit risk estimates, produced by internal rating systems, to determine the minimum capital to be set aside to cover the risk associated with lending. To satisfy the requirements of the Accord, there has been a greater emphasis on model calibration in addition to model discrimination. Prior to the Accord, credit scoring models needed to have strong *discrimination* – the ability to risk rank borrowers accurately. Since the introduction of the Accord, the focus has widened to include also *calibration* performance – the accuracy of the probability of default (PD) estimates themselves (Basel Committee on Banking Supervision, 2005b; Thomas, 2010).

There have been particular concerns in this area about the suitability of corporate models being applied to consumer lending and

* Corresponding author. Address: Southampton Management School, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom. Tel.: +44 (0)23 8059 2561; fax: +44 (0)23 8059 3844.

E-mail addresses: e.tong@soton.ac.uk (E.N.C. Tong), c.mues@soton.ac.uk (C. Mues), l.thomas@soton.ac.uk (L.C. Thomas).

0377-2217/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2011.10.007

retail loan portfolios, as the behaviour of corporate business is dissimilar to that of an individual (Crook et al., 2007). An alternative to the corporate model is to consider the potential losses arising from a portfolio of retail clients where a dynamic approach could be used to model losses over time. One modelling approach that combines credit scoring of individuals with modelling at a portfolio level is again based on the principles of hazard functions (Thomas et al., 2002).

In recent credit research, it has been demonstrated that using time-varying macroeconomic variables in Cox PH models further improved the accuracy of PD estimates in credit scoring (Bellotti and Crook, 2009). Similarly, Thomas (2009) suggested the hazard function approach to estimate the credit risk of whole portfolios of consumer loans as opposed to credit scoring at an account-specific level. Malik and Thomas (2009) then exploited the parallels between individual behavioural scores in retail lending and the ratings given to corporate bonds, by deciding to include the former alongside macroeconomic factors in a Cox PH model to estimate retail credit risk at a portfolio level.

In standard survival analysis, the survival function is the probability of observing a survival time greater than some stated value of t, denoted $S(t) = P(T > t)$. $S(t)$ is also equal to one minus the cumulative distribution function, i.e., $S(t) = 1 - F(t)$. This presumes that $S(t)$ tends to zero as time extends and that all observations will eventually experience the event of interest. However, there are examples in credit scoring where a substantial proportion of accounts may not experience the default event during the lifetime of the loan or credit line and hence $S(t)$ will plateau to non-zero levels. Such accounts can be thought of as non-susceptible to default – in other words, long-term survivors.

Mixture cure models are an extension to the standard survival model which have been recently used in medicine to model survivors in cancer clinical trials in terms of two distinct subpopulations (Sy and Taylor, 2000). In one subpopulation, patients are non-susceptible (cured) and cancer-free after treatment, while the other subpopulation contains patients that are susceptible (uncured) and, in time, will experience cancer remission. Such data typically exhibit heavy censoring at the end of the follow-up period. Empirical evidence of such a non-susceptible subpopulation has been found in a number of trials. In a clinical study of breast cancer by Farewell (1986), 139 patients were observed for time to relapse or death due to the disease for three treatment arms of adjuvant therapy. The Kaplan–Meier survival curves from these three treatment groups were shown to level off above 0.4, resulting in a large proportion of right-censored observations from patients who survived the follow-up period and may possibly be cured of the disease. Hence, the standard PH model would be less appropriate, as the cancer survival distribution would not reach zero as time extends infinitely.

The theory behind the mixture cure model was introduced to the field of medical statistics by Farewell (1982). The general model incorporates two components – an *incidence* model for predicting which are the susceptible individuals and a *latency* model predicting survival times of individuals conditional on their being susceptible. The incidence component is essentially a binary classification model (e.g., a logistic regression). For the latency part, the original mixture model proposed by Farewell (1982) used a parametric survival model based on the Weibull distribution to estimate survival times. However, in the last decade, semi-parametric methods have also been developed, which offer more flexibility as they do not require a survival distribution to be specified for the incidence component. The latter methods often use an EM algorithm to find maximum likelihood estimates of the parameters for a combined likelihood function of a binary regression model and the proportional hazards model (Peng and Dear, 2000; Sy and Taylor, 2000).

Although the Cox PH approach has been widely established in the consumer credit literature, the former clearly suggests there might be situations where standard survival analysis may be insufficient, particularly if a substantial proportion of account observations is right-censored simply because they would not experience default during the lifetime of the loan or credit line. In that case, unlike the mixture cure approach, standard survival analysis does not allow one to distinguish between those covariates that affect whether an account is susceptible to default and those that may affect the timing of default. Interestingly, there have already been some recent studies demonstrating the superior performance of mixture cure models on corporate bankruptcy prediction (Topaloglu and Yildirim, 2009) and default prediction in corporate real estate loans (Yildirim, 2008).

The incidence component for binary classification in the mixture cure model has so far been implemented with generalized linear models with the use of the logistic or probit link functions. A natural benchmark for model comparison would then be the standard logistic regression model. In addition, many more advanced binary classification methods have been applied to credit scoring, including tree-based methods, clustering algorithms, support vector machines, neural networks and bagging/boosting methods (Baesens et al., 2003; Hastie et al., 2009; Tsai and Chen, 2010; Yeh and Lien, 2009). These methods have not been incorporated in the present study, but they could be considered in future research. However, unlike the score based methods such as logistic regression and proportional hazards modelling, some of these are not able to be used in practice because of the legal requirements that one must be able to explain why an applicant is refused credit.

In the present paper, we therefore investigate the use of a mixture cure model on a large dataset of consumer credit accounts from a personal loans portfolio of a major UK retail bank. The consumer accounts will be modelled as two distinct subpopulations – those that are cured and will not default during the lifetime of the loan or credit line, along with those that are uncured and will eventually experience a default at some time point during the loan term. We develop default prediction models with logistic regression, standard Cox PH and the newer mixture cure approach. The results are examined and compared across the three methods and performance is evaluated through cross validation. As increasingly both discrimination and calibration performance are important, the objective of the study is twofold: the first aim is to identify a superior method for risk-ranking accounts at various time intervals. The second goal is to evaluate the calibration performance of the methods by comparing the accuracy of the predicted survival probabilities against the observed survival probabilities. These aims are particularly relevant as the ongoing global financial crisis which started in 2007 has increased awareness of consumer credit models.

The organization of the remainder of the paper is as follows. In Section 2, the various statistical models used in the study, including the mixture cure model, will be outlined. Furthermore, it will be explained how these models were trained and validated. The data set used in the study is described in Section 3. Next, the results of our experiments are discussed in Section 4. Section 5 will conclude the paper and identify some topics for further research.

2. Statistical models

The mixture cure, standard Cox PH and logistic models are outlined in the following subsections.

2.1. Mixture cure model

The mixture cure model (Farewell, 1982; Peng and Dear, 2000) proposes to distinguish between two subpopulations of accounts based on susceptibility to an event of interest – a segment that

shall not experience the event of default during the loan term (long-term survivors) and another segment for those that will eventually default. Hence, a binary random variable Y is defined for the default event with $Y = 0$ denoting that the account is non-susceptible and a long-term survivor, while $Y = 1$ states that the account is susceptible and will default at some time point, though it may be censored in the dataset. Let us also define a censoring indicator δ, where $\delta = 1$ indicates non-censored accounts and $\delta = 0$ indicates censored accounts. There are two possibilities for accounts that do not default in the exposure period of the study. Accounts that do not default in the dataset either will not default in the future or they are right-censored at the end of the data period and would eventually default given sufficient exposure time. There are then three possible states of the data as follows:

$\delta = 1$ and $Y = 1$: non-censored, susceptible, hence the account is observed to default;
$\delta = 0$ and $Y = 1$: censored, susceptible, hence the account would eventually default;
$\delta = 0$ and $Y = 0$: censored, non-susceptible, hence long-term survivor.

2.1.1. Model formulation

The mixture cure model is then given by

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

where $S(t|\mathbf{x},\mathbf{z})$ is the marginal (unconditional) survival function of T for the entire population, incidence $\pi(\mathbf{z})$ denotes the proportion of accounts susceptible to default given a covariate vector $\mathbf{z} = (z_1,\ldots,z_p)$, and $S(t|Y=1,\mathbf{x}) = P(T > t|Y=1,\mathbf{x})$ is the latency or survival function conditional on the account being susceptible to default given a covariate vector $\mathbf{x} = (x_1,\ldots,x_q)$ which may or may not comprise the same covariates as $\mathbf{z}$.

It should be noted that $S(t|\mathbf{x},\mathbf{z}) \to 1 - \pi(\mathbf{z})$ as $t \to \infty$. When there is no non-susceptible (cured) fraction the mixture model also reduces to the standard PH survival model, i.e. $\pi(\mathbf{z}_i) = 1$ for all $\mathbf{z}_i$.

2.1.2. If defaults occur: the incidence model component (logistic model)

The proportion of susceptible accounts, given by $\pi(\mathbf{z}) = P(Y = 1|\mathbf{z})$, may be modelled using a binary regression model. Possible link functions include the logit, probit and the less commonly used complementary log–log link. The logit link was used in the present study because it has convenient parameter interpretations based on the odds ratio and is also well known to the credit risk community:

$$\log\left(\frac{\pi(\mathbf{z})}{1-\pi(\mathbf{z})}\right) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p = \mathbf{z}^T\boldsymbol{\beta} \tag{2}$$

where $\boldsymbol{\beta}$ is the vector of regression parameters associated with $\mathbf{z}$.

2.1.3. When defaults occur: the latency model component (proportional hazards)

There have been a number of proposed parametric and semi-parametric estimations of $S(t|Y = 1)$. The present study will implement the semi-parametric estimate approach for $S(t|Y = 1)$, as this has the closest relation to the standard Cox PH model.

Hence, like in the Cox PH model, the conditional distribution of T is represented by

$$\begin{aligned} S(t|Y=1) &= S_0(t|Y=1)^{\exp(\mathbf{x}^T\mathbf{b})} \\ &= \exp\left(-\exp(\mathbf{x}^T\mathbf{b})\int_0^t h_0(u|Y=1)du\right) \end{aligned} \tag{3}$$

where $S_0(t|Y = 1)$ and $h_0(t|Y = 1)$ are the conditional baseline survival and hazard functions, respectively. In a semi-parametric model, the $h_0(t|Y = 1)$ function is an arbitrary and unspecified hazard function and is not a function of the $\mathbf{x}$ covariates. Through estimates of $\boldsymbol{\beta}$ and $\mathbf{b}$, the mixture cure model allows to separate the covariate effects on the incidence and latency, respectively, thus providing a more flexible class of models when there is evidence to suggest a non-susceptible group of accounts.

2.1.4. Likelihood function

Let the data take the form $(t_i,\delta_i,\mathbf{x}_i,\mathbf{z}_i)$, $i = 1,\ldots,n$ where δ_i is the censoring indicator with $\delta_i = 1$ if t_i is uncensored and $\delta_i = 0$ if otherwise. For account i, the likelihood contribution is $\pi(\mathbf{z}_i)f(t_i|Y_i = 1, \mathbf{x}_i)$ for $\delta_i = 1$ and $(1 - \pi(\mathbf{z}_i)) + \pi(\mathbf{z}_i)S(t_i|Y_i = 1, \mathbf{x}_i)$ for $\delta_i = 0$, where $f(\cdot)$ is the conditional probability density function of T.

Hence the observed full likelihood is given by:

$$\begin{aligned} L(\mathbf{b},\boldsymbol{\beta}) = \prod_{i=1}^{n} &\{\pi(\mathbf{z}_i)f(t_i|Y_i=1,\mathbf{x}_i)\}^{\delta_i} \\ &\times\{(1-\pi(\mathbf{z}_i)) + \pi(\mathbf{z}_i)S(t_i|Y_i=1,\mathbf{x}_i)\}^{1-\delta_i} \end{aligned} \tag{4}$$

The full likelihood comprises a logistic and PH component. When the non-susceptible fraction does not exist, then $\pi(\mathbf{z}_i) = 1$ and the mixture cure likelihood function reduces to the likelihood function for the standard survival model. In that sense, the standard survival model is a special case of the mixture cure model.

Using the relationship $f(t) = h(t)S(t)$ and as long as Y is known, the account-level log likelihood function is given by the sum of the following two components:

$$l_l(\boldsymbol{\beta};y) = \log\prod_{i=1}^{n}\pi(\mathbf{z}_i)^{y_i}(1-\pi(\mathbf{z}_i))^{1-y_i} \tag{5}$$

$$l_L(\mathbf{b},H_0;y) = \log\prod_{i=1}^{n}h(t|Y=1,\mathbf{x}_i)^{\delta_i y_i}S(t|Y=1,\mathbf{x}_i)^{y_i} \tag{6}$$

where $H_0(t|Y=1) = \int_0^t h_0(u|Y=1)du$ is the conditional cumulative baseline hazard function.

It can be seen that (5) is the log likelihood function of a binary regression model and (6) is the log likelihood function of the proportional hazards model with the inclusion of conditioning on susceptible accounts with an offset variable $\log(y_i)$. However, Y is only partially observed in reality, as only δ is known. When $\delta = 0$, Y can be considered missing. The EM algorithm may be used to circumvent such a missing data problem.

2.1.5. EM algorithm

Non-censored accounts ($\delta_i = 1$) in the data correspond to observed defaults whereas censored accounts ($\delta_i = 0$) have an unobserved event of default. In order to maximize Eq. (4), an estimate of the random variable Y_i is necessary.

To estimate $\mathrm{E}(Y_i)$, we have:

$$\mathrm{E}(Y_i) = \begin{cases} 1 & \text{if } \delta_i = 1, \\ \dfrac{\pi(\mathbf{z}_i)S(t_i|Y_i=1,\mathbf{x}_i)}{1-\pi(\mathbf{z}_i)+\pi(\mathbf{z}_i)S(t_i|Y_i=1,\mathbf{x}_i)} & \text{if } \delta_i = 0 \end{cases} \tag{7}$$

The Y_i's represent the fractional allocation to the susceptible group and can be considered case weights for the individual accounts in the likelihood function. The expectation maximization (EM) algorithm deals with such missing data by estimating the unobserved variable for the censored accounts. EM is an iterative maximization algorithm. In the expectation (E) step, the likelihood equation in (4) is estimated with the best guess of the incomplete data, after which the expected values for the incomplete variable Y_i are computed. The second step, i.e., the maximization (M) step, involves estimating parameters using the expected values for Y_i found in the previous step. The steps are repeated by replacing estimated parameters back into the equation and iterating until convergence on the estimates of $\mathbf{b}$, $\boldsymbol{\beta}$ and $S_0(t|Y = 1)$.

We can substitute the expected values from (7) into the log likelihood function in (5) and (6). The log likelihood function will then be maximized when convergence occurs.

To approximate (6) without specifying the baseline hazard function, Peng and Dear (2000) proposed a partial likelihood estimator following a similar method to Breslow (1974):

$$\log \prod_{j=1}^{k} \frac{\exp(s_j \mathbf{b})}{\sum_{i \in R_j} y_i \exp(\mathbf{x}_i \mathbf{b})^{d_j}} \tag{8}$$

where k is the number of distinct default times, d_j the number of tied uncensored accounts at t_j, s_j the sum of covariate vectors associated with the uncensored accounts in d_j and R_j is the risk set at t_j, i.e. the set of uncensored accounts alive just prior to t_j.

Due to the EM algorithm, the standard errors of the coefficients were not readily available; hence the inverse of the observed information matrix was used to compute the standard errors of the coefficients (Sy and Taylor, 2000). Corbière and Joly (2007) found this method competitive with the bootstrapping and multiple imputation approaches described in the literature.

2.2. Standard approaches – Cox PH regression and logistic regression

As it is arguably the most popular survival analysis technique in the consumer credit risk literature, a standard Cox PH model was fitted to the dataset so that the mixture cure model could be compared against it. The semi-parametric approach in hazard form is given by:

$$h(t|\mathbf{x}) = h_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta})$$

where $h(t|\mathbf{x})$ is the hazard or default intensity at time t conditional on a vector of covariates $\mathbf{x}$, and in which $h_0(t)$ is the baseline hazard, i.e., the propensity of a default occurring when all covariates are zero. In the Cox PH model, the baseline hazard is arbitrary and unspecified.

A series of standard logistic regression models were also fitted for comparison purposes, since this technique is traditionally widely used in the credit scoring industry (Thomas, 2009). There were two types of logistic regression methods developed based on the sampling of the dataset. In the first method – Logistic 1 –, all observations in the relevant loan term sample, regardless of the time duration for which they were exposed to the default event, were used to develop each model. Thus a Good was a borrower who did not default during the loan and a Bad was someone who did default during the loan irrespective on when this happened. The resulting model thus estimates the probability of default occurring at some time point during the agreed loan term.

In the second method Logistic 2, for each loan one looked at the situation on the anniversary of the loan starting and defined a Bad outcome to be a default in the next 12 months and a Good outcome to be that the loan was still paying up to date 12 months later. Thus a three year loan could give rise to three different data points in this case, i.e. following the approach proposed in Stepanova and Thomas (2002).

2.3. Model building and validation

The mixture cure, Cox PH and logistic regression models were developed and evaluated using 100-fold cross validation to get unbiased predictive performance estimates for each method. To provide an approximate measure of variance for these measures, bootstrapping was applied to the resulting validation sample to produce 95% confidence intervals with the percentile method. Confidence intervals were created from 1000 bootstrapped samples for each measure. The aforementioned method of cross validation with bootstrapping was derived from Finlay (2010), who originally suggested leave-one-out cross validation. However, as the mixture cure method was highly computationally intensive due to the EM algorithm, the choice of cross validation with 100 folds was viewed as a reasonable compromise.

Since the loan term covariate was found to be the strongest predictor of survival time, and as it restricts the time period in which default can actually occur, it was decided to stratify all models by loan term. Thus three modelling runs were included for the mixture cure approach, i.e., for 12-, 24- and 36-month loans, and another three for both the Cox PH approach and Logistic 1. Six Logistic 2 model runs were required for each loan term and 12-month time window within that loan term (leading to 1, 2 and 3 models for 12-, 24- and 36-month loans, respectively). Continuous variables were fitted linearly in all the models to allow consistent comparisons across the approaches. Variable selection was performed using stepwise backward elimination with a significance threshold of 5%. For backward elimination selection, a full model was initially fitted with all candidate covariates included. The covariates were then tested for statistical significance and the largest p-values were removed one by one until the remaining covariates had p-values of less than 5%.

A number of discrimination measures for assessing the risk ranking of accounts were used. Industry standard measures such as the Area Under the Receiver Operating Characteristic Curve (AUC) and the Kolmogorov–Smirnov (KS) statistic were computed. The H measure, proposed by Hand (2009), was also computed on the validation samples. Hand (2009) argues that the H measure is a superior alternative to the AUC and KS because it is a coherent estimator of discrimination performance. Unlike the AUC, it is not sensitive to the empirical score distributions of the default and non-default groups in the sample. The AUC has a deficiency in that it uses different misclassification cost distributions for different classifiers, which implies that the AUC uses different metrics to evaluate different classification algorithms. However, the H measure maintains coherence because the misclassification cost distributions functions are given by a pre-specified beta distribution, where the default symmetric $beta(x;2,2)$ has been proposed. If there would be any disagreement between these discrimination measures, Hand (2009) has argued that the H measure should be considered the measure of choice to compare the performance of each method.

For the survival approaches, the score function was derived as one minus the survival probability at the time point of interest. The performance was also measured at intermediate time points for the 24- and 36-month loan term models. For example, the 24-month model had discrimination assessed at 12 months and at 24 months conditional on survival to 12 months (i.e., could the model predict default in the second year of the loan term given that the account had not defaulted in the first year).

As a calibration measure, the concordance correlation (Lin, 1989, 2000; Steichen and Cox, 2002) was used to assess the agreement of the predicted survival probabilities from a given model relative to the observed survival probabilities as computed by the Kaplan–Meier (KM) estimator. The Pearson correlation coefficient was not used because it is a linear measure of association. If the KM survival estimates were plotted against the model-based survival estimates, a well-calibrated model would produce estimates that fall on a 45° line through the origin. The Pearson correlation would fail to detect departure from the 45° line but agreement measures such as the concordance correlation and intra-class correlations will correct for this. In addition, marginal survival plots of applicants stratified by home owner status were also used to graphically assess goodness of fit; a similar analysis can be conducted for other discrete input variables.

The competing risk of early repayment was adjusted for in the calibration results through the Kaplan–Meier estimate for the

competing event. The adjustment for early repayment was necessary because a high rate of early repayment would decrease the number of defaults observed in a portfolio and vice versa. Models that do not account for such a competing risk would overestimate the number of defaults. Hence, the expected cumulative number of defaults at time T for a given model was computed as

$$E(D_T) = \sum_{i=1}^{n} \sum_{t=3}^{T} (S(t-1|i) - S(t|i)) \times KM_e(t) \quad (9)$$

where $S(t|i)$ is the Cox PH or mixture cure model based survival estimate at time t for account i and $KM_e(t)$ is the Kaplan–Meier based estimate of early repayment at time t. The computation started at $t = 3$ because defaults could not occur prior to three months and was summed over the length of the period T for each account and for the total number of accounts n in the loan term.

The Logistic 1 models were assessed for calibration by annualizing or scaling the PD estimates to a one year window. The Logistic 2 models were not assessed because they did not offer an equivalent method to adjust for early repayment on a monthly basis.

The models were fitted in SAS 9.2 software (SAS Institute Inc., Cary, NC, USA). The SAS macro written by Corbière and Joly (2007) was used to estimate the parameters of the mixture cure model via the EM algorithm described in Section 2.1.5. Analyses involving the bootstrapping, discrimination and calibration measures were run in R 2.13.0 (R Development Core Team, Vienna, Austria) and Stata 11.0 (StataCorp, College Station, TX, USA).

3. Data

The dataset consisted of 27,527 observations of consumer accounts from a major UK retail bank previously used in Stepanova and Thomas (2002). The dataset was derived from a personal loan portfolio which comprised three loan term durations of 12, 24 and 36 months. The dataset contained covariates relating to application characteristics and loan repayment activity over a maximum follow-up period of 36 months. The sample was collected from all consumers that were accepted and subsequently taken up for a personal loan over the observation period.

An account was classed as a default (Bad) account if it was at least 90 days in arrears. Accounts that were not in arrears or in arrears for less than 90 days were classed as non-default (Good). As Table 1 showed, the bad rate was low across all loan terms. The event of early repayment was also collected in the dataset. Fourteen application variables were available as candidate covariates in model development, viz. age of applicant, amount of loan, time at current address (years), time at current employer (years), gender, number of dependent children, frequency of salary payments, home phone status, amount of insurance premium, type of loan (single, joint), marital status, term of loan, homeowner status and purpose of loan.

4. Results

Next, we present the discrimination and performance results for all loan term models. Thereafter, an example will be shown of the model parameters obtained for one of the mixture cure models and for its Cox PH model counterpart.

Table 1
Summary statistics of defaults and early repayments for all loan terms.

Loan term	N	No. of defaults (%)	No. of early repayments (%)
12	10027	274 (2.7)	2450 (24.4)
24	9979	473 (4.7)	3838 (38.5)
36	7521	376 (5.0)	2992 (39.8)

4.1. Discrimination performance

In each loan segment, discrimination performance for each of the four modelling approaches was assessed by computing the AUC, KS and H measure on the validation samples obtained from the cross validation procedure. The results shown in Table 2 suggest that the Cox PH, mixture cure and the two logistic methods performed well on all discrimination measures. The bootstrapped 95% confidence intervals suggest that the performance for all the methods were broadly similar. For the 36-month loan models, the discrimination ability of all the models was sustained even at the 36-month time point (i.e., where the probability of default in the third and final year, conditional on survival up to the start of year 2, is estimated). Based on the AUC measure, both survival approaches performed similarly with an apparent marginal improvement over the logistic models on the 36-month model at 36 months conditional on survival to 24 months. However, the 95% confidence intervals were wide which suggests the models performed similarly.

The Logistic 2 models tended to only marginally underperform compared to both survival approaches and the Logistic 1 models in later time periods of assessment, e.g. loan term 36 at time point assessments of 24|12 and 36|24. Whereas each of the survival and Logistic 1 models were built on the whole loan segment, the available sample for the Logistic 1 models gets progressively smaller for the later time periods; however, this did not result in a significant loss of discrimination performance relative to those other methods, suggesting that logistic regression can be fairly robust to smaller samples.

4.2. Calibration performance

The calibration performance of the Logistic 1 and survival approaches were assessed on cross validated samples for each loan term. Table 3 tabulates the expected number against the observed number of defaults after the adjustment for early repayment given in (9) for the survival approaches. Results are reported for the three loan term segments, i.e., 12, 24 and 36 months, some of which at intermediate time points. The default figures reported in Table 3 are cumulative. For example, based on the 24-month loan term model, through cross validated predictions, the mixture cure approach predicted there would be 260.8 defaults by the end of the first year and 514.3 defaults by the end of the loan term.

Overall, the close agreement found between observed and expected values indicate a high level of calibration performance for both survival approaches at various time points. Both survival methods appear to be closely competitive, although the Cox PH model marginally predicted defaults better than the mixture cure model for the 36 month loan term model at 36 months. The survival approaches were especially well calibrated in the first 12 months of the 24 and 36 month loan term models. In terms of calibration, both survival approaches were superior to the Logistic 1 method at the intermediate time intervals for the 24 months and the 36 month loan term models. The absolute error and percentage error was lower for both survival approaches in the intermediate scenarios. However, the Logistic 1 model performed better for end of loan term estimates, for example the expected number of defaults at 24 months for the 24 month loan term.

As an example to further illustrate calibration performance of the survival approaches, Figs. 1–3 displayed the marginal survival functions against Kaplan–Meier estimates, each of which stratified by homeowner status, for the three loan terms on a random sample of the dataset. Note that a similar analysis can be made for other (discrete) covariates at the interest of the analyst. The upper and lower functions on the graphs are associated with homeowners and non-homeowners, respectively, indicating that homeowners

Table 2
Discrimination measures on 100-fold cross validation samples for Cox PH, mixture cure and logistic models.

Model	Loan term	Time	AUC (95% CI)†	*H* measure (95% CI)†	KS (95% CI)†	Score function
Cox	12	12	0.798 (0.770–0.823)	0.032 (0.023–0.056)	0.490 (0.447–0.542)	1-S(12)
Mixture Cure			0.798 (0.772–0.820)	0.033 (0.023–0.054)	0.482 (0.442–0.541)	1-S(12)
Logistic 1			0.796 (0.769–0.820)	0.032 (0.022–0.056)	0.473 (0.433–0.531)	
Logistic 2			0.799 (0.771–0.824)	0.037 (0.025–0.062)	0.498 (0.452–0.555)	
Cox	24	12	0.753 (0.725–0.782)	0.020 (0.011–0.040)	0.374 (0.339–0.439)	1-S(12)
Mixture Cure			0.753 (0.725–0.782)	0.019 (0.010–0.035)	0.385 (0.343–0.447)	1-S(12)
Logistic 1			0.749 (0.721–0.779)	0.018 (0.010–0.037)	0.379 (0.338–0.445)	
Logistic 2			0.753 (0.725–0.782)	0.013 (0.009–0.029)	0.382 (0.347–0.446)	
Cox	24	24|12*	0.717 (0.680–0.750)	0.024 (0.015–0.044)	0.347 (0.304–0.411)	1-S(24)/S(12)
Mixture Cure			0.719 (0.684–0.754)	0.024 (0.017–0.046)	0.355 (0.300–0.427)	1-S(24)/S(12)
Logistic 1			0.718 (0.683–0.751)	0.023 (0.015–0.044)	0.342 (0.299–0.418)	
Logistic 2			0.713 (0.677–0.747)	0.022 (0.014–0.043)	0.338 (0.294–0.411)	
Cox	36	12	0.756 (0.723–0.792)	0.017 (0.007–0.036)	0.380 (0.338–0.459)	1-S(12)
Mixture Cure			0.759 (0.726–0.789)	0.011 (0.007–0.026)	0.390 (0.348–0.465)	1-S(12)
Logistic 1			0.759 (0.726–0.794)	0.011 (0.007–0.029)	0.390 (0.346–0.468)	
Logistic 2			0.764 (0.730–0.798)	0.013 (0.007–0.034)	0.402 (0.351–0.475)	
Cox	36	24|12	0.728 (0.689–0.761)	0.026 (0.011–0.051)	0.356 (0.307–0.437)	1-S(24)/S(12)
Mixture Cure			0.727 (0.690–0.763)	0.026 (0.013–0.057)	0.345 (0.296–0.419)	1-S(24)/S(12)
Logistic 1			0.724 (0.685–0.758)	0.025 (0.011–0.049)	0.347 (0.301–0.422)	
Logistic 2			0.719 (0.683–0.756)	0.026 (0.013–0.061)	0.335 (0.284–0.410)	
Cox	36	36|24	0.667 (0.571–0.763)	0.056 (0.019–0.194)	0.293 (0.194–0.490)	1-S(36)/S(24)
Mixture Cure			0.665 (0.570–0.752)	0.037 (0.014–0.127)	0.344 (0.210–0.506)	1-S(36)/S(24)
Logistic 1			0.657 (0.564–0.742)	0.040 (0.014–0.146)	0.296 (0.186–0.450)	
Logistic 2			0.655 (0.551–0.754)	0.035 (0.012–0.120)	0.285 (0.192–0.477)	

* E.g. 24|12 indicates survival (non-default) up to 24 months conditional on survival to 12 months.
† Bootstrapped percentile 95% confidence intervals.

Table 3
Observed versus expected cumulative number of defaults on 100-fold cross validation samples with adjustment for early repayment.

Method	Model	Time	Observed	Expected	Abs (obs–exp)	% Error
Logistic	12	12	274	274.1	0.1	0.1
Cox			274	285.3	11.3	4.1
Mixture			274	283.5	9.5	3.5
Logistic	24	12	260	241.3	18.7	−7.2
Cox			260	260.0	0.0	0.0
Mixture			260	260.8	0.8	0.3
Logistic	24	24	473	473.0	0.0	0.0
Cox			473	515.0	42.0	8.9
Mixture			473	514.3	41.3	8.7
Logistic	36	12	176	129.2	46.8	−26.6
Cox			176	176.0	0.0	0.0
Mixture			176	175.7	0.3	−0.2
Logistic	36	24	341	254.5	86.5	−25.4
Cox			341	359.8	18.8	5.5
Mixture			341	358.3	17.3	5.1
Logistic	36	36	376	376.0	0.0	0.0
Cox			376	452.8	76.8	20.4
Mixture			376	460.2	84.2	22.4

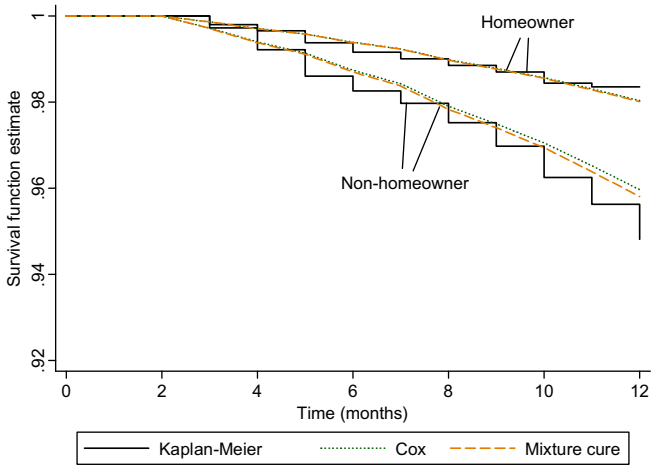

Fig. 1. An example of a marginal survival function stratified by home owner status for the 12 month loan term on a representative sample.

had higher survival rates across all time points and were thus found to be lower risk borrowers (which is in line with common credit scoring intuition). The homeowner characteristic was found to be statistically significant in all loan term segments according to both the Cox PH and mixture cure methods.

Figs. 1–3 suggests that both survival methods fairly accurately estimate marginal survival of homeowners and non-homeowners, as shown by their close agreement with the observed functions of the Kaplan–Meier estimator.

4.3. Regression model parameters

To illustrate how parameter estimate results are to be interpreted, Table 4 shows the results of two example runs for both survival approaches applied to the entire 36 month loan term accounts. The 12 and 24 month models were omitted for brevity. The survival component (latency model) of the mixture cure model identified two covariates that influenced the time to default – home phone and loan type. For example, having a joint loan decreased the hazard of default by 43% and customers with home phones had a 47% decrease in their hazard of default. The logistic component (incidence model) found similar covariates to the Cox PH model with the exception of age and loan type. The mixture cure results thus suggest that the 12 covariates found in the logistic component are significant predictors of the probability of being susceptible to default, whereas the survival component suggests only two covariates are predictive of *when* a default will occur given that the borrower is susceptible to default. None of the covariates found to predict survival time are found to influence susceptibility. A further comparison between the two survival approaches indicates that the

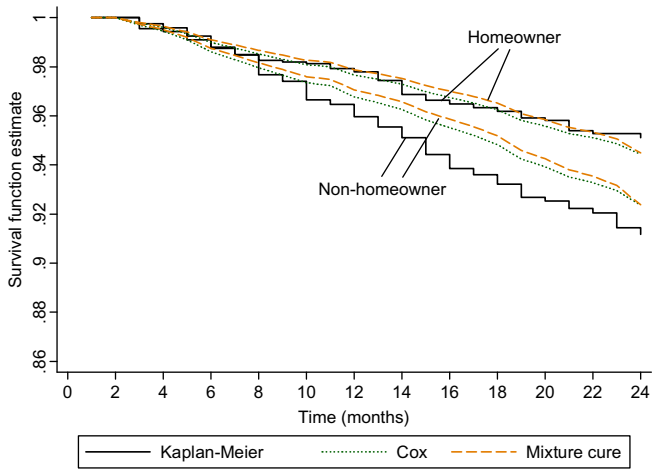

Fig. 2. An example of a marginal survival function stratified by home owner status for the 24 month loan term on a representative sample.

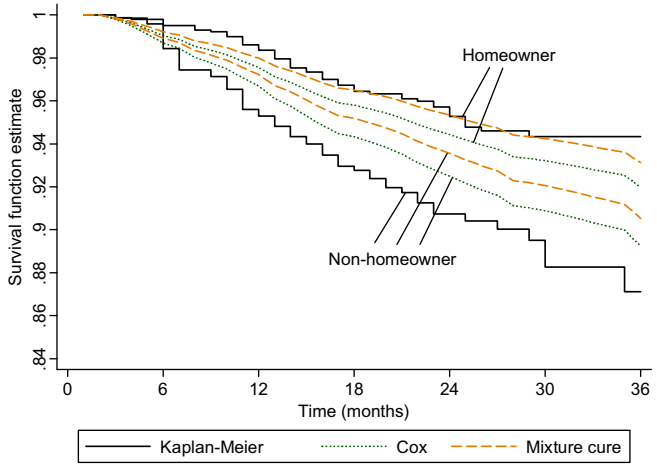

Fig. 3. An example of a marginal survival function stratified by home owner status for the 36 month loan term on a representative sample.

majority of covariates found in the Cox PH model were more suitable for predicting susceptibility in the mixture cure model, considering that they were mostly found in the logistic component.

5. Conclusions and future research

When faced with empirical evidence of a subpopulation in a loan portfolio that is not susceptible to default, the mixture cure approach has the advantage of being able to distinguish between two subpopulations in the modelling. The separation of subpopulations allows inference on the covariates that influence susceptibility to default and those that influence time to default given susceptibility. This provides an alternative interpretation to the standard survival model, by allowing one to distinguish between the factors that drive default and those that explain whether default happens earlier or later into the loan. In other words, the mixture cure interpretation focuses on both *if* and *when* a default will occur. For example, in the mixture cure model given in the previous section, 12 covariates were found to predict if a default will occur and another two distinct covariates were found to predict time to default given susceptibility to default. The standard logistic regression and Cox PH approach would not offer such interpretation separately.

The present study evaluated the discrimination and calibration performance of the mixture cure model on a personal loans portfolio and compared them to the standard Cox PH and two logistic modelling variants. The mixture cure, Cox PH and logistic models performed well and were closely competitive to one another in terms of discrimination performance.

The findings of the study reinforce previous research that indicated survival approaches can perform equivalently to logistic regression in terms of discrimination ability by being able to use all available information in the data. In addition, compared to logistic regression, there are substantial additional benefits to the use of survival analysis in its ability to predict time to an event through conditional survival estimates. These benefits relate to the estimation of the number of defaults arising from the portfolio at various time intervals and the expected profit arising from the portfolio. For calibration, the survival methods in the present study have outperformed the logistic model for the intermediate time intervals because of their ability to estimate the baseline hazard function across time and adjust for competing risks such as early repayment. The logistic model was shown to calibrate well only for the end of loan term estimates. As such, the survival approaches may be useful for accurate PD estimates in the fixed 12 month horizon for various loan terms, which is relevant for PD estimation within the Basel II Accord.

There were a few limitations in the analysis. The dataset used comprised accounts from a personal loans portfolio of a UK bank

Table 4
Parameter estimate results from the mixture cure and Cox proportional hazards models for 36 month loan term based on the entire sample.

Mixture cure				Cox proportional hazards			
Logistic model component	OR[†]	95% CI	*p*-Value		HR[†]	95% CI	*p*-Value
Age[*]	0.99	0.98–1.00	.044	Time at current address[*]	0.97	0.96–0.99	.001
Time at current address[*]	0.98	0.96–0.99	<.001	Time at current employment[*]	0.92	0.90–0.94	<.001
Time at current employment[*]	0.92	0.90–0.94	<.001	No. dependent children (0 vs. 2+)	0.78	0.60–1.00	.045
No. dependent children (0 vs. 2+)	0.75	0.60–0.92	.006	No. dependent children (1 vs. 2+)	1.03	0.75–1.40	.870
No. dependent children (1 vs. 2+)	1.03	0.80–1.32	.843	Frequency of payments	0.60	0.49–0.75	<.001
Frequency of payments	0.58	0.48–0.69	<.001	Homeowner	0.64	0.52–0.80	<.001
Homeowner	0.64	0.53–0.77	<.001	Insurance premium[*]	1.00	1.00–1.00	<.001
Insurance premium[*]	1.00	1.00–1.01	<.001	Loan type	0.72	0.56–0.91	.007
Loan purpose 1	2.70	2.18–3.34	<.001	Loan purpose 1	2.48	1.93–3.19	<.001
Loan purpose 2	0.77	0.60–0.99	.040	Loan purpose 2	0.80	0.59–1.08	.138
Loan purpose 3	1.15	0.87–1.52	.337	Loan purpose 3	1.14	0.80–1.61	.469
Loan purpose 4	1.15	0.68–1.94	.604	Loan purpose 4	1.12	0.59–2.13	.736
Survival model component	HR[†]	95% CI	*p*-Value				
Home phone (yes vs. no)	0.53	0.35–0.79	.002				
Loan type (joint vs. single)	0.57	0.46–0.72	<.001				

[*] Continuous covariate centred on mean.
[†] Odds ratio (OR), hazards ratio (HR).

with application characteristics. The mixture cure model was fitted to predict the timing of default in the semi-parametric component with these application covariates. These were time-constant covariates, although the time to default could be better estimated with time-varying covariates which may include behavioural scores and macroeconomic factors. However, such a model would become less suitable for certain prediction tasks as future values would have to be provided for any time-varying covariates. Given that a single sample was used in the present study, the results may not be readily generalized to other retail portfolios such as home loans where loan terms of 20 years or more are common. As the data was collected from one UK bank, further studies could evaluate mixture cure methods on non-UK datasets, in addition to other retail loan products.

Long-term survivors are a common occurrence in personal loan portfolios, as a substantial proportion of accounts do not reach default status during the lifetime of the loan. Although it was argued that in theory standard proportional hazards approaches may not be appropriate for such data because in the long run, such methods assume the survival function will approach zero, the results of our study suggest that the discrimination and calibration performance will not be adversely affected by doing so. Hence, the Cox PH model proved robust with regards to this aspect.

Future research could involve applications of the mixture cure model to credit card datasets where loan term durations are indefinitely lengthy because of the revolving nature of credit. Such datasets would have sufficiently long observation periods to have non-susceptible subpopulations that do not reach default status during the credit line, because some customers, classified as transactors, always pay off their balance at the end of each month. In addition, it would also be interesting to further apply mixture models to home loan mortgage defaults, considering that loan terms of 20–30 years are common for such portfolios.

The mixture cure model in the present paper has focused on using logistic regression in the incidence component. Future directions could consider alternative binary classification methods such as tree-based, ensemble classifiers and bagging or boosting techniques in lieu of logistic regression. Such advanced methods may have an impact on the latency component and yield improved performance for the mixture cure method.

Acknowledgments

We thank the three anonymous reviewers for their most useful recommendations which improved the paper.

References

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

Baesens, B., Gestel, T.V., Stepanova, M., Poel, D.V.d., Vanthienen, J., 2005. Neural network survival analysis for personal loan data. Journal of the Operational Research Society 56 (9), 1089–1098.

Banasik, J., Crook, J.N., Thomas, L.C., 1999. Not if but when will borrowers default. Journal of the Operational Research Society 50 (12), 1185–1190.

Basel Committee on Banking Supervision (BCBS), 2005a. International Convergence of Capital Measurement and Capital Standards – A Revised Framework. Bank for International Settlements, Basel.

Basel Committee on Banking Supervision (BCBS), 2005b. Studies on the validation of Internal rating systems. Working Paper 14, Basel.

Bellotti, T., Crook, J., 2009. Credit scoring with macroeconomic variables using survival analysis. Journal of the Operational Research Society 60 (12), 1699–1707.

Breslow, N., 1974. Covariance analysis of censored survival data. Biometrics 30 (1), 89–99.

Brown, J., Larson, C.E., 2007. The issue of retail credit risk seasoning and its impact upon basel II PD estimation. Retrieved from <http://www.promontory.com/assets/0/78/110/168/c19f3640-58b5-4ebe-afe3-abddff624f0e.pdf>.

Corbière, F., Joly, P., 2007. A SAS macro for parametric and semiparametric mixture cure models. Computer Methods and Programs in Biomedicine 85 (2), 173–180. doi:10.1016/j.cmpb.2006.10.008.

Cox, D.R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) 34 (2), 187–220.

Crook, J.N., Edelman, D.B., Thomas, L.C., 2007. Recent developments in consumer credit risk assessment. European Journal of Operational Research 183 (3), 1447–1465. doi:10.1016/j.ejor.2006.09.100.

Farewell, V.T., 1982. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 38 (4), 1041–1046.

Farewell, V.T., 1986. Mixture models in survival analysis: Are they worth the risk? The Canadian Journal of Statistics 14 (3), 257–262.

Finlay, S., 2010. Credit Scoring, first ed. Palgrave Macmillan, Basingstoke.

Hand, D., 2009. Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning 77 (1), 103–123. doi:10.1007/s10994-009-5119-5.

Hand, D.J., Kelly, M.G., 2001. Lookahead scorecards for new fixed term credit products. Journal of the Operational Research Society 52 (9), 989–996.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer, New York, NY.

Hosmer, D., Lemeshow, S., May, S., 2008. Applied Survival Analysis: Regression Modeling of Time to Event Data, second ed. Wiley-Interscience.

Lin, L.I.K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45 (1), 255–268.

Lin, L.I.K., 2000. A note on the concordance correlation coefficient. Biometrics 56 (1), 324–325.

Malik, M., Thomas, L., 2009. Modelling credit risk of portfolio of consumer loans. Journal of the Operational Research Society.

Narain, B., 1992. Survival analysis and the credit granting decision. In: Thomas, L.C., Crook, J.N., Edelman, D.B. (Eds.), Credit Scoring and Credit Control. OUP, Oxford, pp. 109–121.

Peng, Y., Dear, K.B.G., 2000. A nonparametric mixture model for cure rate estimation. Biometrics 56 (1), 237–243.

Steichen, T.J., Cox, N.J., 2002. A note on the concordance correlation coefficient. Stata Journal 2 (2), 183–189.

Stepanova, M., Thomas, L., 2002. Survival analysis methods for personal loan data. Operations Research 50 (2), 277–289.

Sy, J.P., Taylor, J.M.G., 2000. Estimation in a cox proportional hazards cure model. Biometrics 56 (1), 227–236.

Thomas, L.C., 2009. Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press, Oxford.

Thomas, L.C., 2009. Modelling the credit risk for portfolios of consumer loans: Analogies with corporate loan models. Mathematics and Computers in Simulation 79 (8), 2525–2534. doi:10.1016/j.matcom.2008.12.006.

Thomas, L.C., 2010. Consumer finance: Challenges for operational research. Journal of the Operational Research Society 61 (1), 41–52.

Thomas, L.C., Crook, J., Edelman, D., 2002. Credit scoring and its applications: Society for industrial and applied mathematics.

Topaloglu, Z., Yildirim, Y., 2009. Bankruptcy prediction. Working paper. CUNY, Department of Economics. New York. Retrieved from <http://www.efmaefm.org/0EFMAMEETINGS/EFMA%20ANNUAL%20MEETINGS/2009-milan/FULL_v1-1.pdf>.

Tsai, C.-F., Chen, M.-L., 2010. Credit rating by hybrid machine learning techniques. Applied Soft Computing 10 (2), 374–380. doi:10.1016/j.asoc.2009.08.003.

Yeh, I.C., Lien, C.-h., 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications 36 (2, Part 1), 2473–2480. doi:10.1016/j.eswa.2007.12.020.

Yildirim, Y., 2008. Estimating default probabilities of CMBS loans with clustering and heavy censoring. The Journal of Real Estate Finance and Economics 37 (2), 93–111. doi:10.1007/s11146-007-9046-6.