European Journal of Operational Research 277 (2019) 20–31

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Continuous Optimization

A prediction-driven mixture cure model and its application in credit scoring

Cuiqing Jiang[a], Zhao Wang[a,*], Huimin Zhao[b]

[a] School of Management, Hefei University of Technology, No.193, Tunxi Road, Hefei 230009, Anhui, PR China
[b] Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee, WI 53201-0742, USA

ARTICLE INFO

Article history:
Received 25 July 2017
Accepted 31 January 2019
Available online 7 February 2019

Keywords:
Risk analysis
Mixture cure model
Random forests
Time-dependent hazards
P2P lending

ABSTRACT

In the credit market, assessment of a borrower's default risk over time is essential to enabling timely risk management, since borrowers' exposure to risk and the losses that result from defaults are strongly related to the time when they default. Mixture cure models, with their ability to predict not only whether borrowers will default but also when they are likely to default, have been applied to credit scoring. We propose a prediction-driven mixture cure model, which sacrifices interpretability for potentially better prediction performance, and apply it to credit scoring. In the incidence part of the mixture cure model, we substitute the typical statistical incidence model (i.e., logistic regression) with a more flexible, and hopefully more accurate, classification method (i.e., random forests). For the latency part, we propose a survival analysis model, named Time-Dependent Hazards, which accommodates a direct relationship between failure times and covariates and can potentially better predict the probability of default over time than the standard Cox PH model. Empirical evaluation using real-world data from a major P2P lending institution in China shows that both extensions contributed to performance improvement in both discrimination and calibration.

1. Introduction

The primary activity of financial institutions, such as banks and peer-to-peer (P2P) lending institutions, is granting credit to borrowers by generating loans (Guo, Zhou, Luo, Liu, & Xiong, 2016). During a loan term, default may occur when the borrower fails to make required payments. Therefore, credit risk and its management are closely bound up with the business performance of a financial institution. A good credit risk evaluation method can help financial institutions distinguish non-creditworthy loan applications from creditworthy ones, thus alleviating credit risk. Credit scoring, an effective tool for assessing default risk during the lending process, is therefore extremely valuable in the credit market.

Moreover, it is important for financial institutions to be able to estimate a borrower's probability of default (PD) over the entire time horizon of the requested loan (e.g., monthly PD for a one-year loan). First, since loan repayment of most lending is a dynamic process (e.g., paid by installment), borrowers' exposure to risk and the losses that result from defaults are strongly related to the time when they default (Chang, Chang, Chu, & Tong, 2016). Second, the ability to estimate the time to default enables more timely risk management and control (Alves & Dias, 2015). Third, estimating the dynamic PD over time enables financial institutions to provide lenders with an estimated dynamic repayment ability of a borrower (in case the loan is issued), so as to better support the investment decision making of lenders.

Credit scoring in traditional financial lending contexts (e.g., bank loans and credit cards) has been the subject of extensive research for a long time (Baesens et al., 2003; Bhattacharya, Wilson, & Soyer, 2019; Djeundje & Crook, 2018; Djeundje and Crook 2019; Lessmann, Baesens, Seow, & Thomas, 2015). Earlier research treated credit scoring as a binary classification problem, where a loan is classified as either creditworthy or non-creditworthy (see Hand & Henley, 1997; Rosenberg & Gleit, 1994 for surveys of earlier research). Subsequently, plenty of classification methods, such as logistic regression, neural networks, and support vector machines, have been used to estimate a borrower's probability of default (see Lessmann et al., 2015 for a survey).

However, not only whether borrowers default on their loans, but also when they are likely to default, is important. Survival analysis, with its ability to model the dynamic process that leads to a borrower's default, has been applied in the context of credit scoring. It has been widely applied in the area of medicine and was first introduced to credit scoring by Narain (1992). With

* Corresponding author.
E-mail addresses: jiangcuiq@163.com (C. Jiang), xcwangzhao@163.com, xcwangzhao@mail.hfut.edu.cn (Z. Wang), hzhao@uwm.edu (H. Zhao).

https://doi.org/10.1016/j.ejor.2019.01.072
0377-2217/

survival analysis, one can predict the probability of default over any time horizon of choice (i.e., when they are likely to default). Its use in credit scoring has been further developed by Banasik, Crook, and Thomas (1999), Hand and Kelly (2001), and Stepanova and Thomas (2002), who applied and compared various standard survival models, both parametric and non-parametric. Bellotti and Crook (2009) and Im, Apley, Qi, and Shan (2012) further incorporated time-dependency into the proportional hazards model.

An often implicit assumption in most standard survival analysis models is that all borrowers will eventually experience the default event over a sufficiently long period of observation (Zhang & Thomas, 2012). However, in practice, most borrowers may not experience the event of default during the full loan terms, and some of these borrowers may be long-term survivors, who are not susceptible to default. More recently, there has been much research to relax the assumption, leading to an extension of standard survival analysis, generally referred to as mixture cure models, which model borrowers in terms of two distinct subpopulations (De Leonardis & Rocci, 2014; Dirick, Claeskens, & Baesens, 2015; Tong, Mues, & Thomas, 2012). In one subpopulation, borrowers are insusceptible (*cured*) and will never default during the lifetime of the loan, while the other subpopulation consists of borrowers who are susceptible (*uncured*) and will experience the event of default at some point of time.

A mixture cure model consists of two components: an *incidence* part, which predicts whether a borrower will default, and a *latency* part, which predicts the survival time of a borrower conditional on the borrower being susceptible to default. Beran and Djaïdja (2007) proposed a parametric mixture model, which consists of an incidence part with no explanatory variables and a latency part with an exponential distribution function, to analyze the default events in mortgages. Tong et al. (2012) applied a semi-parametric mixture cure model to the analysis of default events in a portfolio of UK personal loans, with the incidence and latency components modeled by logistic regression and Cox proportional hazards (PH) regression (Cox, 1972), respectively. The mixture cure model was further developed by De Leonardis and Rocci (2014) and Liu, Hua, and Lim (2015). De Leonardis and Rocci (2014) replaced the discrete time Cox PH function with a time-varying system-level covariate to capture the underlying macroeconomic cycle. Liu et al. (2015) extended the mixture cure model to a hierarchical Bayesian model for identifying future defaulters.

In this paper, we further explore the mixture cure model approach to credit scoring. Aiming to improve prediction performance, we propose extensions to the standard mixture cure model in both the incidence and latency components. In the incidence part, we substitute the typical statistical incidence model, which has been implemented with generalized linear models using logistic, log–log, or probit link functions, with a more flexible and hopefully more accurate classification method, i.e., random forests, which does not assume a certain data distribution and can better deal with potential nonlinear relationships. For the latency part, we propose a survival analysis model, named Time-Dependent Hazards (TDH), which accommodates a direct relationship between failure times and covariates in calculating the baseline survival function and can potentially better predict PD over time (e.g., monthly PD for a one-year loan). Despite the obvious intuitive interpretation and popularity, the Cox PH model does not directly accommodate the interaction between time and covariates, limiting the performance when predicting PD over time. In the proposed TDH, the time-related effect is considered in borrower-specific survival levels and the monthly baseline survival rate in each time interval is generated by averaging the borrower-specific survival levels over all observations at the corresponding time interval.

Both extensions are motivated by the need for more accurate prediction, since even a small (e.g., 1%) improvement in the prediction performance can yield a great decrease in loss for financial institutions (Hand & Henley, 1997). In addition to explanatory understanding of the effects of various factors, the ability to predict PD per se can also be valuable in some scenarios (Huang, Zhao, & Zhu, 2012). The proposed extensions relax the assumptions of the standard models (e.g., linearity and the way the effects of time and covariates are accommodated), are more flexible in fitting training data, and hopefully can also predict more accurately. However, the potential improvement in prediction performance is achieved while sacrificing interpretability of the standard models. The proposed more complex models (e.g., TDH and random forests) are hardly interpretable. In a scenario where interpretability is a mandatory requirement (e.g., when the financial institution is required by legal regulation to explain why a loan application is denied), the proposed extensions may not be an intuitive option, although interpretability can still be accommodated to some extent by some other measures, such as the feature importance measure of random forests. In such case, the proposed method might not work as a model to calculate regulatory capital, but can be used to improve the internal risk management practices of a financial institution.

We have evaluated our proposed mixture cure model using a large dataset on personal loans from a major P2P lending institution in China. We compared our method with the standard mixture cure model (i.e., logistic regression for the incidence component and Cox PH for the latency component), in terms of both discrimination performance (the ability to risk rank borrowers accurately) and calibration performance (the accuracy of the PD estimates themselves) (Tong et al., 2012). We also used a full factorial design to examine whether and how much each of the two proposed attempts (random forests for the incidence component and TDH for the latency component) contributes to overall performance improvement. Note that, although we applied and evaluated the proposed mixture cure model in P2P lending, the model itself is general and can be applied in other contexts, such as traditional bank loans. Also note that although we tested random forests as an example in our evaluation, there are numerous other classification methods developed in the machine learning field, which may be adapted and tested in the future.

The remainder of this paper is organized as follows. In the next section, we provide further background on the standard mixture cure model. We then present our proposed extensions in Section 3. We describe the empirical evaluation in Section 4 and report on the results in Section 5. Finally, we conclude the paper by summarizing our contributions and discussing future research directions.

2. Background

2.1. Notation

In mixture cure models, borrowers are divided into two subpopulations: cured (insusceptible) and uncured (susceptible). Borrowers in the insusceptible subpopulation will never default and will be censored at the end of any observation period. Borrowers in the susceptible subpopulation will eventually default. Let y_i be a binary random variable defined for the default event, with possible values $y_i = 0$ denoting that borrower i is cured and will never experience the default event and $y_i = 1$ denoting that borrower i is uncured and will experience the default event during or after the observation period. Let δ_i be a censored indicator, with $\delta_i = 1$ denoting that borrower i defaults within the observation period, and $\delta_i = 0$ otherwise. Hence, there are three possible states of borrowers (Table 1).

Table 1
The possible states of borrowers.

δ_i	y_i	Description
1	1	Non-censored. Borrower i is observed to have defaulted.
0	1	Censored, and borrower i would eventually default.
0	0	Censored, and borrower i will never default.

2.2. Standard mixture cure model

The standard mixture cure model is given by:

$$S_i(t) = 1 - p_i + p_i * S_i(t|y_i = 1). \tag{1}$$

$S_i(t)$ denotes the probability that borrower i $(i = 1, 2, \ldots, N)$ survives (i.e., has not yet defaulted) beyond time t. p_i is referred to as *incidence* and denotes the probability that the borrower will eventually default. $S_i(t|y_i = 1) = P(T_i > t|y_i = 1)$, where T_i denotes the default time of borrower i since the loan is approved, is referred to as *latency* and denotes the conditional probability that the borrower survives beyond time t given that the borrower will eventually default.

It should be noted that an advantage of mixture cure models is that the incidence component and the latency component are modeled separately. Hence, it is possible to distinguish variables that influence the probability of default from those that affect the default time distribution (Tong et al., 2012).

The incidence component, given by $p_i = P(y_i = 1)$, may be modeled using a regression model. Possible link functions include logit, probit, and the less commonly used log-log link. Following most of the existing credit scoring literature (e.g., Tong et al., 2012), we choose logistic regression to model the incidence component of the standard mixture cure model for model comparison:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_n z_{in} = \boldsymbol{\beta} \cdot \boldsymbol{z}_i, \tag{2}$$

where $\boldsymbol{z}_i = (1, z_{i1}, z_{i2}, \ldots, z_{in})$ is the vector of explanatory variable values observed on borrower i in the incidence model, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_n)$ is a vector of regression parameters associated with $\boldsymbol{z}_i$.

There have been several estimations of the latency component, $S_i(t|y_i = 1)$, both parametric and semi-parametric. Compared to a parametric estimation, which needs to make a specific assumption on the distribution of default time (e.g., Weibull, exponential, and gamma), a semi-parametric estimation is more flexible because it only needs to make a (less restrictive) assumption on the form of the survival function. Hence, we choose the semi-parametric Cox PH model for estimation of $S_i(t|y_i = 1)$, as a benchmark for model comparison:

$$S_i(t|y_i = 1) = S_0(t|y_i = 1)^{\exp(\alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_m x_{im})}, \tag{3}$$

where $S_0(t|y_i = 1)$ is referred to as the conditional baseline survival function, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ is the vector of explanatory variable values observed on borrower i in the latency model, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ is a vector of regression parameters associated with $\boldsymbol{x}_i$. Eqs. (1)–(3) constitute the basic framework of the standard mixture cure model.

3. Proposed prediction-driven mixture cure model

We propose extensions to both the incidence component and the latency component of the standard mixture cure model, in an attempt to improve prediction performance.

3.1. Time-dependent hazards for the latency component

The process of modeling the effect of covariates on default time in the standard Cox PH model is that the baseline hazards function is estimated on the basis of the ranked event times and then the effect of the covariates is to multiply the baseline hazard by a (borrower-specific) factor. Generally, the failure time of the baseline hazards function $S_0(t|y_i = 1)$ is independent of the covariates. However, this can be too restrictive in practice. For example, default due to a personal financial crisis may be accompanied with a lag period. That is, if the account balance of a borrower is not enough to pay the remaining installment, the borrower may still strive to avoid default, which would leave a bad credit record. The borrower may eventually default in the future, but maybe not right away. This lag period may also depend on the characteristics of the borrower. This time-related effect may not show up in a long predicting time interval (e.g., the lag period may be contained in an interval), but it may be inevitable in a fine-grained time interval (e.g., monthly). In this and similar situations, the baseline hazards function needs to be adjusted according to the interaction between the failure time and covariates. One potential approach to account for the interaction between the failure time and covariates is to incorporate time-dependent covariates, such as macroeconomic variables (Bellotti & Crook, 2009). One drawback of this approach is that it is an indirect adjustment, i.e., the effect of time-dependent covariates is applied on top of the baseline hazard instead of considering the interaction in the baseline hazard function.

In the proposed TDH model, we preserve the framework of Cox PH and accommodate the effect of covariates on default time in calculating the baseline survival function using the Accelerated Failure Time (AFT) model (Zhang & Peng, 2007). The difference between a Cox PH model and a parametric PH model is that Cox PH does not make any assumption about the baseline survival function $S_0(t)$, which is the non-parametric part of the Cox PH model. In typical estimations of the baseline survival function of Cox PH, such as the Breslow-type estimator (Peng, 2000) and smooth estimator (Ma, Heritier, & Lô, 2014), the time-related effect is out of consideration. A more flexible baseline survival function based on AFT, which takes into account the effect of covariates on default time, has the potential to give better prediction performance in the context of credit scoring.

The process of TDH can be divided into the following three steps:

(1) Estimating the borrower-specific survival levels over time, $SB_i(t|y_i = 1)$.
(2) Calculating the baseline survival levels, $S_0^{TDH}(t|y_i = 1)$ by averaging the borrower-specific survival levels over all observations at corresponding time intervals.
(3) Estimating the latency or conditional survival function $S_i^{TDH}(t|y_i = 1)$.

In the first step, we calculate the borrower-specific survival level over time for every borrower. The borrower-specific survival function of borrower i is estimated as:

$$SB_i(t|y_i = 1) = SB_0(t * \exp(-(\gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_m x_{im}))), \tag{4}$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_m)$ is a vector of regression parameters associated with $\boldsymbol{x}_i$, and SB_0 is the baseline survival function of AFT.

With the form of survival function, the first step of the proposed TDH model is able to capture the interaction between the effect of failure time and the effect of borrower specifics. In other words, the conditional survival function in Eq. (4) is able to generate different survival levels $SB_i(t|y_i = 1)$ for different borrowers and the time-related effect is directly manifested. In the standard Cox PH model, the covariates $\boldsymbol{x}_i$ appear in the exponent of the conditional survival function (3) only, while the time t affects the baseline survival function only. The baseline survival level $(S_0(t|y_i = 1))$ in (3) is time specific but the failure time is not related to the covariates.

At the second step, all observations are divided into groups according to observation time t, and then the baseline survival level at time t is calculated by taking the mean of borrower-specific survival levels over all observations observed at time t.

$$S_0^{TDH}(t|y_i=1)=\frac{1}{|O(t)|}\sum_{i\in O(t)}SB_i(t|y_i=1), \tag{5}$$

where $O(t)$ denotes the set of observations observed at time t.

Then, like in the Cox PH model, the conditional survival function is estimated as:

$$S_i^{TDH}(t|y_i=1)=S_0^{TDH}(t|y_i=1)^{\exp(\alpha_1x_{i1}+\alpha_2x_{i2}+\cdots+\alpha_mx_{im})}, \tag{6}$$

where $S_i^{TDH}(t|y_i=1)$ and $S_0^{TDH}(t|y_i=1)$ are the conditional survival function and the baseline survival levels, respectively, $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_m)$ is a vector of regression parameters associated with $\boldsymbol{x}_i$.

The conditional survival function $S_i^{TDH}(t|y_i=1)$ has the same form as that in the standard Cox PH model (3). However, the baseline survival function in the standard Cox PH model is independent of the covariates, whereas that in the proposed TDH model captures the interaction between the default time and the covariates through borrower-specific survival levels. This makes TDH more flexible than the standard Cox PH.

The parameters related to mixture cure models, with TDH in the latency component, are estimated using a framework of the expectation maximization (EM) algorithm (see details in Section 3.3). A partial likelihood method proposed by Peng (2000) is used to estimate the parameters $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_m)$ without specifying the baseline hazard function. With the partial likelihood of TDH, the estimation function for the latency component is similar to the log-likelihood function of the standard Cox PH with an additional offset variable $\log(E(y_i))$ (Cai, Zou, Peng, & Zhang, 2012). The estimation of $E(y_i)$ in an iterative process involves the survival function $S_i^{TDH}(t|y_i=1)$, which in turn involves the baseline survival function $S_0^{TDH}(t|y_i=1)$. Hence, the estimated parameters $\boldsymbol{\alpha}$ for TDH, as the latency component of mixture cure models, are generally different from those for the standard Cox PH.

A rank-based estimation method proposed by Zhang and Peng (2007) is used to estimate the parameters $\boldsymbol{\gamma}=(\gamma_0,\gamma_1,\ldots,\gamma_m)$. A derivation of the estimation process is provided in Appendix A. Note that the first step of our proposed TDH model is identical to the baseline survival function of AFT. However, unlike the accelerating failure time effect in AFT, where explanatory variables act as acceleration factors to speed up or slow down the survival process, the event of default on a fixed-term loan cannot occur after the end of the loan term, thus rendering AFT not directly applicable. In our proposed TDH model, AFT is used as an intermediate step to estimate the baseline survival function in the framework of Cox PH and is not needed when the model is applied out-of-sample.

Note that we focus on the event of default only when presenting the mixture cure model in this paper. However, in a real scenario, not only the event of default but also early repayment will lead to an "incomplete" repayment period. The mixture cure model can be extended to deal with the competing risk setting. For example, the mixture cure model can be performed separately on time until default T_1 and time until early repayment T_2, and then the predicted lifetime of the loan is estimated as $T=\min\{T_1, T_2, \text{term of the loan}\}$ (Banasik et al., 1999). Alternatively, the competing risk can also be accommodated by extending the single-event mixture cure model into a multiple-event mixture cure model (Dirick et al., 2015).

3.2. Random forests for the incidence component

For the incidence component, we substitute the typical logistic regression model with a more complex, and hopefully also more accurate, classifier. Many classification methods have been developed in the field of machine learning. In this paper, we consider random forests (Breiman, 2001) for two reasons. First, as a tree-based method, random forests can deal with potential nonlinear relationships among the attributes. Second, as a tree ensemble method, random forests may further improve performance over a single tree. Of course, empirical evaluation in the actual application is needed to test whether random forests outperform logistic regression in terms of classification accuracy and whether the higher accuracy, if any, leads to performance improvement for the entire mixture cure model.

Ensemble methods build an ensemble (or *committee*) of multiple classifiers, which may (or may not) perform better than any single classifier (Dietterich, 2000). For example, in credit scoring, Finlay (2011) found that ensemble classifiers often outperform single classifiers. Specifically in the context of P2P lending, Malekipirbazari and Aksakalli (2015) found that random forests outperformed several other classification methods, including logistic regression, in an empirical evaluation. Some ensemble methods, such as bagging and boosting, are general and work with any type of base classifiers (Dietterich, 2000). Others, such as random forests, work with particular types of base classifiers.

Random forests can be seen as an enhanced bagging method since it builds multiple classification and regression trees on bootstrapped samples, as is done in bagging. Different from bagging of trees, at each node while training the trees, the best splitting attribute is chosen from a randomly selected subset of z attributes (referred to as *random subspace*) rather than the full set of Z attributes. In other words, each tree can be seen as an expert in a narrow domain. It is expected that the aggregate performance of the committee of multiple experts (i.e., forest of trees) will exceed the performance of a single expert (i.e., a single tree). In each tree, splitting continues until the tree is grown to the maximum size and no further splits are possible. The outputs of all trees are aggregated to produce the final prediction.

Random forests also construct an out-of-bag estimate. When constructing a random forest from any training set, the entire training set is first divided into a series of bootstrapped training sets and about 1/3 of the cases are left out in each bootstrapped training set, which are called out-of-bag samples. The out-of-bag estimate for each case in the training set is the aggregation of the outputs of the trees that have not seen this case. Then, the out-of-bag estimate for the generalization error is calculated by comparing these out-of-bag estimates to the actual outcomes. It has been shown that unlike cross-validation, where bias is present but its extent is unknown, the out-of-bag estimates are unbiased (Breiman, 2001).

The effectiveness of random forests in prediction, compared to that of growing a single decision tree, can be explained through learning theory (i.e., bias-variance decomposition) (Svetnik et al., 2003). In order to avoid over-fitting and get the right model complexity for optimal prediction performance, pruning is usually needed for a single decision tree (i.e., a tradeoff between bias and variance). However, random forests grow an ensemble of decision trees without pruning (i.e., low bias, high variance), and then average them (i.e., lower variance). Mitigating the detrimental effect of variance through model averaging, random forests with unpruned trees reduce both bias and variance. Moreover, the reduction in variance depends on the correlation across the trees; lower correlation tends to lead to higher reduction in variance. Injecting the right kind of randomness through bootstrapping and random subspace, random forests reduce the correlation between the trees, resulting in even lower variance. In addition, as the Strong Law of Large Numbers shows, random forests always converge as the size (i.e., number of trees) increases so that overfitting is not a big problem.

3.3. Estimation

Given the observation data $(t_i, \delta_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$ for borrower i $(i = 1, 2, \ldots, N)$, we need to train the random forests and estimate the parameters of the TDH model. The complete likelihood function can be expressed as follows (Tong et al., 2012).

$$\prod_{i=1}^{N} [1 - p(\boldsymbol{z}_i)]^{1-y_i} \times p(\boldsymbol{z}_i)^{y_i} \times h(t_i|y = 1, \boldsymbol{x}_i)^{\delta_i y_i} \times S(t_i|y = 1, \boldsymbol{x}_i)^{y_i}, \tag{7}$$

where $p(\boldsymbol{z}_i) = p_i$ is the probability that borrower i will eventually default, and $h(t_i|y = 1, \boldsymbol{x}_i)$ is the hazard function corresponding to $S(t_i|y = 1, \boldsymbol{x}_i)$.

The log likelihood function can be written as the sum of two parts, each of which is related to only the incidence component or the latency component:

$$L_I = \sum_{i=1}^{N} (y_i \log[p(\boldsymbol{z}_i)] + (1 - y_i) \log[1 - p(\boldsymbol{z}_i)]), \tag{8}$$

$$L_L = \sum_{i=1}^{N} (y_i \delta_i \log[h(t_i|y = 1, \boldsymbol{x}_i)] + y_i \log[S(t_i|y = 1, \boldsymbol{x}_i)]). \tag{9}$$

It can be seen that y_i is partially observed because only δ_i is known from the observation data. A non-censored observation $(\delta_i = 1)$ corresponds to an observed default $(y_i = 1)$, but y_i is uncertain when an observation is censored $(\delta_i = 0)$. Hence, it is necessary to estimate the random variable y_i. The expectation maximization (EM) algorithm (Sy & Taylor, 2000), with its ability to estimate the maximum likelihood function with unobserved variables, can be adapted to estimate the conditional expectation of y_i. The expectation of y_i, $E(y_i)$, can be written as:

$$E(y_i|t_i, \delta_i, \boldsymbol{z}_i, \boldsymbol{x}_i) = \delta_i + (1 - \delta_i) \frac{p(\boldsymbol{z}_i) S(t_i|y = 1, \boldsymbol{x}_i)}{1 - p(\boldsymbol{z}_i) + p(\boldsymbol{z}_i) S(t_i|y = 1, \boldsymbol{x}_i)}. \tag{10}$$

EM is an iterative algorithm consisting of an expectation (E) step and a maximization (M) step. In the E-step, the expectation $E(y_i)$ in Eq. (10) is estimated. The next M-step is to maximize the likelihood function Eqs. (8) and (9) with respect to the unknown parameters. The two steps are iterated by replacing the estimated parameters in Eqs. (8) and (9) back into Eq. (10) until convergence on the estimation of the unknown parameters.

EM is a broadly applicable algorithm that provides an iterative procedure for computing Maximum Likelihood (ML) estimations. ML estimation in the standard mixture cure model (i.e., estimation for logistic regression and Cox PH), with the absence of some additional data (i.e., y_i), would be straightforward. Furthermore, the EM algorithm can be easily modified to produce the maximum a posteriori estimation (Nevat, Peters, & Yuan, 2008), the maximum penalized likelihood estimation (Liu, Levine, & Zhu, 2009), or even the minimum loss approximation, which is often needed in machine learning (Sela & Simonoff, 2012). For example, Hajjem, Bellavance, and Larocque (2011) presented an extension of regression tree (RT) algorithms, such as CART, to the case of clustered data. It is essentially an iterative call to a standard RT algorithm within the framework of the EM algorithm (Larocque, 2014). A possible generalization to a mixture cure model with random forests in the incidence component consists in replacing the logistic regression during each iteration with a forest of trees. Hence, the estimation of the proposed mixture cure model (i.e., RF-TDH) can be implemented within a framework of the EM algorithm. In the estimation process, like in the standard mixture cure model, the expectation $E(y_i)$ in Eq. (10) is first estimated. The random forests model and TDH model are then rebuilt based on the expectation $E(y_i)$. The two steps are iterated by replacing the estimated $p(\boldsymbol{z}_i)$ and $S(t_i|y = 1, \boldsymbol{x}_i)$ back into Eq. (10) until convergence.

It is worth to note that, unlike in the standard mixture cure model, where the estimated coefficients of variables can be straightforwardly used in the convergence indicator, there is no parameter directly quantifying convergence in random forests. We use the increment of the out-of-bag output of random forests (the out-of-bag samples are kept identical over all iterations) for indicating convergence. Let ${\boldsymbol{O}_{oob}}^{(r)}$ denote the out-of-bag output vector of random forests in the rth iteration, then the increment of the out-of-bag output vector of random forests, $\Delta {\boldsymbol{O}_{oob}}^{(r)}$, can be calculated as $({\boldsymbol{O}_{oob}}^{(r)} - {\boldsymbol{O}_{oob}}^{(r-1)})$. Same understanding applies to $\Delta {\boldsymbol{S}_0}^{(r)}$ and $\Delta \boldsymbol{\alpha}^{(r)}$, where $\boldsymbol{S}_0 = (S_0(t_1), S_0(t_2), \ldots, S_0(t_l))$ is the vector of the baseline survival levels, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ is a vector of regression parameters associated with $\boldsymbol{x}_i$. The discriminant value of convergence ($V_{convergernce}$) can be expressed as:

$${V_{convergernce}}^{(r)} = \left\| \Delta {\boldsymbol{O}_{oob}}^{(r)} \right\|_2^2 + \left\| \Delta {\boldsymbol{S}_0}^{(r)} \right\|_2^2 + \left\| \Delta \boldsymbol{\alpha}^{(r)} \right\|_2^2. \tag{11}$$

Pseudocode for the estimation process in the EM framework is given in Fig. 1.

4. Empirical evaluation

We have evaluated our proposed mixture cure model, in comparison with the standard model (using logistic regression and Cox PH in the incidence and latency components, respectively), using data collected from a major P2P lending institution in China. Credit risk evaluation for mitigating information asymmetry is extremely valuable in the P2P lending market. Hence, in this paper, we select P2P lending as our empirical evaluation context.

The experiments were performed in R version 3.3.1. The standard mixture cure model was fit using the "smcure" package proposed by Cai et al. (2012). The proposed model was implemented by modifying the "smcure" package, calling the random forests method in the "ranger" package. We used the "mlr" package for other methods involved in the empirical evaluation, including CART, bagging, and XGBoost.

4.1. Data

The data were collected from the P2P lending platform between January 2013 and December 2015. The dataset used in our evaluation consists of a sample of 52,573 12-month loans. The sample was collected from all borrowers who applied and subsequently got funded for a 12-month personal loan over the observation period. The attributes of application characteristics and historical repayment activities available in the dataset are described in Table 2. The platform assigns a Credit Grade to each borrower. For categorical attributes, some minority categories were merged into one category, such as "other" for occupation type and AA (AA+ and AA were merged into AA) for credit grade. A borrower was classified as default if the borrower had payment overdue for at least three months, and non-default otherwise. There are 6079 default observations and 46,494 non-default observations in the dataset. Since the dataset does not contain information about early repayment (data on early repayment, if any, could not be collected from the platform), we only consider the event of default in our empirical evaluation. A case, which may or may not have reached the end of the 12-month period, is treated as censored if no default is observed at the observation time. Table 3 shows the distribution of censored observations and non-censored observations at different times.

Input: Observation Data $(t_i, \delta_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$;
Convergence Threshold $\theta = 10^{-7}$;
Maximum Number of Iterations $R = 50$

Output: $p(\boldsymbol{z}_i)$ and $S(t_i|y = 1, \boldsymbol{x}_i)$

$y_i = \delta_i$

$\mathrm{RF} = \mathrm{RFTrain}(y_i \sim \boldsymbol{z}_i)$

$p(\boldsymbol{z}_i) = \mathrm{RFPredict}(\mathrm{RF}, \boldsymbol{z}_i)$

$S(t_i|y = 1, \boldsymbol{x}_i) = \mathrm{TDH}(y_i, t_i \sim \boldsymbol{x}_i)$

$V_{convergence} = \propto,\ r = 1$

while $V_{convergence} > \theta \ \wedge \ r < R$ **do**

$$\mathrm{E}(y_i)^{(r)} = \delta_i + (1 - \delta_i)\frac{p(\boldsymbol{z}_i)S(t_i|y=1,\boldsymbol{x}_i)}{1-p(\boldsymbol{z}_i)+p(\boldsymbol{z}_i)S(t_i|y=1,\boldsymbol{x}_i)}$$

$$\mathrm{RF} = \mathrm{RFTrain}(\mathrm{E}(y_i)^{(r)} \sim \boldsymbol{z}_i)$$

$$p(\boldsymbol{z}_i) = \mathrm{RFPredict}(\mathrm{RF}, \boldsymbol{z}_i)$$

$$S(t_i|y = 1, \boldsymbol{x}_i) = \mathrm{TDH}(\mathrm{E}(y_i)^{(r)}, t_i \sim \boldsymbol{x}_i)$$

$$\mathrm{V}_{convergernce} = \left\|\Delta \boldsymbol{O}_{oob}{}^{(r)}\right\|_2^{\ 2} + \left\|\Delta \boldsymbol{S}_0{}^{(r)}\right\|_2^{\ 2} + \left\|\Delta \boldsymbol{\alpha}^{(r)}\right\|_2^{\ 2}$$

$$r = r + 1$$

end while

Fig. 1. The estimation process in the EM framework.

Table 2
Attributes used in analysis.

No.	Attributes		Summary statistics		
	Continuous		Min	Max	Mean
1	Borrower age		20	64	38.52
2	Amount of loan		500	2,000,000	59719.87
3	Interest rate (%)		10	25	18.78
4	Number of successful loan applications		0	9	1.58
5	Number of failed loan applications		0	10	0.07
6	Number of loans paid off		0	5	0.78
	Categorical	Number of categories	Values		
7	Gender	2	{Male, female}		
8	Loan type	4	{PC, BG, CT, other}		
9	Annual income	6	{<2w, 2–6w, 6–12w, 12–24w, 24–40w, >40w}		
10	Credit line	7	{None, <3k, 3k–6k, 6k–2w, 2w–5w, 5w–10w, >10w}		
11	Credit grade	8	{AAA, AA, A, BB, B, CC, C, HR}		
12	Years with credit records	4	{No record, 1–3 years, 3–5 years, >5 years}		
13	Insurance status	2	{Yes, no}		
14	Credit report	2	{Yes, no}		
15	Guarantor	2	{Yes, no}		
16	House guarantee	2	{Yes, no}		
17	Car guarantee	2	{Yes, no}		
18	Living area	5	{RR, UC, CIC, COC, other}		
19	Home ownership	6	{R/N, PH, ME<40w, ME>40w, ON<100w, ON>100w}		
20	Years with employer	5	{Unemployed, <1 year, 1–3 years, 3–5 years, >5 years}		
21	Job title	7	{S/U, LS, JE, ME, SE, BS, other}		
22	Occupation type	6	{S/J/U, COS, CIS, PI, FI, other}		
23	Education level	5	{JHS or below, SHS/TSS, JC, BR, MS or above}		
24	Marital status	4	{UM, MC, MWC, DD}		
25	Years with social security	5	{Unpaid, <1 year, 1–3 years, 3–5 years, >5 years}		

*The unit of money is RMB (¥), 1k = ¥1000, 1w = ¥10,000.
**Abbreviations: PC = personal consumption, BG = borrowing, CT = capital turnover; RR = rural residential, UC = urban community, CIC = city community, COC = commercial community; R/N = renting/no house, PH = parent's house, ME = mortgage, ON = own; S/U = student/ uncertain title, LS = laborious staff, JE = junior executive, ME = middle executive, SE = senior executive, BS = business; S/J/U = student/ job-waiting/ unemployed, COS = company staff, CIS = civil servant, PI = public institution, FI = financial institution; JHS = Junior high school, SHS/TSS = senior high school/ technical secondary school, JC = junior college, BR = bachelor, MS = master; UM = unmarried, MC = married childless, MWC = married with children, DD = divorced.

Table 3
Numbers of censored and non-censored observations.

Observation time	Censored	Non-censored	Sum
1	4897	0	4897
2	4204	0	4204
3	7365	1068	8433
4	4553	706	5259
5	4540	826	5366
6	3800	670	4470
7	3506	496	4002
8	3674	523	4197
9	2400	467	2867
10	2530	491	3021
11	815	204	1019
12	4210	628	4838
Sum	46,494	6079	52,573

4.2. Weight of evidence transformation

The dataset contains both continuous attributes and categorical attributes. Categorical attributes are normally recoded into several dichotomous (dummy) variables in statistic models, such as logistic regression. A tree-based method, such as random forests, deals with a categorical attribute through tree splitting and thus does not need dummy encoding. In order to keep the input attributes and their values identical across methods for a fair comparison, we used the Weight of Evidence (WOE) transformation in a preprocessing step. WOE is an alternative attribute transformation method and has been shown to be superior to dummy encoding in several studies (Moeyersoms & Martens, 2015).

The WOE value of the kth category (interval) in a categorical (continuous) attribute is calculated as:

$$\mathrm{WOE}_k = \ln\left(\frac{Bad\ distribution_k}{Good\ distribution_k}\right), \tag{12}$$

where $Bad\ distribution_k$ is the proportion of defaulting borrowers in the kth category (interval) among all defaulting borrowers ($Bad\ distribution_k = \frac{Number\ of\ Defaulters_k}{Total\ Number\ of\ Defaulters}$), and $Good\ distribution_k$ is the proportion of non-default borrowers in the kth category (interval) among all non-default borrowers ($Good\ distribution_k = \frac{Number\ of\ Non-defaulters_k}{Total\ Number\ of\ Non-defaulters}$). We used the "woe" package in R with the default value (10) for the parameter "count of bins to be computed".

4.3. Pre-evaluation of classification methods

Before evaluating our proposed mixture cure model, we first tested whether random forests outperform logistic regression in a binary classification problem (i.e., classifying a borrower as either a defaulter or not). We also tested whether bagging and boosting (Dietterich, 2000), which are general ensemble methods, improve performance of logistic regression and decision tree. We therefore included the following seven methods in the comparison: logistic regression (LR), decision tree (DT), bagging of logistic regression (Bagging-LR), boosting of logistic regression (Boosting-LR), bagging of decision tree (Bagging-DT), boosting of decision tree (Boosting-DT), and random forests (RF). We used the CART algorithm for decision tree and the XGBoost algorithm for boosting.

We used the default settings for the single classifier methods in the R packages. For example, by default, the logistic regression method uses all attributes without attribute selection, and the parameter z for random forests is $Z/3$. For the ensemble methods, including bagging, XGBoost, and random forests, we tuned some hyper-parameters through a preliminary experiment. We tuned each parameter using a grid search and chose a value that seemed to give the best performance (in terms of AUC) or when the performance seemed to have stopped improving. The performance under each parameter setting was estimated using 10 rounds of training and validation (90% of the data randomly selected for training and 10% for validation). The number of trees for each model was chosen using grid search from 10 to 500 in increment of 10. The proportion of training examples or attributes (if applicable) for each base classifier was chosen using grid search from 0.1 to 1.0 in increment of 0.1. The chosen parameter values are as follows. For random forests, the number of trees was set to 500, as the performance estimate had converged before the number of trees reached 500. For XGBoost, the number of iterations was set to 100. For Bagging-LR, the number of base classifiers was set to 500, the proportion of training examples for each base classifier was set to 0.1, and the proportion of attributes for each base classifier was set to 0.7. For Bagging-DT, the number of base classifiers was set to 500, the proportion of training examples for each base classifier was set to 0.1, and the proportion of attributes for each base classifier was set to 1.0. While there are many hyper-parameters related to random forests and XGBoost, we only tuned some necessary hyper-parameters using grid search and retained the default values recommended by the corresponding packages for the other hyper-parameters[1,2]. Further tuning on the other hyper-parameters might further improve the performance slightly. However, the current parameter settings could already reveal the advantage of random forests over logistic regression in this evaluation, although the comparison between random forests and XGBoost might not be conclusive.

Table 4
Performance of classification methods.

Method	Mean AUC	(95% confidence interval)
LR	0.705	(0.702–0.707)
DT	0.673	(0.670–0.676)
Bagging-LR	0.675	(0.673–0.678)
Bagging-DT	0.699	(0.696–0.701)
Boosting-LR	0.705	(0.703–0.707)
Boosting-DT	0.716	(0.714–0.718)
RF	**0.733**	(0.730–0.735)

After the parameter tuning, we performed 10 independent 10-fold cross validations to estimate the performance (AUC) of each classification method (Finlay, 2012; Molinaro, Simon, & Pfeiffer, 2005), resulting in 100 performance estimates. Performance results (mean and 95% confidence interval) reported later are based on the 100 estimates. During each 10-fold cross validation, the dataset was divided into 10 equal-sized subsets (called folds). Each fold was used to estimate the performance of the classifier trained on the other 9 folds. The splitting of folds was kept identical across classification methods. The results are summarized in Table 4. A single decision tree was not more accurate than logistic regression. Bagging and boosting of logistic regression was not productive; the ensemble classifiers were not more accurate than a single logistic regression classifier. Bagging and boosting significantly improved the performance of decision tree. Random forests outperformed all base and ensemble methods. In particular, the AUC of random forests was statistically significantly higher than that of logistic regression according to the estimated confidence interval. The results confirmed that random forests indeed outperformed logistic regression in the binary classification problem.

Note that in order to have a fair comparison between random forests and logistic regression without other confounding factors that may influence performance, we kept the input attributes and

[1] http://xgboost.readthedocs.io/en/latest/parameter.html (XGBoost).
[2] https://cran.r-project.org/web/packages/ranger/index.html (Random Forests).

Table 5
LR with/without attribute selection.

	Mean AUC	(95% confidence interval)
Without attribute selection	0.705	(0.702–0.707)
With attribute selection	0.705	(0.702–0.707)

Table 6
RF with/without WOE transformation.

	Mean AUC	(95% confidence interval)
Without transformation	0.737	(0.735–0.740)
With transformation	0.733	(0.730–0.735)

their values identical across methods. However, there are other factors that may influence performance. For example, attribute selection may influence the performance of logistic regression, while random forests deal with categorical attributes naturally and may actually work better on the original attributes without transformation. As a robustness check, we also examined some of these factors. We compared logistic regression with and without attribute selection. The attribute selection was done using a backward stepwise AIC selection method (Yamashita, Yamashita, & Kamimura, 2007). We also compared random forests with and without the WOE transformation. The results are summarized in Tables 5 and 6, respectively. The influence of attribute selection on logistic regression was negligible (the difference in AUC was less than 0.001). The performance of random forests on the original attributes without transformation was marginally better (the difference in AUC was 0.004). The robustness check results further confirm that random forests indeed outperformed logistic regression in this evaluation.

4.4. Evaluation of proposed mixture cure model

After confirming that random forests indeed outperformed logistic regression in binary classification, we then evaluated our proposed mixture cure model, in comparison with the standard mixture cure model. To test whether and how much each of our proposed extensions (random forests in place of logistic regression in the incidence component and TDH in place of Cox PH in the latency component) affects the performance, we used a full factorial design, including the following settings: logistic regression and Cox PH (LR-Cox), logistic regression and TDH (LR-TDH), random forests and Cox PH (RF-Cox), and random forests and TDH (RF-TDH). LR-Cox corresponds to the standard mixture cure model (Tong et al., 2012), whereas RF-TDH corresponds to our proposed model. In random forests, we again used 500 trees. Parameters of all models were estimated using the EM framework.

We compared the methods in terms of their ability to predict the monthly PD over the one-year loan term. Note that with our definition of default (i.e., if a borrower had payment overdue for at least three months), defaulting would not happen in the first two months. Therefore, the one-year loans have 10 months, denoted T_1 to $-T_{10}$, to be predicted (note that T_1 corresponds to the third month, T_2 corresponds to the fourth month, and so on). The score function for each month, t ($t = 3, 4, \ldots, 12$), was derived as $1 - S(t)/S(t-1)$. Most previous studies (e.g., Tong et al., 2012) in traditional loan settings considered the yearly PD (e.g., estimating the PD in the second year of the loan term given that the borrower has not defaulted in the first year). In P2P lending, month is a more appropriate forecast cycle, as the loan term is typically shorter than bank loans and most loans are repaid in monthly installments.

We examined both discrimination performance (the ability to risk rank borrowers accurately) and calibration performance (the accuracy of the PD estimates themselves) (Tong et al., 2012). We used AUC, Kolmogorov–Smirnov (KS) statistic, and H measure (Hand, 2009) as the discrimination performance measures (Tong et al., 2012). For calibration performance, we examined the agreement between the survival probability (or the number of defaulters) predicted by each model and the empirical survival proportion according to the Kaplan–Meier (KM) estimator.

We again estimated the performance of each model using 10 independent 10-fold cross validations, resulting in 100 performance estimates. Performance results (mean and 95% confidence interval) reported later are all based on the 100 estimates.

5. Results

5.1. Discrimination performance

Table 7 summarizes the discrimination performance of the four mixture cure models. The third column (Time) is the predicting time. The results have very similar patterns across the three performance measures, showing that the results are quite robust. In all predicting months (T_1 to $-T_{10}$), the proposed RF-TDH outperformed all other alternatives in terms of all three performance measures. Specifically, using logistic regression for the incidence component, TDH outperformed Cox PH in all months (i.e., LR-TDH vs. LR-Cox). Using random forests for the incidence component, TDH outperformed Cox PH in all months (i.e., RF-TDH vs. RF-Cox). Using Cox PH for the latency component, random forests outperformed logistic regression in all months (i.e., RF-Cox vs. LR-Cox). Using TDH for the latency component, random forests outperformed logistic regression in all months (i.e., RF-TDH vs. LR-TDH). The results show that both RF and TDH contributed to improvement in discrimination performance.

To analyze the influence of the choice of the incidence model (LR or RF) and the latency model (Cox or TDH) on the overall discrimination performance, we used the repeated-measure ANOVA with the incidence method (LR or RF) and the latency method (Cox or TDH) as two main factors and the predicting month as a between-subject factor. Although ANOVA is a parametric test, it allows us to examine not only the main effects of the factors but also their interaction effects. Of course, when interpreting the results of ANOVA, we should keep in mind its drawbacks (Mchugh, 2011), such as the assumption of normality, the assumption of equal population means from each group, and the assumption of equal variances from each group. Since the discrimination performance of each model is measured on 10 independent groups based on the observation month (month 3–12), the predicting month (T_1 to $-T_{10}$) is a between-subject factor. Table 8 summarizes the results of repeated-measure ANOVA. Fig. 2 visualizes the marginal effects of the two main factors. The effects of both factors in terms of all three performance measures were statistically significant ($p < 0.001$), providing statistically significant evidence for improvement in discrimination performance due to RF and TDH. The interaction between the two factors was also statistically significant for AUC and KS, but the effect size was very small (partial $\eta^2 < 0.1$).

5.2. Calibration performance

The calibration performance of each method was measured using the KM estimator and the concordance correlation coefficient. The KM estimator is a widely used descriptive estimator. We estimated the survival function of the test group by the KM estimator using the predicted number of defaulters and observed number of defaulters separately. We calculated the predicted number of defaulters by summing up the predicted probability (Liu et al., 2015), avoiding the need for a discriminative threshold on each observation. Fig. 3 shows the agreement between the predicted survival

Table 7
Discrimination performance of mixture cure models.

Incidence	Latency	Time	AUC (95%CI)	KS (95%CI)	H (95%CI)	Score function
LR	Cox	T_1	0.683(0.676–0.690)	0.317(0.308–0.326)	0.179(0.171–0.187)	1-S(3)
	TDH		0.730(0.725–0.735)	0.361(0.352–0.369)	0.227(0.219–0.235)	
RF	Cox		0.708(0.701–0.714)	0.339(0.329–0.349)	0.200(0.191–0.208)	
	TDH		**0.739**(0.734–0.745)	**0.386**(0.377–0.396)	**0.242**(0.233–0.251)	
LR	Cox	T_2	0.670(0.662–0.677)	0.293(0.282–0.304)	0.172(0.163–0.181)	1-S(4)/S(3)
	TDH		0.730(0.722–0.739)	0.385(0.371–0.399)	0.248(0.235–0.261)	
RF	Cox		0.680(0.672–0.687)	0.302(0.291–0.313)	0.182(0.173–0.191)	
	TDH		**0.735**(0.727–0.742)	**0.389**(0.377–0.401)	**0.255**(0.244–0.266)	
LR	Cox	T_3	0.627(0.618–0.636)	0.254(0.242–0.266)	0.123(0.114–0.132)	1-S(5)/S(4)
	TDH		0.690(0.682–0.699)	0.354(0.342–0.367)	0.192(0.182–0.203)	
RF	Cox		0.657(0.648–0.666)	0.290(0.278–0.303)	0.152(0.142–0.163)	
	TDH		**0.711**(0.704–0.719)	**0.360**(0.347–0.372)	**0.218**(0.207–0.228)	
LR	Cox	T_4	0.694(0.684–0.703)	0.346(0.330–0.361)	0.202(0.189–0.215)	1-S(6)/S(5)
	TDH		0.696(0.687–0.705)	0.352(0.338–0.366)	0.203(0.192–0.214)	
RF	Cox		0.730(0.721–0.738)	0.393(0.379–0.407)	0.257(0.245–0.270)	
	TDH		**0.741**(0.733–0.750)	**0.414**(0.399–0.428)	**0.273**(0.259–0.286)	
LR	Cox	T_5	0.681(0.671–0.691)	0.357(0.342–0.373)	0.197(0.182–0.211)	1-S(7)/S(6)
	TDH		0.704(0.695–0.713)	0.390(0.375–0.405)	0.222(0.209–0.235)	
RF	Cox		0.707(0.697–0.717)	0.376(0.362–0.391)	0.232(0.218–0.246)	
	TDH		**0.735**(0.726–0.744)	**0.421**(0.406–0.436)	**0.275**(0.260–0.289)	
LR	Cox	T_6	0.695(0.687–0.703)	0.367(0.354–0.379)	0.204(0.193–0.216)	1-S(8)/S(7)
	TDH		0.710(0.703–0.718)	0.372(0.360–0.384)	0.222(0.212–0.233)	
RF	Cox		0.731(0.722–0.739)	0.403(0.389–0.416)	0.256(0.244–0.268)	
	TDH		**0.748**(0.740–0.756)	**0.419**(0.406–0.432)	**0.286**(0.274–0.299)	
LR	Cox	T_7	0.675(0.662–0.688)	0.345(0.327–0.364)	0.212(0.195–0.228)	1-S(9)/S(8)
	TDH		0.685(0.673–0.696)	0.363(0.345–0.381)	0.219(0.203–0.235)	
RF	Cox		0.717(0.707–0.728)	0.387(0.371–0.403)	0.259(0.244–0.273)	
	TDH		**0.725**(0.714–0.736)	**0.397**(0.381–0.414)	**0.269**(0.254–0.284)	
LR	Cox	T_8	0.707(0.697–0.717)	0.391(0.375–0.407)	0.241(0.226–0.255)	1-S(10)/S(9)
	TDH		0.715(0.705–0.725)	0.406(0.390–0.422)	0.258(0.244–0.272)	
RF	Cox		0.730(0.720–0.741)	0.412(0.396–0.428)	0.271(0.256–0.287)	
	TDH		**0.741**(0.731–0.751)	**0.426**(0.410–0.441)	**0.290**(0.276–0.305)	
LR	Cox	T_9	0.676(0.656–0.697)	0.416(0.388–0.444)	0.284(0.257–0.311)	1-S(11)/S(10)
	TDH		0.701(0.681–0.722)	0.455(0.426–0.484)	0.321(0.292–0.351)	
RF	Cox		0.701(0.680–0.722)	0.443(0.416–0.470)	0.319(0.292–0.347)	
	TDH		**0.728**(0.708–0.748)	**0.475**(0.447–0.503)	**0.358**(0.329–0.387)	
LR	Cox	T_{10}	0.641(0.632–0.650)	0.284(0.269–0.298)	0.137(0.127–0.147)	1-S(12)/S(11)
	TDH		0.701(0.692–0.710)	0.377(0.363–0.392)	0.220(0.207–0.232)	
RF	Cox		0.739(0.730–0.747)	0.420(0.406–0.433)	0.294(0.282–0.306)	
	TDH		**0.751**(0.742–0.760)	**0.437**(0.424–0.450)	**0.337**(0.325–0.349)	

Table 8
Results of repeated-measure ANOVA.

Source	AUC			KS			H measure		
	F	*p*	*Partial* η^2	*F*	*p*	*Partial* η^2	*F*	*p*	*Partial* η^2
Incidence	1094.6	<0.001	0.525	462.3	<0.001	0.318	990.5	<0.001	0.500
Latency	1721.4	<0.001	0.635	1110.5	<0.001	0.529	1412.6	<0.001	0.588
Incidence*Latency	88.1	<0.001	0.082	30.1	<0.001	0.029	0.1	0.814	0.000

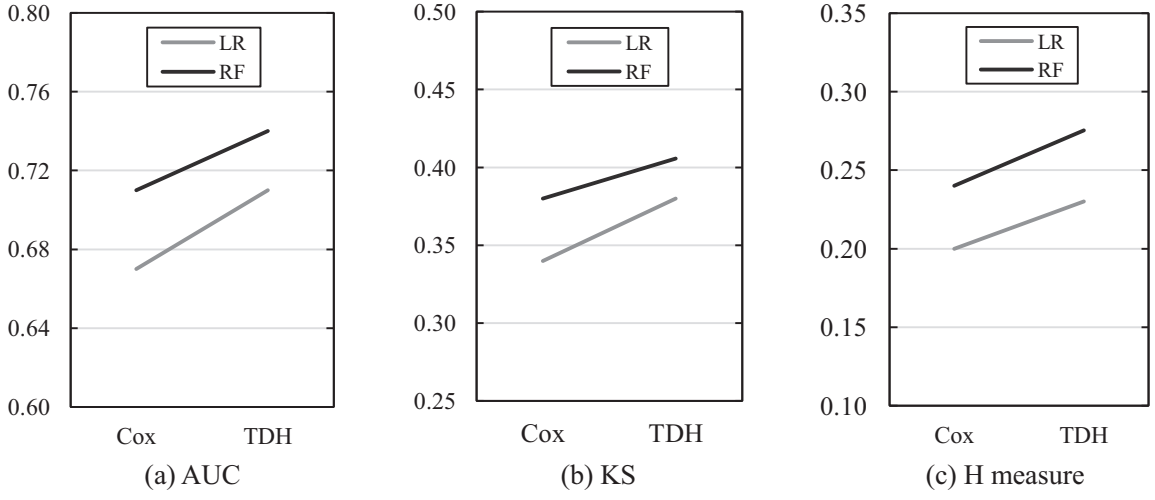

Fig. 2. The effects of incidence method and latency method on discrimination performance.

Table 9
Concordance correlation coefficient.

	LR-Cox	LR-TDH	RF-Cox	RF-TDH
Concordance correlation coefficient	0.921	0.935	0.930	0.944
95% confidence interval	(0.913–0.929)	(0.928–0.942)	(0.922–0.937)	(0.937–0.949)
Pearson ρ	0.966	0.974	0.969	0.977
Bias correlation factor C_b	0.954	0.960	0.960	0.966

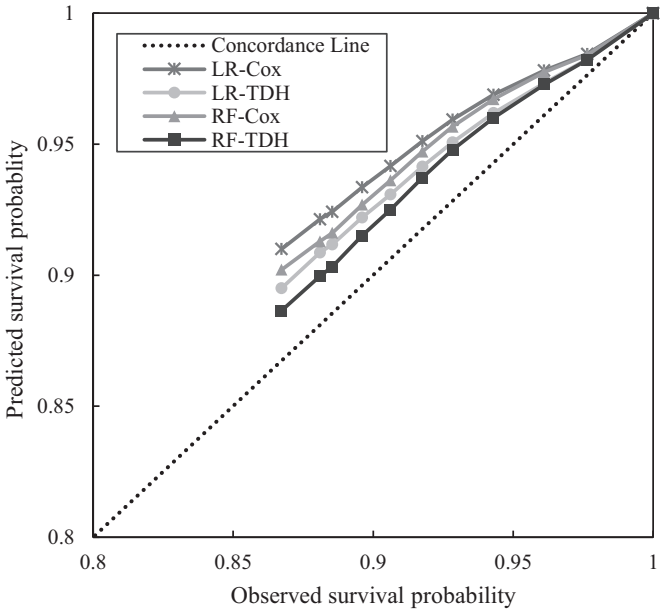

Fig. 3. Agreement between observed and predicted survival probabilities.

probability and the observed survival probability (based on the 100 estimates of repeated cross validations). Overall, the proposed RF-TDH outperformed all other alternatives (i.e., was the closest to the concordance line) throughout the loan term in both cases. Specifically, using either LR or RF in the incidence component, the proposed TDH outperformed Cox PH throughout the loan term. Using either Cox PH or TDH in the latency component, RF outperformed LR throughout the loan term.

Table 9 summarizes the results of concordance correlation between the predicted number of defaulters and the observed number of defaulters. The concordance correlation coefficient ρ_c is associated with a measurement of precision ρ and accuracy C_b; $\rho_c = \rho C_b$ (Lin, 1989). Pearson correlation coefficient ρ measures how far each pair of observed and predicted numbers of defaulters deviates from the best-fit line. C_b is a bias correction factor measuring how far the best-fit line deviates from the 45° line through the origin (i.e., the concordance line). The proposed RF-TDH outperformed all other alternatives (i.e., had the highest concordance correlation with the observed number of defaulters). Specifically, using either LR or RF in the incidence component, the proposed TDH outperformed Cox PH. Using either Cox PH or TDH in the latency component, RF outperformed LR. The results show that both our proposed extensions contributed to improvement in calibration performance too.

6. Conclusion

We have proposed a prediction-driven mixture cure model and applied it to predict time-dependent probability of loan default. We proposed extensions to the standard mixture cure model in both the incidence and latency components, which relax the assumptions of the standard methods and sacrifice interpretability for potential improvement in prediction performance. Empirical evaluation on an actual dataset from a major P2P lending institution in China shows that both extensions indeed contributed to prediction performance improvement.

Our work has several limitations, which may be addressed in future research. First, as discussed earlier, we did not consider the competing risk of early repayment since our dataset does not contain information on early repayment. In practice, the adjustment for early repayment is necessary because a high rate of early repayment would decrease the number of defaults observed in a portfolio. Models that do not account for such a competing risk may overestimate the number of defaults. Future research may extend our proposed model to further consider the competing risk of early repayment. Second, we only evaluated random forests, as an example, for the incidence component of a mixture cure model. Many classification methods have been developed in the field of machine learning. Some of them may be adapted and tested in future research. Third, we only evaluated our method in one dataset from one P2P lending institution in China, future research may conduct a more comprehensive evaluation using multiple datasets collected from multiple institutions in multiple countries to validate the generalizability of our findings. Finally, the proposed mixture cure model is general and may be applied and evaluated in other contexts (e.g., traditional bank loans) in the future to validate its applicability.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Grant nos. 71731005, 71571059) and the Humanities and Social Sciences Fund Research Planning of the Ministry of Education (Grant no. 15YJA630010). We are grateful to the anonymous reviewers for their constructive feedback, which helped us improve the quality of the paper considerably.

Appendix A. Estimation of the parameters in the borrower-specific survival function of TDH

The rank-based method for estimating the parameters in the borrow-specific survival functions of TDH, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_m)$, is based on Zhang and Peng (2007). We briefly outline the estimation process below. For more details, please refer to Zhang and Peng (2007).

Let ε_i denote $\log t_i - \boldsymbol{\gamma} \cdot \boldsymbol{x}_i$. The relationship between ε_i and the borrower-specific survival function (Eq. (4)) can then be derived from:

$$P(T > t) = P\left(e^{\boldsymbol{\gamma}\cdot\boldsymbol{x}_i+\varepsilon_i} > t\right) = P\left(e^{\varepsilon_i} > t * e^{-\boldsymbol{\gamma}\cdot\boldsymbol{x}_i}\right) = SB_0\left(t * e^{-\boldsymbol{\gamma}\cdot\boldsymbol{x}_i}\right). \tag{13}$$

Then the log failure time can be expressed as:

$$\log(T_i) = \boldsymbol{\gamma} \cdot \boldsymbol{x_i} + \varepsilon_i^*, \tag{14}$$

where the hazard function of ε_i^* is $E(y_i)^* h(\varepsilon_i^*)$, $E(y_i)$ denotes the expectation of y_i (its estimation is described in Section 3.3), and $h(\cdot)$ denotes the hazards function.

The $h(\varepsilon_i^*)$ can be estimated using the rank-based estimation method from the PH model:

$$h_{PH}(\varepsilon_i^*) = E(y_i)h(\varepsilon_i^*)\exp\left(\boldsymbol{\theta}\cdot\boldsymbol{x}_i\right). \tag{15}$$

The derivative of the logarithm of the partial likelihood function for (15) with respect to $\boldsymbol{\theta}$ is:

$$\psi\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{N}\delta_i\left(\boldsymbol{x}_i - \frac{\sum_{j=1}^{N}\boldsymbol{x}_j E\left(y_j\right)e^{\boldsymbol{\theta}\cdot\boldsymbol{x}_j}I\left(\varepsilon_j^* \geq \varepsilon_i^*\right)}{\sum_{j=1}^{N}E\left(y_j\right)e^{\boldsymbol{\theta}\cdot\boldsymbol{x}_j}I\left(\varepsilon_j^* \geq \varepsilon_i^*\right)}\right), \tag{16}$$

where $I(\cdot)$ is the indicator function. If the parameter $\boldsymbol{\theta}$ is 0, $\psi(0) = 0$ can be used as a linear rank-based estimation equation for $\boldsymbol{\gamma}$. Thus, $\psi(0)$ can be rewritten as $\psi(\boldsymbol{\gamma}, k(\cdot))$:

$$\psi(\boldsymbol{\gamma}, k(\cdot)) = \sum_{i=1}^{N}\delta_i k(\varepsilon_i^*)\left(\boldsymbol{x}_i - \frac{\sum_{j=1}^{N}\boldsymbol{x}_j E\left(y_j\right)I\left(\varepsilon_j^* \geq \varepsilon_i^*\right)}{\sum_{j=1}^{N}E\left(y_j\right)I\left(\varepsilon_j^* \geq \varepsilon_i^*\right)}\right), \tag{17}$$

where $k(\cdot)$ is a weight function which is predictable. Fygenson and Ritov (1994) showed that the estimation Eq. (16) is monotone under the Gehan weight function: $k(u) = \sum_{j=1}^{N} I(\varepsilon_j^* \geq u)/N$.

Then the estimation Eq. (17) can be simplified under the definition of a Gehan-type weight function:

$$\psi(\boldsymbol{\gamma}, k(\cdot)) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_i\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)E\left(y_j\right)I\left(\varepsilon_j^* \geq \varepsilon_i^*\right). \tag{18}$$

Another advantage of the defined Gehan-type weight function is that the estimation Eq. (18) can be taken as the gradient of a convex function (Jin, Lin, Wei, & Ying, 2003):

$$L_G(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_i E\left(y_j\right)\left|\varepsilon_i^* - \varepsilon_j^*\right|I(\varepsilon_i^* < \varepsilon_j^*). \tag{19}$$

Hence, finding the root of $\psi(\boldsymbol{\gamma}, k(\cdot)) = 0$ is equivalent to minimizing this convex function, which can be conveniently carried out by the linear programming method.

Given the estimate of the $\boldsymbol{\gamma}$ from the above procedure, the survival function $SB_0(\varepsilon|y=1)$ can be estimated using the Breslow log-likelihood function (Breslow, 1974). Let $\tau_1 < \tau_2 < \ldots < \tau_k$ be the distinct uncensored failure residuals ε_i, t_i denote the observation time of borrower i, d_{τ_j} denote the number of failures, and $R(\tau_j)$ denote the risk set at τ_j. An estimator of $SB_0(\varepsilon|y=1)$ is given by:

$$SB_0(\varepsilon|y=1) = \exp\left(-\sum_{j:\tau_j\leq\varepsilon}\frac{d_{\tau_j}}{\sum_{i\epsilon R(\tau_j)}E(y_i)}\right). \tag{20}$$

Because the estimator $SB_0(\varepsilon|y=1)$ may not approach 0 as $t \to \infty$, we set $SB_0(\varepsilon|y=1) = 0$ when $\varepsilon > \tau_k$.

References

Alves, B. C., & Dias, J. G. (2015). Survival mixture models in behavioral scoring. *Expert Systems with Applications, 42*(8), 3902–3910.

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.

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.

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

Beran, J., & Djaïdja, A. Y. K. (2007). Credit risk modeling based on survival analysis with immunes. *Statistical Methodology, 4*(3), 251–276.

Bhattacharya, A., Wilson, S. P., & Soyer, R. (2019). A Bayesian approach to modeling mortgage default and prepayment. *European Journal of Operational Research, 274*(3), 1112–1124.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

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

Cai, C., Zou, Y., Peng, Y., & Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine, 108*(3), 1255–1260.

Chang, Y. C., Chang, K. H., Chu, H. H., & Tong, L. I. (2016). Establishing decision tree-based short-term default credit risk assessment models. *Communications in Statistics-Theory and Methods, 45*(23), 6803–6815.

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

De Leonardis, D., & Rocci, R. (2014). Default risk analysis via a discrete-time cure rate model. *Applied Stochastic Models in Business & Industry, 30*(5), 529–543.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of international workshop on multiple classifier systems* (pp. 1–15). Heidelberg, Berlin.

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.

Djeundje, V. B., & Crook, J. (2018). Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. *European Journal of Operational Research, 271*(2), 697–709.

Djeundje, V. B., & Crook, J. (2019). Dynamic survival models with varying coefficients for credit risks. *European Journal of Operational Research, 275*(1), 319–333.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, 210*(2), 368–378.

Finlay, S. (2012). *Credit scoring, response modelling and insurance rating: A practical guide to forecasting consumer behavior*. Palgrave Macmillan.

Fygenson, M., & Ritov, Y. (1994). Monotone estimating equations for censored data. *Annals of Statistics, 22*(2), 732–746.

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research, 249*(2), 417–426.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523–541.

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.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning, 77*(1), 103–123.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters, 81*(4), 451–459.

Huang, Z., Zhao, H., & Zhu, D. (2012). Two new prediction-driven approaches to discrete choice prediction. *ACM Transactions on Management Information Systems (TMIS), 3*(2), 9.

Im, J. K., Apley, D. W., Qi, C., & Shan, X. (2012). A time-dependent proportional hazards survival model for credit risk analysis. *Journal of the Operational Research Society, 63*(3), 306–321.

Jin, Z., Lin, D. Y., Wei, L. J., & Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika, 90*(2), 341–353.

Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation & Simulation, 84*(6), 1313–1328.

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 1–32.

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

Liu, F., Hua, Z., & Lim, A. (2015). Identifying future defaulters: A hierarchical Bayesian method. *European Journal of Operational Research, 241*(1), 202–211.

Liu, L., Levine, M., & Zhu, Y. (2009). A functional EM algorithm for mixing density estimation via nonparametric penalized likelihood maximization. *Journal of Computational & Graphical Statistics, 18*(2), 481–504.

Ma, J., Heritier, S., & Lô, S. N. (2014). On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. *Computational Statistics & Data Analysis, 74*(5), 142–156.

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications, 42*(10), 4621–4631.

Mchugh, M. L. (2011). Multiple comparison analysis testing in ANOVA. *Biochemia Medica, 21*(3), 203–209.

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems, 72*(8), 72–81.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics, 21*(15), 3301–3307.

Narain, B. (1992). Survival analysis and the credit granting decision. *Credit scoring and credit control* (pp. 109–121). Oxford: Oxford University Press.

Nevat, I., Peters, G. W., & Yuan, J. (2008). Maximum a-posteriori estimation in linear models with a random Gaussian model matrix: A Bayesian-EM approach. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing* (pp. 2889–2892).

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

Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: A survey. *Operations Research, 42*(4), 589–613.

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning, 86*(2), 169–207.

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

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

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification

and QSAR modeling. *Journal of Chemical Information and Computer Sciences, 43*(6), 1947–1958.
Tong, E. N. C., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research, 218*(1), 132–139.
Yamashita, T., Yamashita, K., & Kamimura, R. (2007). A stepwise AIC method for variable selection in linear regression. *Communications in Statistics - Theory and Methods, 36*(13), 2395–2403.
Zhang, J., & Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine, 26*(16), 3157–3171.
Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting, 28*(1), 204–215.