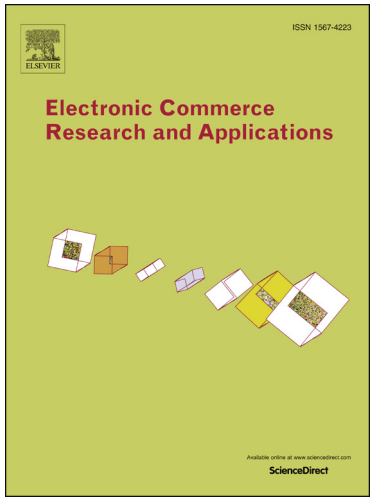


A novel behavioral scoring model for estimating probability of default over time in Peer-to-Peer lending

Zhao Wang, Cuiqing Jiang, Yong Ding, Xiaozhong Lv, Yao Liu

PII: S1567-4223(17)30099-6
DOI: https://doi.org/10.1016/j.elerap.2017.12.006
Reference: ELERAP 749

To appear in: *Electronic Commerce Research and Applications*

Received Date: 12 September 2017
Revised Date: 18 December 2017
Accepted Date: 18 December 2017

Please cite this article as: Z. Wang, C. Jiang, Y. Ding, X. Lv, Y. Liu, A novel behavioral scoring model for estimating probability of default over time in Peer-to-Peer lending, *Electronic Commerce Research and Applications* (2017), doi: https://doi.org/10.1016/j.elerap.2017.12.006

A NOVEL BEHAVIORAL SCORING MODEL FOR ESTIMATING PROBABILITY OF DEFAULT OVER TIME IN PEER-TO-PEER LENDING

Zhao Wang [a], Cuiqing Jiang (corresponding author) [a], Yong Ding [a], Xiaozhong Lv [a], Yao Liu [a]

[a] School of Management, Hefei University of Technology, Hefei 230009, Anhui, P.R. China

Last revised: December 18, 2017

ABSTRACT

Traditional behavioral scoring models applying classification methods that yield a static probability of default may ignore the borrowers' dynamic characteristics because borrower repayment behavior evolves dynamically. In this study, we propose a novel behavioral scoring model based on a mixture survival analysis framework to predict the dynamic probability of default over time in peer-to-peer (P2P) lending. A random forest is utilized to identify whether a borrower will default, and a random survival forest is introduced to model the time to default. The results of an empirical analysis on a Chinese P2P loan dataset show that the proposed *ensemble mixture random forest* (EMRF) has a better performance in terms of predicting the monthly dynamic probability of default, while compared with standard mixture cure model, Cox proportional hazards model and logistic regression. It is also concluded that the proposed EMRF model provides a meaningful output for timely post-loan risk management.

Keywords: Behavioral scoring; dynamic probability, P2P lending; random forest; random survival forest; risk management; survival analysis.

---

1. INTRODUCTION

Peer-to-peer (P2P) *lending* is the practice of providing loans to individuals or businesses through online platforms that match lenders directly with borrowers, bypassing the need for a traditional financial intermediary, such as a bank. By operating fully online, P2P lending incurs lower overhead costs than do traditional bank loans, often resulting in higher returns for lenders and lower interest rates for borrowers. Thus P2P lending is especially attractive to individuals and small businesses (Guo et al., 2016). P2P lending has undergone phenomenal development in recent years and has now become an important alternative to loan services provided by traditional financial institutions in countries such as the U.S. and China (Bachmann et al., 2011). However, P2P lending also faces great challenges. The information asymmetry inherent to P2P lending—in which lenders know limited information about borrowers while borrowers know considerably more about their own risk levels—attracts riskier borrowers and misleads lenders to fund them, leading to higher default rates compared with bank loans (Chen and Han, 2012).

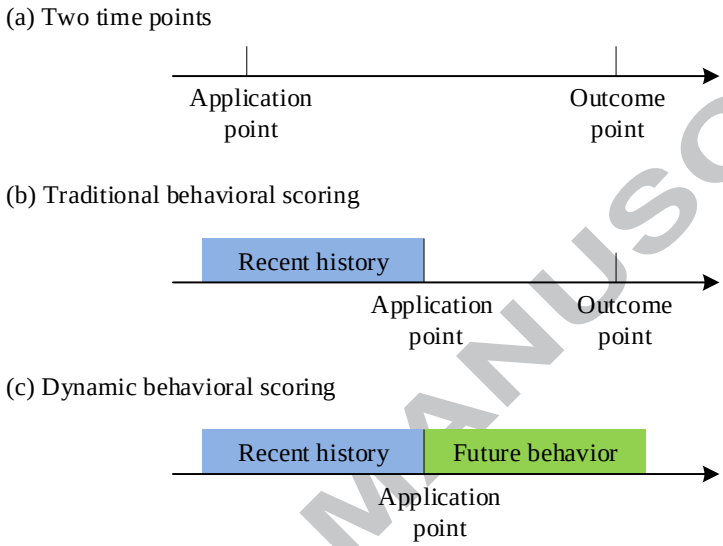
Financial institution business performance is heavily dependent on the management of credit risk. Credit scoring models are important tools for decision support systems in the financial industry, and have been widely applied to make decisions concerning whether to grant credit to new applications (Crook et al., 2007; Wang et al., 2012). Although credit scoring models can effectively filter out most risky borrowers, some borrowers still default even after their loan application has been approved. Therefore, most financial institutions subsequently apply behavioral scoring models to monitor borrowers' repayment behaviors, especially in P2P lending (Alves and Dais, 2015). An accurate behavioral scoring model for borrowers is extremely valuable in the P2P market.

Both credit scoring and behavioral scoring are a process of determining how likely borrowers are to default with their repayment, that is, predicting the *probability of default* (PD). *Credit scoring* is defined as a connection between two snapshots of the state of the borrower indicating the characteristics at the time of application and the state of the loan at a later date respectively (Savvopoulos, 2010). (See Panel (a) in Figure 1). The future performance (i.e., the performance of repayment behavior) is indicated as a static PD for the entire loan.

In a behavioral scoring model, the borrowers' future performance depends not only an initial snapshot of their risk condition at the time of application but also on their recent performance; thus, contrary to credit scoring, the PD takes a dynamic property (Alves and Dais, 2015). By adding information about the borrower's past performance such as the number of successful loans, failed loans and paid-off loans, the traditional behavioral scoring model replaces the first snapshot by a description of the dynamics of the borrower's performance (Panel (b) in Figure. 1), but the second snapshot remains. Given the dynamic property of repayment behavior, a more appropriate behavioral scoring model should be able to utilize the past behavior to estimate subsequent repayment performance over a future time interval, not just at a specific future time (Panel (c) in Figure 1). Thus, to develop dy-

namic behavior scoring one needs to estimate the dynamic PD over time, namely predicting not only *whether* but also *when* the borrowers are likely to default. Especially, this framework can offer guidance in a timely manner and improve the effectiveness of post-loan risk management and control. Experience has shown that early intervention can both effectively minimize the arrears and reduce the number of accounts that become bad debts and the resulting losses (Sarlija et al., 2009).

Figure 1. Behavioral Scoring Time Line



In the present paper, we propose a novel behavioral scoring model based on a mixture random forest ensemble to predict the PD over time. First, we propose a *mixture random forest* (MRF) *model* based on survival analysis to predict the PD over time. The proposed MRF model consists of two components: the *incidence* component, which is modeled by a random forest, predicts whether a borrower will default, and the *latency* component, which is modeled by a random survival forest, predicts the survival time (i.e., the time of the occurrence if it were to occur) of a borrower conditional on the borrower being susceptible to default. The two components are combined by a conditional probability formula to estimate the PD over time. In contrast to an ordinary behavioral scoring model, which poses a classification problem in which each loan is classified as default or non-default, the proposed mixture random forest model outputs a two-dimensional matrix of the probabilities of default over time (i.e., the PD at different times). Second, after obtaining the MRF model, we consider using a combination of models to reduce the variance and improve the stability.

An *ensemble* is a supervised learning paradigm in which multiple base learners are combined to form an improved learner. The trained ensemble represents a single hypothesis that is not necessarily contained within the hypothesis space of the models from which it is built. Thus, ensembles have greater flexibility to represent various functions. This flexibility can, in theory, enable ensembles to obtain relatively higher classification accuracies compared to a single model (Twala, 2010). In this paper, we utilize *tree-based ensemble methods* (i.e., a random forest and a random survival forest) to improve the performances of the incidence and latency components. The averaging ensemble, which

is a simple and effective model combination strategy, is used to improve the robustness of the proposed model.

We evaluated our proposed *ensemble mixture random forest* (EMRF) model on a real-world dataset collected from a major P2P platform in China. We compare our proposed model with a standard mixture cure model (Tong et al., 2012), the Cox proportional hazards model and logistic regression in terms of both their *discrimination* performances (the ability to risk-rank borrowers accurately over time) and their *calibration* performances (the accuracy of the probability of the default estimates themselves over time).

The remainder of this paper is organized as follows. In Section 2, relevant previous studies on behavioral scoring and P2P lending are presented. In Section 3, we propose an averaging ensemble model based on a mixture random forest for behavioral scoring. We describe the experimental design 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 in Section 6.

2. RELATED WORKS

2.1. Credit Assessment Models

In the scope of credit risk management, two types of scoring models have been widely used to make decisions on new loan applications and to monitor borrowers' repayment behaviors; the former model is known as a credit scoring model (also known as an application scoring model) while the latter model is known as a behavioral scoring model (Kao et al., 2012; Alves and Dias, 2015). Credit scoring and behavioral scoring models are commonly built on accepted loan applicants; however, credit scoring models are faced with the problem of sample bias because the sample of accepted applicants is different from those pertaining to practical applications. Thus, some researchers focus on reject inference to solve the problem of sample bias (e.g., Banasik, 2007; Chen and Åstebro, 2012). In addition to those two scoring model types, other types of implementations are also popular, such as the profit scoring model (e.g., Serrano-Cinca and Rrez-Nieto, 2016).

Despite their different purposes, behavioral scoring models usually treat borrowers similarly to credit scoring models, which pose a classification problem to estimate the PD based on classification models (see Baesens et al., 2003; Lessmann et al., 2015 for surveys of earlier research), such as logistic regression (Bensic et al., 2005), a neural network (Tsai and Wu, 2008;), or a support vector machine (Martens et al., 2007). In addition to individual classification models, multiple classifier systems in credit risk assessment, especially ensemble learning, have attracted extensive scholarly attention in recent years (Lessmann et al., 2015). Finlay (2011) and Lessmann et al. (2015) showed empirical evidence that multiple classifier architectures often predict credit risk with high accuracy. However, Lessmann et al. (2015) also found that sophisticated ensemble methods do not necessarily improve accuracy.

2.2. Survival Analysis

Survival analysis is a statistical method of data analysis in which the outcome of interest is the time to the occurrence of an event, often referred to as default in the context of credit assessment (Zhang and Thomas, 2012; De Leonardis and Rocci, 2014). Survival analysis introduces the time factor in the modeling of the event of concern (Yildirim, 2008). More generally, survival analysis involves the modeling of time to event data; in the context of credit scoring, the default is considered an "event" in the literature (Dirick et al., 2015; Liu et al., 2015). Survival analysis is well established in medical fields and was first introduced to credit risk assessment by Narain (1992). Subsequently, Banasik et al. (1999), Hand and Kelly (2001) and Stepanova and Thomas (2002) applied and compared various standard parametric and non-parametric survival models. Bellotti and Crook (2009) and Im et al. (2012) further extended the use of survival analysis in credit risk assessment by incorporating time-dependent variables to construct a proportional hazards model.

More recently, Tong et al. (2012) proposed the mixture cure model to take long-term survivals into account because a substantial proportion of borrowers do not experience default during the loan lifetime. They latently divide the borrowers into two subpopulations; in one subpopulation borrowers never default during the loan lifetime and in the other subpopulation borrowers default at some point during the loan. The two subpopulations are respectively modeled by incidence using logistic regression and latency using the semi-parameter Cox proportional hazards model (Cox, 1972). An empirical analysis showed that the mixture cure model obtains a better performance in terms of estimating the PD over different years (e.g., predicting default in the second year of a loan given that the borrower did not experience default in the first year of the loan).

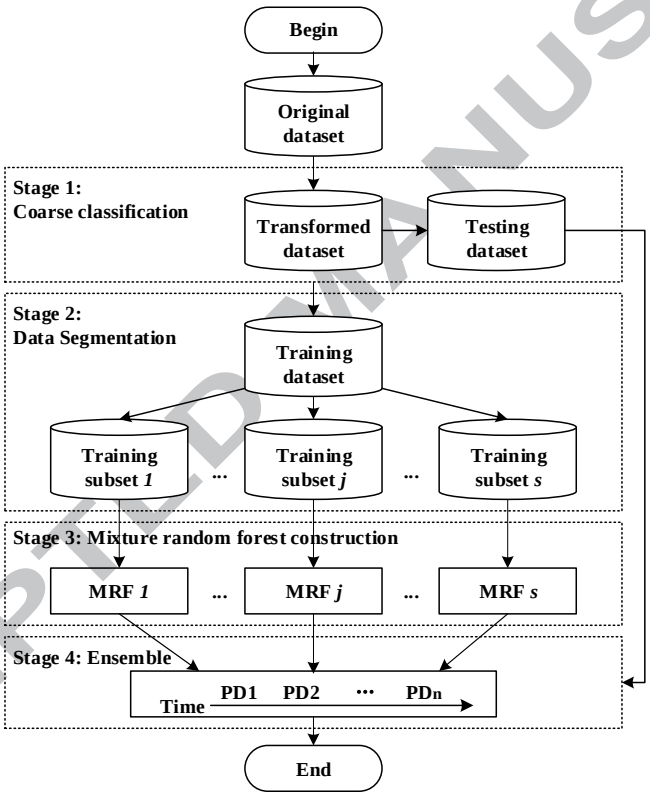
From these studies, it can be concluded that most previous research focuses on identifying whether borrowers will default based on various classification methods or predicting the time to default using survival analysis respectively. A mixture modeling of both whether and when borrowers are likely to default is more benefit for predicting the dynamic PD and can function as a credit risk decision support tool. In this study, we focus on modeling the two aspects of default together (i.e., whether and when) based on a mixture survival framework. Besides, we focus on a more fine-grained timeframe, namely, the PD as a function of time by month, because P2P lending primarily involves short-term loans, such as a 12-month loan. Two representative survival analysis models, the mixture cure model and the Cox proportional hazards model, are selected as benchmarks for comparison. Our study tries to illustrate the suitability of mixture survival models for providing guidance for timely and effective post-loan risk management in P2P lending.

3. PROPOSED METHODOLOGY

We next discuss how the *ensemble mixture random forest model* (EMRF) assessment process is implemented. The process of predicting the PD over time including variables preprocessing is illustrated in Figure 2. Specifically, we start by using the *weight of evidence* (WOE) t*ransformation* to coarsely classify the original variables. In the WOE transformation, all variables are converted into

orderly discrete variables, in which the order is based on the default rate of the categories in each variable. Then, the transformed dataset is divided into training and testing datasets using 10-fold cross validation, and each training set is further bootstrapped into several groups. To provide guidance for timely and effective post-loan risk management and control, we propose a mixture random forest model that outputs a two-dimensional matrix showing the probabilities of default over time. The proposed mixture random forest model is established on each training dataset. Because the model combination is intended to reduce the variance and improve stability, we chose the averaging ensemble strategy to integrate the output of the MRF used for each training dataset. In the remainder of this section, we discuss the EMRF in more detail.

Figure 2. General Process of Ensemble Mixture Random Forest Model



3.1. Weight of Evidence

The WOE transformation is based on information theory and was initially intended for scorecard development. Compared with dummy coding, another popular discretization method that adds multiple binary dummy variables for each nominal variable, WOE avoids dimensionality increases; thus, it has been shown to be superior to dummy coding (Moeyersoms and Martens, 2015).

As a recoding technique for raw data, WOE first needs to discretize the continuous variables. Then, it recodes the variable values into discrete categories and assigns a unique value to each category, a *WOE value*, with the goal of producing the largest differences between the recoded values. For category *i* in a discrete variable, the WOE value is calculated as follows:

$$\mathrm{WOE}_i = \left[\ln\left(\frac{Bad\ distribution_i}{Good\ distribution_i}\right)\right] \times 100, \quad (1)$$

where $Bad\ distribution_i$ and $Good\ distribution_i$ are respectively calculated as:

$$Bad\ distribution_i = \frac{Number\ of\ Bad_i}{Total\ Number\ of\ Bad}, \quad (2)$$

$$Good\ distribution_i = \frac{Number\ of\ Good_i}{Total\ Number\ of\ Good}, \quad (3)$$

where Bad denotes a default during the loan repayment lifetime, and $Good$ denotes non-default during the loan repayment lifetime.

3.2 Mixture Random Forest

3.2.1 Model Specification

The MRF model is based on the mixture cure survival analysis architecture, in which borrowers are divided into two subpopulations: cured and uncured. Borrowers in the cured subpopulation never default during the loan lifetime, while borrowers in the uncured subpopulation will eventually default at some point during the loan term. Because survival analysis is a dynamic data analysis method and it considers the status of each analysis object in different periods, let us define a censored indicator function δ_i for credit assessment where $\delta_i = 1$ denotes that borrower i defaults within the observation period and $\delta_i = 0$ otherwise. Let y_i denote the default indicator during entire loan period, where $y_i = 1$ denotes that borrower i will experience a default event during or after the observation period and $y_i = 0$ denotes that borrower i will never experience a default event. This condition, it will be censored at the end of any observation period. In the definition of δ and y, borrowers may be in one of three possible states (see Table 1).

Table 1. Possible States of a Borrower

δ	y	Description
1	1	Non-censored, and the borrower is observed to default
0	1	Censored, and the borrower will eventually default
0	0	Censored, and the borrower will never default

To start, we provide the expression for the mixture random forest (MRF). Let $PD(\boldsymbol{z})$ denote the PD given the vector of covariates $\boldsymbol{z}$. Additionally, we define $S(t|y = 1, \boldsymbol{x})$ as the survival PD at time t given a certain covariate vector $\mathbf{x}$ condition that the borrower will eventually default. Specifically, $S(t|y = 1, \boldsymbol{x}) = P(T > t|y = 1, \boldsymbol{x})$, where T denotes the elapsed time until borrower default from loan approval. The mixture random forest expression is

$$S(t) = \big(1 - PD(\boldsymbol{z})\big) + PD(\mathbf{z})S(t|y = 1, \boldsymbol{x}), \quad (4)$$

where $\mathrm{S}(t)$ denotes the probability of borrower survival after time period t. $PD(\boldsymbol{z})$ is referred to as the incidence component and $S(t|y = 1, \boldsymbol{x})$ is referred to as the latency component. These two components are respectively modeled by a random forest and a random survival forest. Thus, it is possible to distinguish the factors that affect the PD from factors that affect the distribution of default time.

Starting with the incidence component of MRF, the effect of the attributes vector $\mathbf{z}$ on the PD is modeled using the random forest (Breiman, 2001). A random forest, which builds multiple classification and regression trees on bootstrapped samples, can be considered as an enhanced bagging method.

In contrast to ensemble methods such as *bagging and boosting*, which can generally work with any type of base learner, random forest works with a particular type of learner: the classification and regression tree. In the context of P2P lending, it has been demonstrated that random forest outperforms several other state-of-the-art classification methods such as logistic regression and support vector machines (Malekipirbazari and Aksakalli, 2015). The pseudocode of the random forest algorithm is given in Figure 3.

Compared with growing a single decision tree, a random forest has some unique efficiencies. On one hand, while selecting the splitting attributes, all the attributes are tested at each node in the growing algorithm of a single decision tree, while the random forest itself needs to test only those selected as the m candidate splitting attributes. Because the number of candidate split attributes m, is typically smaller than the full set of M attributes ($m < M$), the search process of a random forest is very fast. On the other hand, pruning is usually needed when growing a single tree to avoid overfitting and achieve the optimal prediction strength with an appropriate complexity. The full random forest performs no pruning. Mitigating the detrimental effect of variance through model averaging, random forest with unpruned trees reduces both bias and variance (Breiman, 2001).

Figure 3. The Random Forest Algorithm

Input: Dataset $D = \{(\mathbf{z_1}, \mathrm{y}_1), (\mathbf{z}_2, y_2), \ldots, (\mathbf{z_n}, \mathrm{y}_n)\}$;
The number of classification and regression trees $ntree$;
The number of split attributes $mtry$.

Process: For $k = 1,2, \ldots, ntree$;
$D_k = Bootstrap(D)$; % Draw a bootstrap sample from dataset *D*
$\boldsymbol{m}_k = RS(M)$; % Randomly select $mtry$ attributes from the full set of *M* attributes
$\{B_{k\_1}, B_{k\_2}, \ldots, B_{k\_m}\}$; % Choose the best split attribute from candidate attributes
$H_k = L(D_k, \boldsymbol{m}_k)$; % Let the tree grow to the maximum size without pruning
end.

Output: $PD_i = \frac{1}{ntree}\sum_{j=1}^{ntree} H_k\,(y_i = 1)$. % Aggregate the results of $ntree$ trees.

The latency component of MRF, which models the effects of the attributes vector $\boldsymbol{x}$ on the survival distribution of the uncured subpopulation, is implemented by the random survival forest model (Ishwaran et al., 2011). A random survival forest is an ensemble survival tree method used to analyze censored survival data. The process of building a survival forest is similar to that of building random forest in that it starts by drawing multiple bootstrap samples. However, in contrast to a random forest, which working with the *classification and regression tree* (CART) paradigm, a random survival forest consists of multiple survival trees based on bootstrap samples. Then, each grown survival tree calculates a *cumulative hazard function* (CHF) and average to obtain the ensemble CHF.

As central elements of random survival forest, growing survival trees and calculating the CHFs are discussed in detail. Similar to classification and regression trees (CART), the process of growing survival trees is implemented by recursively splitting the tree nodes. First, the root node is split into a left and right daughter node according to a predetermined survival criterion whose purpose is to maximize the survival difference between daughters. Then, each daughter node is subsequently split

into child nodes. This process is repeated recursively for each daughter node until the survival tree finally reaches a saturated state, at which point no new daughter nodes can be formed under the constraint condition that each node must contain at least one unique uncured observation. We define the most extreme nodes in a saturated tree as terminal nodes, TN.

Let $\{(T_{1\_h}, \delta_{1\_h}), (T_{2\_h}, \delta_{2\_h}), \dots, (T_{n\_h}, \delta_{n\_h})\}$ be the survival times and censored indicator values for n borrowers at a terminal node h ($h \in TN$). Specifically, let $t_{1\_h} < t_{2\_h} < \cdots < t_{N(h)\_h}$ be the $N(h)$ distinct default times. The cumulative hazard function (CHF) for terminal node h is estimated as follows:

$$\widehat{H}_h(t) = \sum_{t_{l\_h} \le t} \frac{d_{l\_h}}{Y_{l\_h}}, \quad (5)$$

where d_{l_h} denotes the number of borrowers in default at time t_{l_h}, and Y_{l_h} denotes the number of borrowers at risk at time t_{l_h}.

To determine the cumulative hazard function (CHF) for each borrower i, we drop the attribute vector $\boldsymbol{x}_i$ of each borrower i down the survival tree. Because of the binary splitting mechanism, the attribute vector $\boldsymbol{x}_i$ will wind up at a unique terminal node, h. The cumulative hazard function (CHF) for each borrower i is given by

$$H(t|\boldsymbol{x}_i) = \widehat{H}_h(t), \text{ if } \boldsymbol{x}_i \in h. \quad (6)$$

3.2.2 Model Estimation

Let $(t_i, \delta_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$ denote the observed information for the borrower i, $i = 1,2, \dots, n$, where t_i denotes the observed survival time, δ_i denotes the censored status, and $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ respectively denote the attribute vector for the random forest (incidence component) and the random survival forest (latency component). The complete likelihood function of Equation (4) can be expressed as follows:

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

where $h(t_i|y = 1, \boldsymbol{x}_i)$ is the hazard function corresponding to $S(t_i|y = 1, \boldsymbol{x}_i)$.

The log complete likelihood function can be divided into two parts: the log-likelihood function of the incidence component L_I, and the log likelihood function of the latency component L_L:

$$L_I = \sum_{i=1}^{n} (y_i \log [PD(\boldsymbol{z}_i)] + (1 - y_i) \log [1 - PD(\boldsymbol{z}_i)]), \quad (8)$$

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

According to the log-likelihood functions of the incidence and latency components, the expectation of y_i can be written as:

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

Because y_i is an unobserved variable, we use the *expectation-maximization* (EM) *method* (Sy and Taylor, 2000) to estimate the expectation of y_i. The pseudocode for the *expectation maximization* (EM) *algorithm* is given in Figure 4.

Figure 4. The Expectation Maximization Algorithm

Input: Observation data $(t_i, \delta_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$;
The number of iterations m with an initial value of 1 and a maximum value of M;
Predetermined criterion: θ.

Process: While $\{\Delta^2 < \theta$ or $m \leq M$;
$\mathrm{E}(L_I) = \sum_{i=1}^{n}(y_i{}^{(m)} \log [PD(\boldsymbol{z}_i)] + (1 - y_i{}^{(m)}) \log [1 - PD(\boldsymbol{z}_i)])$; % M-step
$\mathrm{E}(L_L) = \sum_{i=1}^{n}(y_i{}^{(m)} \delta_i \log[h(t_i|y = 1, \boldsymbol{x}_i)] + y_i{}^{(m)} \log [S(t_i|y = 1, \boldsymbol{x}_i)])$; %M-step
$E(y_i|t_i, \delta_i, \boldsymbol{z}_i, \boldsymbol{x}_i)^{(m)} = \delta_i + (1 - \delta_i) \frac{PD(\boldsymbol{z}_i) S(t_i|y=1, \boldsymbol{x}_i)}{1 - PD(\boldsymbol{z}_i) + PD(\boldsymbol{z}_i) S(t_i|y=1, \boldsymbol{x}_i)}$. % E-step
end.

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

3.3 Averaging Ensemble

After obtaining the single MRF, we improve the stability and the performance of the MRF by mode combination. An averaging ensemble is one of the most intuitive and simplest types of mode combinations and has been shown to have surprisingly good performance (Finlay, 2011). Let O denote the dataset processed by WOE transformation. The process of creating multiple subsets $(O_1, O_2, \ldots, O_s)$ is implemented by sampling with replacement from O. Specifically, each subset contains the same number of observations as O after sampling. Then, each O_j of the multiple subsets $(O_1, O_2, \ldots, O_s)$ is expected to have a fraction $\left(1 - \frac{1}{e} \approx 63.2\%\right)$ of the unique observations of O and the rest are duplicated. The final output of the ensemble mixture random forest (EMRF) is a two-dimensional PD matrix in which each row represents a borrower and each column represents an observation time point. The output of the EMRF, supplied by the $M_{ensemble}$, is derived as follows:

$$M_{ensemble} = \frac{1}{s} \sum_{j=1}^{s} M_{MRF_j}, \quad (11)$$

where M_{MRF_j} denotes the output PD matrix of mixture random forest built on the jth observation subset O_j.

4. EMPIRICAL EVALUATION

The empirical evaluation was performed on a PC with a 3.20 GHz Intel Core i5-3470 CPU and 12 GB of RAM using the Windows 7 operating system. The data mining toolkit from R version 3.31 was used for the experiment. R is a free software environment for statistical computing and graphics.

4.1. Real World Dataset of P2P Lending

We evaluated our proposed ensemble mixture random forest (EMRF) model by using a large dataset from a major peer-to-peer platform in China. The data were collected between January 2013 and December 2015. The dataset used in evaluation consists of 52,573 approved applications that were

subsequently funded for 12-month loans. The definition of a default is when repayment is at least three months overdue; otherwise, loans are non-default. As a result, the dataset contains 6,079 defaults and 46,494 non-default loans in our dataset; the default rate is approximately 11.56% (i.e., 6,079/52,573). The attributes used in the analysis, including the loan information, borrower information and historical information (A4~A7), are listed in Table 2.

Table 2. Description of Attributes

No.	Attributes	Summary Statistics			
	Continuous	*Min*	*Max*	*Mean*	*SD*
1	Age	20	64	38.52	8.60
2	Loan Amount	500	2,000,000	59,719.87	23,485.88
3	Interest Rate (%)	10	25	18.78	1.25
4	Number of Successful Loans	0	9	1.58	0.71
5	Number of Failed Loans	0	10	0.07	0.33
6	Number of Loans Paid Off	0	9	0.80	0.69
7	Number of Remaining Repayment Periods (month)	0	24	4.37	4.41
	Categorical	*Number of Values*	*Categories*		
8	Gender	2	{male, female}		
9	Loan Type	4	{personal consumption, borrowing, capital turnover, other}		
10	Annual Income	6	{<20 k, 20 ~60 k, 60~120 k, 120~240 k, 240~400 k, > 400 k}		
11	Platform Assigned Grade	8	{AAA, AA, A, BB, B, CC, C, HR}		
12	Credit Line	7	{None, <3 k, 3~6 k, 6~20 k, 20~50 k, 50~100 k, > 100 k}		
13	Living Area	5	{rural residential, urban community, city community, commercial community, other}		
14	Home Ownership	6	{renting/no house, parent's house, mortgage <400 k, mortgage ≥ 400 k, own < 1,000 k, own ≥ 1,000 k}		
15	Credit Report	2	{yes, no}		
16	House Guarantee	2	{yes, no}		
17	Car Guarantee	2	{yes, no}		
18	Guarantor	2	{yes, no}		
19	Insurance Status	2	{yes, no}		
20	Yrs w/ Soc. Sec.	5	{unpaid, <1 y, 1~3 y, 3~5 y, > 5 y}		
21	Yrs w/ Credit Records	4	{No record, 1~3 y, 3~5 y, > 5 y}		
22	Years with Employer	5	{unemployed, <1 y, 1~3 y, 3~5 y, > 5 y}		
23	Marital Status	4	{unmarried, married childless, married with children, divorced}		
24	Education Level	5	{junior high school or below, senior high school/technical secondary school, junior college, bachelor, master or above}		
25	Occupation Type	6	{student/job-waiting/unemployed, company staff, civil servant, public institution, financial institution, other}		
26	Job Title	7	{student/uncertain title, staff labor, junior executive, middle executive, senior executive, business, other}		

[*] The unit of money is RMB (¥), 1 k = ¥1,000.

[**] Note that some minority categories were merged into one category (e.g., category "other" for Loan Type and "AA", "AA+" and "AA" were merged into "AA", for Platform Assigned Grade, etc.)

4.2. Evaluation of Proposed Method

We evaluated our proposed EMRF model in comparison with two dynamic models, the *mixture cure model* (MCM) (Tong et al., 2012) and the *Cox proportional hazards* (Cox PH) model, both of which predict the PD over time, as well as one typically static model, *logistic regression* (LR), which

predicts only the PD over the entire repayment period. The MCM is a state-of-art model for survival analysis that utilizes logistic regression and Cox proportional hazard regression to model the incidence and latency components, respectively. The Cox PH model is a classical method for modeling the time of an event. In addition, LR, which estimates the probability of a binary response using a logistic function, is one of the most popular credit risk evaluation models, and it is widely used as a benchmark model in credit risk modeling.

We compared the performances of the proposed EMRF MCM, Cox PH and LR models in terms of their ability to predict the PD over time for a 12-month loan. The time interval is one month because borrowers' repayments are due on a monthly basis, leading to multiple discrete time intervals for a 12-month loan. Note that with by our definition of default (repayment overdue for at least three months) there are 10 discrete time intervals, denoted TI_1 to TI_{10} (where TI_1 corresponds to the third month). In the EMRF, the size of the ensemble was 100 and the size of the forest was 500 trees. Variable selection for EMRF was performed according to the GINI-index in the random forest and the C-index in the random survival forest. Variable selection for MCM, Cox PH and LR were performed using backward stepwise elimination at a 5% significance level.

We examined both the discrimination performance and the calibration performance among the four models (i.e., EMRF, MCM, Cox PH and LR). The discrimination performance is a measure of the model's ability to distinguish bad borrowers from good borrowers. In our dynamic assessment scenario, we compared the discrimination performance of the four models (i.e., EMRF, MCM, Cox PH and LR) by examining the performance for risk-ranked borrowers at various time intervals in terms of the following three measures: (1) the *area under the receiver operating characteristic curve* (AUC), which is calculated as the area under the receiver operating characteristic curve and ranges in value from 0 to 1; (2) the *Kolmogorov–Smirnov* (KS) *statistic*, which is the maximum difference between the cumulative true positive and cumulative false positive rate; and (3) the *H-measure*, proposed by Hand (2009), who argued that the H-measure is a coherent estimator that is insensitive to the empirical score distributions of the default and non-default groups. The H-measure should be considered the measure of choice to compare the performance of each method.

To better reflect the dynamic repayment behavior, we also examined the calibration performance among the four models. The calibration performance refers to the accuracy of the PD estimates themselves over time (Liu et al., 2015). We evaluated the calibration performance by comparing the expected cumulative number of defaulters against the observed number of defaulters from TI_1 to TI_{10}. Specifically, the expected cumulative number of defaulters in each time interval was calculated by the sum of the default probabilities of a borrower defaulting before the end of the corresponding time interval:

$$E(D_{TI}) = \sum_{i=1}^{n}\big(1 - S(TI|i)\big). \tag{12}$$

During the process of estimating the performance of each model (i.e., EMRF, MCM, Cox PH and

LR), we performed repeated cross validation (REPCV), specifically, 10 times for the independent 10-fold cross validations, to achieve unbiased estimates. Molinaro et al. (2005) argued that 10 random fold splits are sufficient to realize most of the achievable variance reduction. During each 10-fold cross validation, the total samples are randomly divided into 10 equally sized folds, 9 of which are selected to train each model and the remaining fold is used for validation. Note that the training sets and testing sets were kept consistent across the four models.

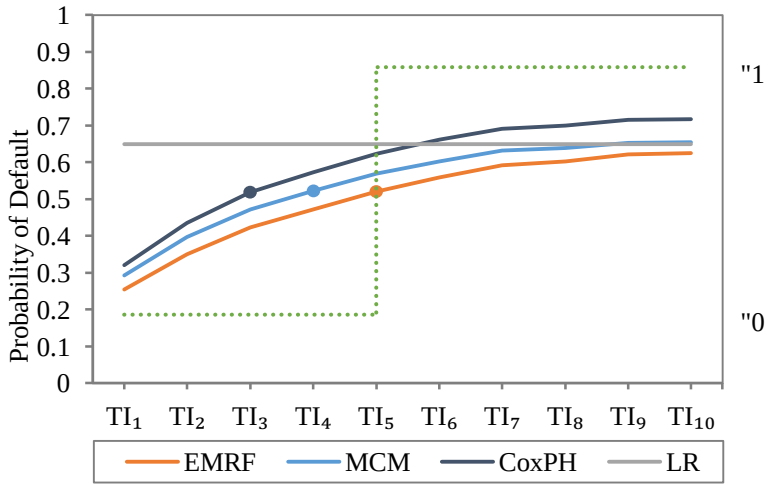
5. RESULTS

5.1. Dynamic Repayment Process

Before examining the discrimination and calibration performance of each model, we first provide an example to illustrate how the dynamic PD reflects the dynamic repayment process. Figure 5 displays the predicted PD over time compared to the actual repayment process. The solid lines depict the predicted PD of four models (i.e., EMRF, MCM, Cox PH and LR) over time. The dotted lines ("0" and "1" denote "non-default" and "default", respectively) depict the actual repayment process, indicating that the borrower was on a regular repayment schedule until the fifth time period (i.e., the 7th month of the entire 12-month loan), at which point he defaulted.

The dynamic PD reflects the dynamic repayment process by predicting not only whether the borrowers will default but also when they are likely to default. The marked points in Figure 5 denote the time of default using the default threshold of 0.5 (i.e., a loan is treated as "default" when the predicted PD is greater than 0.5). The outputs of EMRF, MCM and Cox PH indicate that the borrower defaulted at the 5th, 4th and 3rd time periods, respectively, and the prediction result of the proposed EMRF is consistent with the actual repayment process, where the borrower defaulted at the fifth time period. As mentioned above, logistic regression can generate only a static PD over the entire loan period; thus, the dynamic PD manifests as a straight line. Although it provides an accurate estimation indicating that the borrower will default, it cannot forecast the dynamic repayment process.

Figure 5. Example of Predicted PD over Time against Actual Repayment Process on a Sample



5.2. Discrimination Performance

Table 3 summarizes the AUC results on repeated 10-fold cross validations for the four models (in terms of mean values and 95% confidence intervals). In the time intervals from TI_3 to TI_9, the proposed EMRF model performed significantly better than all the other models in terms of the mean value of AUC, which indicates its robust performance. The performance of MCM was superior to that of Cox PH for all the time intervals except TI_3 and TI_7. Among the three survival methods (i.e., EMRF, MCM and Cox PH) over the different time intervals (TI_1 to TI_{10}), the Cox PH model performed worst at nearly every time interval, which may be caused by the assumption of the standard survival method (e.g., Cox PH) that all borrowers will eventually default given sufficient observation time. The mixture modes (i.e., EMRF and MCM), which both consider the non-defaulting subgroup and model the default and non-default subgroups separately, outperformed Cox PH for most of the time intervals. Because the LR model could generate results of one default probability for the whole loan rather than yielding the PD over time, we simply replicated the static result of LR 10 times to correspond to the times from TI_1 to TI_{10}. The LR performance fluctuates; it performed best at TI_2 but worst at TI_9. These unstable LR results indicate that static binary output cannot effectively reflect the dynamic change of PD in the scope of dynamic evaluation.

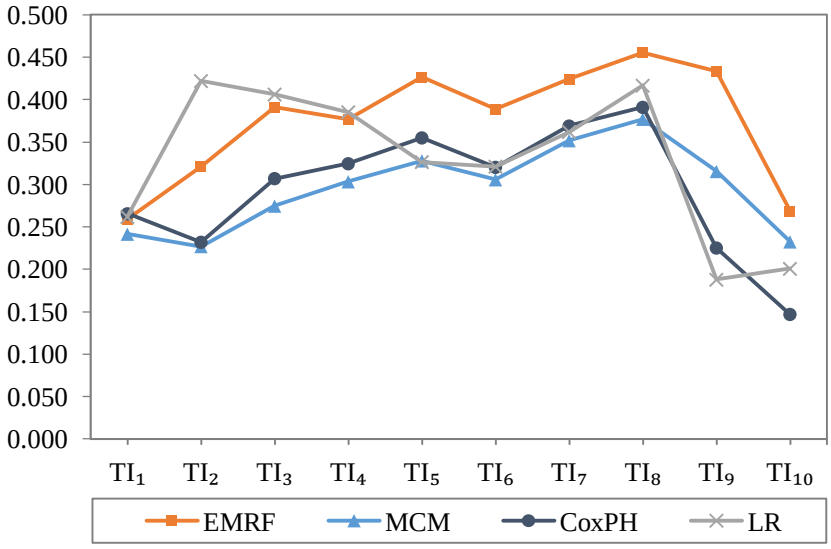
Table 3. AUCs of Repeated 10-fold Cross Validations for Four Models, Different Time Intervals

Time Interval	Model (95% CI)			
	EMRF	**MCM**	**Cox PH**	**LR**
TI_1	.642(.632-.652)	**.683**(.676-.690)	.644(.636-.651)	.638(.632-.644)
TI_2	.682(.674-.691)	.670(.662-.677)	.595(.585-.605)	**.725**(.719-.731)
TI_3	**.731**(.722-.739)	627(.618-.636)	.669(.659-.679)	.710(.703-.717)
TI_4	**.712**(.702-.722)	.694(.684-.703)	.656(.645-.666)	.696(.687-.705)
TI_5	**.747**(.738-.755)	.681(.671-.691)	.675(.665-.685)	.677(.668-.685)
TI_6	**.717**(.705-.729)	.695(.687-.703)	.658(.645-.671)	.659(.649-.670)
TI_7	**.740**(.731-.750)	.675(.662-.688)	.679(.669-.690)	.695(.688-.702)
TI_8	**.715**(.696-.733)	.707(.697-.717)	.635(.616-.655)	.679(.658-.700)
TI_9	**.751**(.742-.759)	.676(.656-.697)	.562(.552-.573)	.551(.546-.557)
TI_{10}	.629(.620-.638)	**.641**(.632-.650)	.542(.535-.549)	.560(.552-.569)

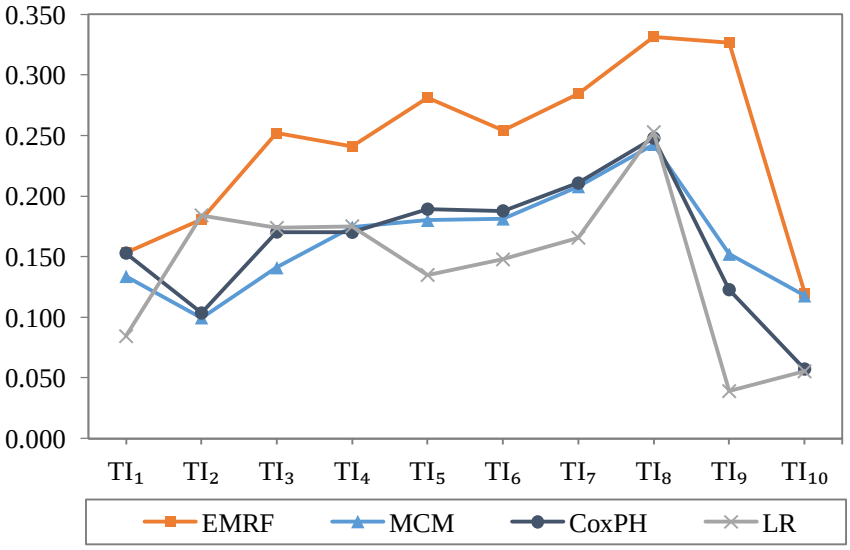
Figure 6 contrasts the four models in terms of KS and H-measure from TI_1 to TI_{10}. The shapes of the curves related to the three survival methods (i.e., EMRF, MCM, Cox PH) are similar for both KS and H-measure, which demonstrates the robustness of the results. Meanwhile, the curves related to EMRF are above those for MCM and Cox PH at nearly all the time intervals, indicating that it performed better. The patterns of LR with respect to the KS and H-measure vary somewhat. Compared with the other three models, the performance of LR in terms of H-measure was relatively inferior to its performance in terms of KS. This result was probably caused by differences in the misclassified cost distributions between the ROC-based metrics (i.e., AUC and KS) and H-measure; the AUC and KS utilize different misclassified cost distributions for different models, but the H-measure pre-sets a beta distribution for misclassified cost functions. In this respect, the H-measure is more reasonable

(Hand, 2009). The AUC, KS and H-measure show that the proposed EMRF performs well for risk-ranking borrowers when predicting dynamic default risk.

Figure 6. Comparison of Four Models in Terms of Mean KS and H-Measure



(a) KS



(b) H-measure

To verify the significance of the difference in model performances, the results of paired-sample t-tests for the models' H-measure results are shown in Table 4. Paired-sample t-tests were performed on 100 performance estimates by each model in each time interval (the output of the 10 independent repeated 10-fold cross validations). The mean obtained H-measure values between EMRF and other three models were statistically significant ($p < 0.01$) in most of the time intervals (70%—7 out of 10 time intervals). In terms of H-measure, MCM and Cox PH yielded statistically similar performances

and the performance of LR was significantly inferior to all the other models in half (50%) of the time intervals (i.e., $TI_1, TI_5, TI_6, TI_7, TI_9$).

Table 4. Full-pairwise Comparison of the Four Models

	M1 vs M2	M1 vs M3	M1 vs M4	M2 vs M3	M2 vs M4	M3 vs M4
TI_1	.000	.955	.000	.000	.000	.000
TI_2	.227	.000	.594	.000	.045	.000
TI_3	.000	.000	.000	.000	.000	.648
TI_4	.000	.000	.000	.001	.002	.563
TI_5	.000	.000	.000	.439	.000	.000
TI_6	.000	.000	.000	.087	.000	.000
TI_7	.000	.000	.000	.932	.000	.000
TI_8	.000	.000	.000	.622	.428	.789
TI_9	.005	.000	.000	.000	.000	.000
TI_{10}	.012	.000	.000	.000	.000	.535

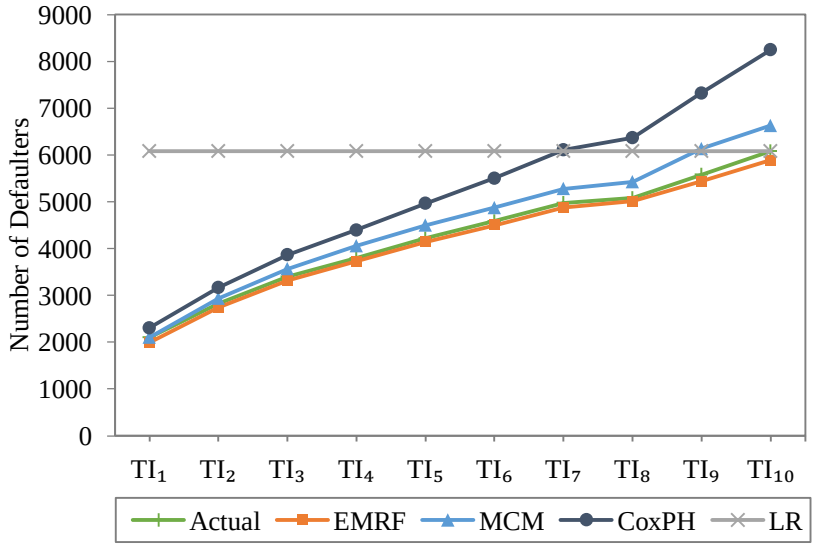
* M1 = EMRF, M2 = MCM, M3 = Cox PH, M4 = LR

5.3. Calibration Performance

To compare the calibration performance among the four models, we performed 10-fold cross validation so that each sample would serve as a testing sample during the 10-fold testing, which generated the PDs for all observations in the dataset and computed the expected cumulative number of defaulters.

Figure. 7 depicts the expected cumulative number of defaulters against the actual number of defaulters at each validation time from TI_1 to TI_{10}.

Figure 7. Actual Versus Expected Number of Defaulters from TI_1 to TI_{10}.



Although LR provides a very precise estimation of the number of defaulters by the end of the loan term, it cannot forecast the number of defaulters at other times intervals because it predicts only a static PD for each loan. However, as shown in Fig. 7, the inclusion of dynamic behavioral scoring models in the EMRF, MCM and Cox PH made such predictions possible. Furthermore, the curve related to EMRF is closer to the true curve compared with the curves related to MCM and Cox PH, in-

dicating a better calibration performance. It also shows that, in the peer-to-peer lending scenario, models such as EMRF and MCM, which model borrowers in terms of two distinct subpopulations (i.e., default and non-default), achieve a performance superior to that of models such as Cox PH, which assumes that all borrowers will eventually experience a default event if given a sufficiently long observation period.

6. CONCLUSION

Behavioral scoring has gradually become an essential tool for financial institutions in today's credit marketplace, which requires highly automated decisions, especially in the P2P market. To monitor the repayment behavior and evaluate the creditworthiness of existing customers, traditional behavioral scoring models have used a variety of classification methods to predict a static PD over a future period. Because borrowers' repayment behavior is a dynamic process, in this paper, we proposed a dynamic ensemble mixture random forest (EMRF) model based on survival analysis to predict the PD over any time horizon. We introduced the random forest and random survival forest models to model the incidence component $\left(PD(\boldsymbol{z})\right)$ and the latency component $\left(S(t|y=1,\boldsymbol{x})\right)$, respectively, with high accuracy and fine time granularity (monthly). An evaluation analysis of data from a major P2P institution in China showed that the performance of the EMRF was significant in terms of both discrimination and calibration, which indicates it can be used as an alternative and superior method for behavioral scoring.

Our research makes several contributions to both research and practice. First, the proposed ensemble mixture random forest model can generate the PD over time for predicting not only whether borrowers will default on their loan repayments but also when they are likely to default. It can be used by risk management practitioners and researcher to improve credit risk evaluation. It also creates opportunities for new research and applications in profit scoring domain since a prediction of dynamic PD can provide the ability to compute the profitability over a borrower's lifetime and perform profit scoring. Second, we modeled the incidence and latency components using tree-based methods as well as ensemble learning method, which can deal with potential non-linear relationship between independent variables and the target variable more naturally and help to improve the performance. Our experiments show that the proposed model is indeed powerful in predicting the PD over time. Third, financial institutions may also be interested in the proposed model because they can strengthen their post-loan management capabilities by re-ranking the borrowers' creditworthiness and predicting the dynamic number of defaulters according to the PD at any feasible time—a capability that cannot be achieved by traditional scoring models. Timely intervention for potential defaulters can effectively reduce loan delinquencies as well as the number of accounts that become bad debts. Those reductions could thus decrease losses at financial institutions.

There are a few limitations with this work. One limitation of EMRF is the weak interpretability of the predicted results (i.e., the black-box problem). It is difficult to transform the discrimination pro-

cess of ensemble methods such as bagging, boosting and random forest into rules that have logical relationships. Future research may consider improving the interpretability of ensembles in the framework of survival analysis. Another limitation lies in the limited available data: we use only the application and history repayment information. Further research may add the variables reflecting the management interfere from P2P platform to further improve the dynamic assessment. In addition, we implemented experiments on a dataset from only one P2P institution in China; future research may collect multiple datasets from different countries and sites, such as LendingClub and Proposer, to comprehensively validate our findings.

Acknowledgments. This work was funded by the National Natural Science Foundation of China (Grant Nos. 71731005, 71571059) and the Humanities and Social Sciences Fund Projects of the Ministry of Education (Grant No. 15YJA630010).

REFERENCES

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

Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., Tibertus, P., Funk, B. (2011). Online peer-to-peer lending: A literature review. Journal of Internet Banking and Commerce, 16(2).

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.

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.

Banasik, J. (2007). Reject inference, augmentation, and sample selection. European Journal of Operational Research, 183(3), 1582-1594.

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

Bensic, M., Sarlija, N., Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. Intelligent Systems in Accounting Finance and Management, 13(3), 133-150.

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

Chen, G. G., Åstebro, T. (2012). Bound and collapse bayesian reject inference for credit scoring. Journal of the Operational Research Society,63(10), 1374-1387.

Chen, D., Han, C. (2012). A comparative study of online p2p lending in the usa and china. Journal of Internet Banking and Commerce, 17(2), 1-15.

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.

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.

Dirick, L., Claeskens, G., Baesens, B. (2015). Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society, 1-14.

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

Guo, Y., Zhou, W., Luo, C., Liu, C., Xiong, H. (2016). Instance-based credit risk assessment for in-

vestment decisions in P2P lending. European Journal of Operational Research, 249(2), 417-426.

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.

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.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2011). Random survival forests. Journal of Thoracic Oncology Official Publication of the International Association for the Study of Lung Cancer,6(12), 1974.

Kao, L. J., Chiu, C. C., Chiu, F. Y. (2012). A bayesian latent variable model with classification and regression tree approach for behavior and credit scoring. Knowledge-Based Systems, 36(6), 245-252.

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), 124–136.

Li, W., Wu, X., Sun, Y., Zhang, Q. (2011). Credit card customer segmentation and target marketing based on data mining. In Proceedings of the International Conference on Computational Intelligence and Security, IEEE Computer Society Press, Washington, DC.

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

Liu, Y., Ram, S., Lusch, R.F., Brusco, M. (2010). Multicriterion market segmentation: A new model, implementation, and evaluation. Marketing Science, 29(5), 880-894.

Martens, D., Baesens, B., Gestel, T.V., Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. European Journal of Operational Research, 183(3), 1466-1476.

Molinaro, A. M., Simon, R., Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. Bioinformatics, 21(15), 3301-7.

Narain, B. (1992). Survival analysis and the credit granting decision. Credit Scoring and Credit Control. Oxford, Oxford University Press, 109-121.

Nosal, J., & Galenianos, M. (2015). Segmentation of information and the credit market. Medicine and Science in Sports and Exercise, 12(3), 189-194.

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

Sarlija, N., Bensic, M., Zekic-Susac, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. Expert Systems with Applications, 36(5), 8778-8788.

Savvopoulos, A. (2010). Consumer credit models: Pricing, profit and portfolios. Journal of the Royal Statistical Society, 173(2), 468-468.

Serrano-Cinca, C., Gutierrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer lending. Decision Support Systems, 89(C), 113-122.

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), 227-36.

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.

Tsai, C. F., Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications, 34(4), 2639-2649.

Twala, B. (2010). Multiple classifier application to credit risk assessment. Expert Systems with Ap-

plications, 37(4), 3326-3336.

Wang, G., Ma, J., Huang, L., Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. Knowledge-Based Systems, 26, 61-68.

Yildirim, Y. (2008). Estimating default probabilities of CMBS loans with clustering and heavy censoring. Journal of Real Estate Finance and Economics, 37(2), 93-111.

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.

- We propose a novel behavioral scoring model for online peer-to-peer lending.
- The proposed model predicts if and when borrowers will default.
- For incidence component, random forest is used to predict whether default.
- For latency component, random survival forest is used to predict when to default.
- The empirical study is implemented on a real-world dataset form P2P platform.