European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Interfaces with Other Disciplines

The profitability of online loans: A competing risks analysis on default and prepayment

Zhiyong Li[a,e], Aimin Li[b,a], Anthony Bellotti[c], Xiao Yao[d,*]

[a] School of Finance and Fintech Innovation Center, Southwestern University of Finance and Economics, 555 Liutai Avenue, Wenjiang, Chengdu 611130, China
[b] School of Economics and Management, Chongqing University of Posts and Telecommunications, No.2 Chongwen Road, Nan'an District, Chongqing 400065, China
[c] School of Computer Science, Faculty of Science and Engineering, University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo 315100, China
[d] Faculty of Business School, Central University of Finance and Economics. 39 South College Road, Beijing 100081, China
[e] Collaborative Innovation Center of Financial Security, Southwestern University of Finance and Economics, China

ARTICLE INFO

Article history:
Received 30 April 2021
Accepted 10 August 2022
Available online xxx

JEL code:
C34
D8
G17

Keywords:
OR in banking
Competing risks
Credit scoring
Profitability
Survival analysis

ABSTRACT

Traditional credit scoring models help lenders to make informed decisions in identifying those borrowers most likely to default. We analyse over one million online loans and find that the rates for both default and prepayment are relatively high compared to traditional bank loans. A preliminary nonparametric life-table estimate shows that loans with different terms exhibit varying patterns of hazards. We use a proportional hazard model with competing risks to predict the time to default and prepayment, and parameterise those covariates affecting the time to both events. Two dimensions of predictive performance, the discriminant power and the probability calibration, are then examined. To further support the primacy of profit-driven decisions, we propose a framework based on competing risks survival analysis to estimate the profitability of loans and the return of loan portfolios. Profitability forecasts incorporating both the time to default and prepayment are compared to the profitability of real assets, and finally a penalty is suggested to compensate for those losses incurred by prepayment.

1. Introduction

Credit scoring models have been widely adopted by financial institutions such as banks to support lending decisions. Standard credit scoring takes loan application decisions as a binary classification problem, expressed as a decision to either accept the 'good' applicants or reject the 'bad' ones, who have high probabilities of default beyond a certain threshold. Online lending is developing rapidly as an alternative credit market, with regard to both the number of lending platforms and the volume of loans in recent years (Jiang et al., 2018). Due to information asymmetry (Jiang et al., 2018; Lin et al., 2013; Serrano-Cinca & Gutierrez-Nieto, 2016; Serrano-Cinca et al., 2015), it is important to classify borrowers into various risk levels so that a rating grade can be applied to a tranche of loan assets in a portfolio. However, what is more important for investors, particularly for institutional investors such as banks and insurers who take up a large proportion stake on the lending platforms, is to estimate the risk-adjusted returns while also taking all potential risks in the portfolios into consideration.

There are two events which may have great impact on the risk-adjusted return of online loans: default and prepayment. Defaults bring severe losses of both the interest and the principal to the lenders, so that forecasting the probability of default has been the main focus of the associated literature (Dorfleitner et al., 2016; Hu et al., 2019; Michels, 2012). However, prepayment only causes partial losses of interests in the remaining payments, which has attracted less attention when compared to default. We find that most studies addressing the issue of prepayment are concentrated on mortgages, such as Beltratti et al. (2017), Ergungor & Moulton (2014); Mayer et al. (2013), Rose (2013), Steinbuks (2015), Varli & Yildirim (2015). However, except for Li et al. (2019) and Li et al. (2021), little research has focussed on online prepayment, despite the fact that the proportion of prepayments for online lending is much higher compared with the mortgages, because access to credit and closure of accounts are much easier and less costly online. As revealed in this paper, over a half of issued loans have been settled off earlier than their maturity, which leads to losses of potential interests to investors. Guo et al. (2016) optimise investment

* Corresponding author.
E-mail address: yaoxiao18@cufe.edu.cn (X. Yao).

https://doi.org/10.1016/j.ejor.2022.08.013
0377-2217/

decision making for online loan portfolios but fail to recognise the potential losses from prepayment. Considering relative high interest rates on online risky assets, the lost interests of online loans incurred from prepayment cannot be neglected. Compared to the existing literature on credit risk and the profitability of online loans, the major contribution of this paper is to propose a new approach to forecasting profitability, which takes both default risk and prepayment risk into consideration and gives new insights into portfolio management.

Essentially the losses caused by default and prepayment depend on the timing of the events, whichever happens first having the most significant impact. An appropriate method to capture this feature is called 'competing risks analysis'. Precisely, the 'time to event' is set to be the target variable, so when modelling time to default (or prepayment), prepayment (or default) is regarded as a competing event that would result in censoring, and vice versa. This allows us not only to calculate the time value of returns, but also to make use of a larger volume of censored data in analysis, which is difficult to handle with conventional statistical or machine learning methods (Allison, 2010). The competing risks model enables us to estimate the profitability of online loans, which is an important issue in banking and finance (Fitzpatrick & Mues, 2021).

Cox (1972) proposes the traditional proportional hazards (PH) model which thereafter has been utilised extensively in fields as diverse as reliability analysis for engineering, lifetime analysis for medicine as well as time-to-event analysis for social sciences. Kumar and Klefsjö (1994) set out the characteristics of the PH model and review its methodological issues, as well as the range of its possible applications, from life sciences to mechanical engineering. The PH model has also applied to economic and financial studies, e.g. Cole & Gunther (1995); Lane et al. (1986). In this paper, we combine the PH model and the competing risks analysis and propose the competing risks proportional hazards (CRPH) model to estimate the time to default and early repayment, and capture the characteristics that influence the time to both events. The CRPH model facilitates estimating risk-adjusted returns and further risk-based pricing for online consumer loan portfolios.

We make three distinct contributions in this paper. First, the empirical hazard functions of default and prepayment are estimated through a nonparametric life-table method. Since no mathematical form or equation is assumed, this approach is more streamlined, not to mention faster in computation, making it a clear candidate for preliminary hazard analysis. It is found that the hazards of default and prepayment are distinct between the events and the terms of the loans. Since term structure matters in the estimation of returns of financial assets, it is necessary to model the time to default and prepayment by term, respectively. Few study has addressed the issue of loan terms in credit scoring and profit scoring before. Second, we employ a CRPH model to capture the characteristics which affect the time to default and prepayment, and estimate the survival probability for accounts at each month-on-book. Two dimensions of the predictive performance of the models are then examined. Third, we further forecast the loan's profitability in order to support profit-focused decisions, by combining the survival estimation of default and prepayment. Out-of-time forecasts show that our model has better predictive accuracy than the traditional profit scoring methods. Several insights are further obtained from the forecasts. For example, risky loans with lower grades may still be profitable. Thus, the loan grade given by the lending platform can be misleading to investors in terms of their profitability. After taking prepayment into account, the profitability forecast developed in this paper can be a good measure of an individual loan's profitability and thus be further employed to predict the return on a given portfolio. Inspired by the loan contract, we frame a possible *ex ante* contract to prevent early repayment and charge prepayment penalties when necessary.

The structure of this paper is as follows. The second section reviews the literature on credit risk in online lending. Section 3 describes the methodology. In Section 4, we present the datasets applied in this paper and perform preliminary analyses. In Section 5, we develop new credit scores and evaluate their performance. The profitability forecast and a method for charging prepayment penalties are given in Section 6. Our conclusions are drawn in the final section.

2. Literature review

In all forms of loans, the future payment behaviour of borrowers is uncertain given the asymmetric information. There are usually two events in the repayment period which may affect a lender's returns - default and prepayment. A great number of tools have been adopted to estimate the probability of default, and classify those 'bad' borrowers who cannot repay loans on time from those 'good' borrowers who can fulfil their obligations. These tools are typically credit scoring models, which help lenders or investors to make wise decisions. In particular, application scoring is the go-to tool for lenders to decide which borrowers may be issued a loan. We will review application scoring models for online lending first.

Statistical models have been the standards of the domain throughout the last four decades. Logistic regression (LR) and probit regression have been proven to be remarkably popular, since they are specially tailored for binary classification. In online lending, LR (Emekter et al., 2015; Serrano-Cinca et al., 2015) and probit regression (Dorfleitner et al., 2016; Hu et al., 2019; Michels, 2012) have also been applied extensively to predict the possibility of default. There are also many studies which are focused on proposing new default predictive models, and these also use LR as a benchmark to assess the performance of their models (Jiang et al., 2018; Malekipirbazari & Aksakalli, 2015). Some other statistical methods, such as linear regression (Iyer et al., 2016; Mild et al., 2015), have also been applied to modelling default risk for online loans. Linear regression is particularly suitable for a continuous measurement of default risk, such as the repayment ratio, which is a ratio of the sum of all payments received over the sum of all liabilities promised (Mild et al., 2015).

However, all the above-mentioned statistical methods are obliged to consider default within a certain time frame and thus cannot deal with censoring and time-dependent covariates. Consequently, some studies transcend these limitations by using survival analysis, especially the PH model (Duarte et al., 2012; Emekter et al., 2015; Lin et al., 2013; Serrano-Cinca et al., 2015; Xu & Chau, 2018). Survival analysis shows some clear advantages in predicting default. First, survival analysis can take any length of time horizon into consideration, since it models the time until an event occurs (Bellotti & Crook, 2009). Second, survival analysis can appropriately consider censored cases and incorporate time-dependent covariates (Allison, 2010). Examples of studies using survival analysis for consumer loans include Andreeva et al. (2007); Banasik et al. (1999); Bellotti & Crook (2009); Im et al. (2012); Malik & Thomas (2010); Stepanova & Thomas (2002); Stepanova & Thomas (2001). Banasik et al. (1999) are some of the first to apply survival analysis to predict consumer loans default and also prepayment. Their results suggest that survival analysis is competitive with the LR model in terms of identifying those who default in the first year. Im et al. (2012) provide evidence that incorporating the time dependency into survival analysis can provide more predictive accuracy and capture dynamic market effects on default behaviour.

In recent years, machine learning techniques or intelligent algorithms have come to represent another group of important methods for default risk classification. Due to the large amount of available data volumes and dimensions, machine learning algorithms are particularly popular in online lending, since they can achieve

relatively higher predictive accuracy compared to statistical methods (Ma et al., 2018; Malekipirbazari & Aksakalli, 2015; Wang et al., 2018). However, machine learning techniques have been criticised for their lack of transparency and potential overfitting (Anderson, 2007). Therefore, statistical methods are still preferred and more widely used today.

If we focus on the event of prepayment only, all the methods applied to default can be applied to prepayment prediction in a similar fashion, since it is essentially still a classification problem in which the primary goal is to assign a label to early repaid loans. However, we find that the event of prepayment, as a competing event of default and one of the possible real outcomes of a loan, has not received much attention in the field of online lending. As mentioned above, studies which have mainly focused on the event of prepayment generally come from the vantage point of the mortgage market (Beltratti et al., 2017; Mayer et al., 2013; Rose, 2013; Varli & Yildirim, 2015). Since prepayment would diminish the profits made from the loan (Banasik et al., 1999), the imposed penalties have been widely used in practice and discussed for mortgage loans (Mayer et al., 2013; Steinbuks, 2015; Varli & Yildirim, 2015). However, in the online lending markets, it is a worldwide standard not to apply a fee to prepayment, as the marginal cost of customer exit and acquisition is much cheaper than traditional lending (Li et al., 2019). Moreover, default and early settlement are both potential outcomes of each online loan, and the occurrence of any one of the two would terminate the contracted payments. An alternative method which would take these potential outcomes into consideration and forecast the outcomes within a framework would be to use multinomial logistic regression, as per Li et al. (2019). The multinomial logistic regression model can address the dependent variable of two or more levels (Li et al., 2019). However, the more popular method is to use competing risks models, as shown next.

Competing risks models have been widely adopted in banks' personal loans (Banasik et al., 1999; Stepanova & Thomas, 2002) and mortgages (Agarwal et al., 2006; Brown, 2016; Deng & Gabriel, 2006; Deng et al., 2000; Ergungor & Moulton, 2014; Quercia & Spader, 2008; Schmeiser & Gross, 2016; Steinbuks, 2015; Thackham & Ma, 2020). Many of these studies have applied a PH framework to competing risks analysis. In online lending, Zhang et al. (2019) propose a mixture cure PH model under competing risks, and fit their model using different assumptions based on online loan data. They have argued that their best-fitted model works well when compared with LR models in terms of both default and prepayment predictive accuracy. However, the limitations of their findings are apparent in that the mixture cure PH model used in their paper has strong assumptions on the model structure, and a Weibull function has to be predetermined as the parametric baseline hazard function. These assumptions are made arbitrarily, while the baseline hazard function is effectively unknown. Tan et al. (2019) use deep neural networks to characterise competing risks in online lending. The empirical experiments in their study show that their approach outperforms classical competing risks survival analysis in terms of risk prediction. Although they combine their machine learning algorithms with competing risks so that the predictive accuracy has been improved, their model still suffers from a lack of transparency and incapability of capturing the term structure of payments.

The competing risks approach has some potential advantages in credit scoring: first, it is specialised for handling several situations where the occurrence of one type of events removes the borrower from any risk from the other competing events (Allison, 2010). In the online lending scenario, this situation arises when default and prepayment are potential outcomes for fixed term loans. Second, competing risks analysis can identify why someone either defaults or pays off a loan given a series of covariates (Banasik et al., 1999). Third, it is able to forecast the survival probability of a loan, which depends on the timing of both events, whichever happening first being the most significant (Banasik et al., 1999; Stepanova & Thomas, 2002). This will further facilitate the estimation of risk-based returns on two types of risks. Last, since competing risks can be adapted to the PH framework, the competing risks approach can inherit all the advantages that the PH model presents.

It is argued that the time of full repayment and losses caused by prepayment cannot be neglected. However, the competing risks analysis of default and prepayment is limited and their corresponding influences on the returns of portfolios are unknown. Byanjankar & Viljanen (2020) use survival analysis-based monthly default probability to predict the expected returns of online loans, but they neglect the time to prepayment and fail to consider competing risks of default and prepayment. Only by forecasting the time of default and prepayment can we estimate the risk-adjusted returns (which is the real concern for lenders) more accurately. This paper adds to the literature on risk-adjusted returns of online loans by using a CRPH model to predict the time to default and prepayment, and by building a model to measure the profitability of online loans. We will introduce the CRPH model next.

3. Methodology

Let random variable T denote the time until an event, either default or prepayment, occurs. The probability distribution of T is described by a hazard function $h(t)$ and a survival function. The former is given by taking the limit of the conditional probability, that a case survives till t but fails in $t + \Delta t$:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t | T \ge t)}{\Delta t}, \tag{1}$$

while the survival function is defined as the probability of surviving past the time t : $S(t) = \Pr(T > t)$. In the lending context, $h(t)$ can be interpreted as the instantaneous probability that a borrower defaults (or prepays) at time t, given that the loan has survived until time t. In effect, $h(t)$ is not only a way of describing the probability distribution, but also shows that the hazard at any time point t corresponds directly to the risk of event occurrence at time t (Allison, 2010). The survival function, on the other hand, is a measure of the likelihood of an unsecured loan being current until at least time t. In addition, the cumulative distribution function $F(t)$ and its probability density function $f(t)$ can alternatively be expressed by the hazard function, as $F(t) = 1 - e^{-\int_0^t h(u)du}$ and $f(t) = h(t)e^{-\int_0^t h(u)du}$.

3.1. Proportional hazards model

If the hazard function $h(t)$ for each borrower is known at any time t, then the lenders would be able to make satisfactory decisions based on their own risk preference. We can estimate the hazard function of default or early repayment using the PH model with application characteristics and other available information:

$$h(t, \boldsymbol{x}) = e^{\boldsymbol{x}'\beta} h_0(t). \tag{2}$$

The PH model is proposed by Cox (1972), in which $\boldsymbol{x}$ is a vector of covariates, $\boldsymbol{\beta}$ is a vector of parameters to be estimated and $h_0(t)$ is a baseline hazard function. If additional information is available to determine which distribution $h_0(t)$ follows, such as being constant over time, the model may be estimated by the maximum likelihood method. Otherwise, the maximum likelihood procedure cannot be applied since the likelihood function is unknown. Nevertheless, Cox (1972) proposes the partial likelihood method to estimate $\boldsymbol{\beta}$ without understanding the form of the baseline hazard function $h_0(t)$. Let $\boldsymbol{X}_i$ be covariate vectors, and we order event occurrence times or censored times for each borrower i as t_i

$(i = 1, 2, \cdots k)$ and $R(t_i)$ is the set of borrowers at risk, namely those individuals who still have outstanding payments remaining at the time t_i. The conditional probability that borrower i will experience an event at time t_i, given the risk set $R(t_i)$, is

$$\frac{\exp(\boldsymbol{x'}_i\boldsymbol{\beta})h_0(t_i)}{\sum_{s\in R(t_i)}\exp(\boldsymbol{x'}_s\boldsymbol{\beta})h_0(t_i)} = \frac{\exp(\boldsymbol{x'}_i\boldsymbol{\beta})}{\sum_{s\in R(t_i)}\exp(\boldsymbol{x'}_s\boldsymbol{\beta})}. \tag{3}$$

The denominator in Eq. (3) is the sum of the hazards for all the borrowers who are at risk of the event at time t_i. Under the assumption that the observations are independent, the likelihood function of the observed data is

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{x'}_i\beta)}{\sum_{s\in R(t_i)}\exp(\boldsymbol{x'}_i\beta)}. \tag{4}$$

Apart from censoring, heavily tied survival times can be found, due to payments being recorded on a monthly basis. At this point, the likelihood function (4) needs to be further modified to include ties. Let us assume that there are d_i borrowers failing at time t_i. Let D_i denote the set of those d_i borrowers, where each element of D_i is a borrower who experiences the event at time t_i. $R(t_i, d_i)$ represents the set of all subsets of d_i borrowers at risk, each element of $R(t_i, d_i)$ is a combination of d_i borrowers who are at risk at time t_i . R_i, which represents the set of d_i borrowers at risk, is one of the elements of $R(t_i, d_i)$. Then, the likelihood function is modified as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left(\sum_{l\in D_i}\boldsymbol{x'}_l\boldsymbol{\beta}\right)}{\sum_{R_i\in R(t_i,d_i)}\exp\left(\sum_{s\in R_i}\boldsymbol{x'}_s\boldsymbol{\beta}\right)}. \tag{5}$$

The most common methods to deal with tied event times in the PH model are Breslow's approximation and Efron's approximation. Both methods are widely used to approximate the likelihood function (5) as they can handle computational difficulties. Breslow's approximation works well when ties are only a few while Efron's approximation is better when data are heavily tied (Allison, 2010). The two exact methods give exact partial likelihoods based on different assumptions on tied event times. Thus Eq. (5) itself is one of these exact methods, but the problem associated with Eq. (5) becomes increasingly computationally demanding as the ties get more. However, neither of these exact methods can escape time-consuming calculations (Allison, 2010). Efron's approximation will be exploited in this paper, as ties are heavy in our data. For the precise formulae for all of these techniques, please see Hsieh (1995).

In this paper, we also produce life-table estimates of the hazard function and product-limit estimators of the survival probability. The life-table estimate is calculated at the midpoint of each interval $[t, t+1)$, where $t = 0, 1, \cdots T'-1$ and T' is the maturity of the loan, by dividing the number of events by the total survival time within each interval. The product-limit estimator is a standard estimator of survival probability (Yang, 1997). For vectors of given covariates $\boldsymbol{x}$ and parameters $\hat{\boldsymbol{\beta}}$ estimated by the partial likelihood method, the product-limit estimator of the survival probability is given by

$$\hat{S}(t, \boldsymbol{x}) = \left[\hat{S}_0(t)\right]^{\exp(\boldsymbol{x'}\hat{\beta})}, \tag{6}$$

where $\hat{S}_0(t)$ is the baseline survival function derived from the likelihood function. For more details on this estimator please see Kalbfleisch and Prentice (2011).

3.2. Competing risks analysis

Once the loan is issued, each borrower may experience one of two types of events - either default or prepayment - during the repayment period, so those who have experienced one event would have no chance to be positioned in the other category before the loan reaches maturity, and vice versa. These are typical competing risks where two events compete as to which will occur first. Even if the balance is cleared off in collection, the account is still recorded as defaulted.

According to Kleinbaum & Klein (2012), a common way to analyse competing risks is to fit the PH models separately for each event and treat observations of the other competing events as right censored cases. There are two main drawbacks to this method - it requires there to be independent competing risks, and it is also somewhat questionable in terms of its interpretation of the survival curves obtained from fitting separate PH models (Kleinbaum & Klein, 2012). These drawbacks are not a problem in this paper. On the one hand, there is no evidence to show that default is related to prepayment, as they are regarded as contrary behaviours in lending. And on the other, the focus of this paper is not the survival curves obtained from the PH model. We follow this way to analyse the competing risks. Two binary dummies are defined to indicate censorship and event occurrence for default and prepayment, respectively when we fit the CRPH model in Eq. (8), and a subscript $j \in \{1, 2\}$ is used to denote the events (1 for default and 2 for prepayment). Now the type-specific hazards for event j can be defined as

$$h_j(t) = \lim_{\Delta t\to 0}\frac{\Pr(t \le T < t+\Delta t, J = j | T \ge t)}{\Delta t}, \quad j = 1, 2. \tag{7}$$

Accordingly, the CRPH model to be fitted is

$$h_j(t, \boldsymbol{x}) = e^{\boldsymbol{x'}\boldsymbol{\beta}_j}h_{0j}(t), \quad j = 1, 2. \tag{8}$$

and the product-limit estimator for event j can be derived from

$$\hat{S}_j(t, \boldsymbol{x}) = \left[\hat{S}_{0j}(t)\right]^{\exp\left(\boldsymbol{x'}\hat{\beta}_j\right)}, \quad j = 1, 2. \tag{9}$$

In addition, since both default and prepayment would result in account termination, the lifetime of loans can be predicted as $T = \min\{T_1, T_2, T'\}$, where T' is the maturity of the loan (Banasik et al., 1999).

4. Data and variables

The application and monthly repayment information of unsecured consumer loans from LendingClub, a market-leading lending platform, is analysed in this research. The loan accounts were generated between January 2013 and December 2017, and followed until December 2019. As briefly summarised in Table 1, out of 1,431,488 loans, 70.93 percent are 36-month term loans, and the remaining 29.07 percent are 60-month term loans.

It is noted that 'default' is defined by the platform as four consecutive missed monthly payments. The intermediate loans which are late by less than 120 days account for only 0.4% of the population, so have been left out of the sample. In the sample where the current loans are taken into account, the average default rate is 16.97%, which is much higher than the delinquency rate of federal commercial bank consumer loans over the same follow-up period – no more than 2.56%.[1] These online loans may generally be considered to be very risky loans. The average prepayment rate is 55.26%, i.e., more than half of borrowers do not wait until the loan is mature. This is much higher than that of mortgage loans.[2] If we

[1] Data was retrieved from FRED, Federal Reserve Bank of St. Louis, Board of Governors of the Federal Reserve System (US), Delinquency Rate on Consumer Loans, All Commercial Banks [DRCLACBS], https://fred.stlouisfed.org/series/DRCLACBS

[2] Since the prepayment rate of federal commercial bank consumer loans is not available, we would like to refer to mortgage loans' prepayment rate. According to the report on Fitch Ratings Global Housing and Mortgage Outlook (2020), the prepayment rate of American mortgage loans is about 20% over the same follow-up period.

Table 1
A summary of loan outcomes.

Term	Observations	Current	Defaulted	Prepaid	Fully paid	All
36-month	Number	88,084	141,054	588,244	198,010	1,015,392
	Proportion in total	6.15%	9.85%	41.09%	13.83%	70.93%
	Proportion in 36-month loans	8.67%	13.89%	57.93%	19.50%	100%
60-month	Number	98,789	101,857	202,863	12,587	416,096
	Proportion in total	6.90%	7.12%	14.17%	0.88%	29.07%
	Proportion in 60-month loans	23.74%	24.48%	48.75%	3.03%	100%
Total	Number	186,873	242,911	791,107	210,597	1,431,488
	Percentage	13.05%	16.97%	55.26%	14.71%	100%

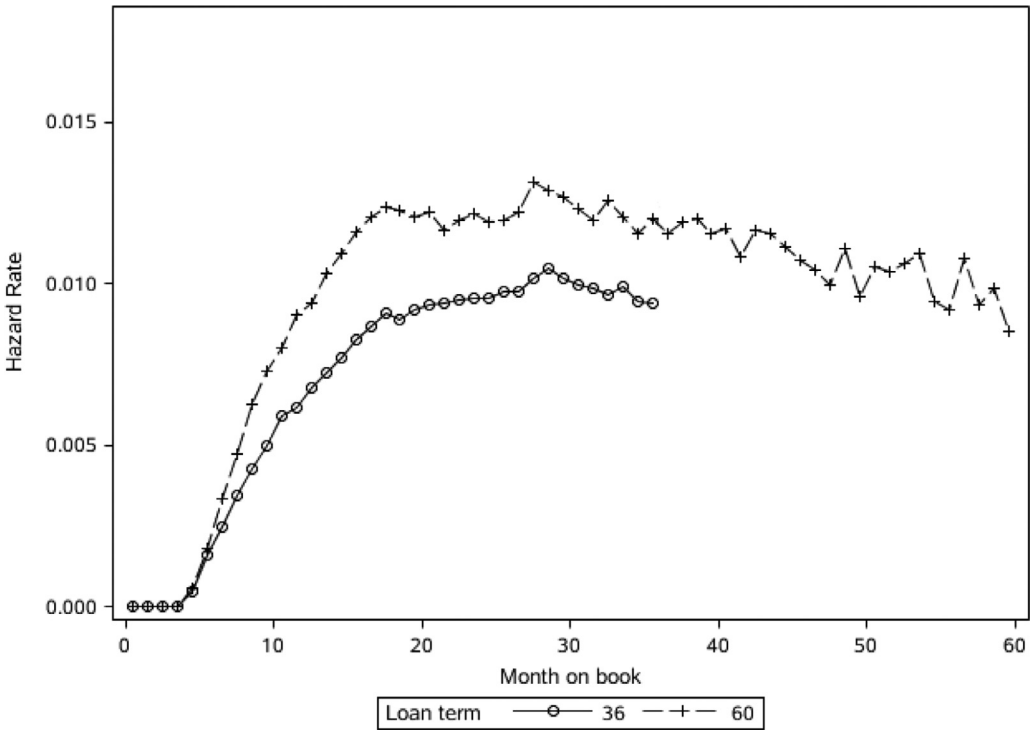


Fig. 1. Estimated hazard functions of default for two loan terms.

only look at the vintage loans with the final outcomes, those online loans' proportions would be even higher.

The hazard functions estimated by the life-table method are shown in Figs. 1 and 2. It is apparent in Fig. 1 that the hazard of default for the 36-month term group is lower than that of the 60-month group, and this gap is maintained until the 36-month term loan matures. Fig. 2 for prepayment also demonstrates a clear gap between two curves, suggesting that loan of two terms behave differently. Therefore, in the following analysis, loans with two types of terms are analysed separately (Panel A and Panel B).

In Figs. 3 and 4, we further estimate the hazard functions of default and prepayment for various loan grades, which have been qualified by the lending platform to capture the risk of default. Grades E, F, G are integrated into Grade EFG as the null group, since they represent only a small portion of all observations (9.4%). In Fig. 3(a) and (b), Grade A loans have the lowest default hazard over the payment period. The lower the grade, the higher the hazard of default. In Fig. 4(a) and (b), curves for prepayment are compact across grades, which shows loan grade has limited power in capturing the prepayment risk. Another phenomenon is clear that when approaching the maturity, the prepayment risk increases over time, i.e., good borrowers are impatient.

The main distinctions between 36-month and 60-month loans lie in the term structure, the loan amount, the interest rate on the loan, and associated default and prepayment risk patterns. In the analysis, the 36-month and 60-month loans are randomly stratified into two data subsets, 70% of observations are kept for training and 30% are for validation. All the models are fitted on the training sample and validated on the corresponding holdout sample. The covariates, including the application information and information regarding monthly repayment behaviours, are available in the database. The macroeconomic variables (MVs) have been extracted from the Economist Intelligence Unit (EIU) Countrydata database. A detailed description for each covariate is listed in Table A1 in the Appendix.

Note that we have not incorporated the interest rate of the loan in the model, since the interest rate is charged on the basis of the loan grade, i.e., the interest rate and the grade are highly correlated. The descriptive statistics of continuous variables are demonstrated in Table A2. On the one hand, those who have a 60-month loan generally make a demand for more loan amounts (approximately $8000) than those taking out a 36-month loan. On the other hand, 60-month loan borrowers have higher annual incomes, an average of $86,413, than those of 36-month loan borrowers (who have an average annual income of $77,491). Other noticeable differences include the fact that 60-month borrowers generally hold a credit history which is 10 months longer than 36-month borrowers. The average debt to income ratio is 17.85 for the 36-month group and 19.39 for the 60-month group. The number of open accounts and total accounts are also different.

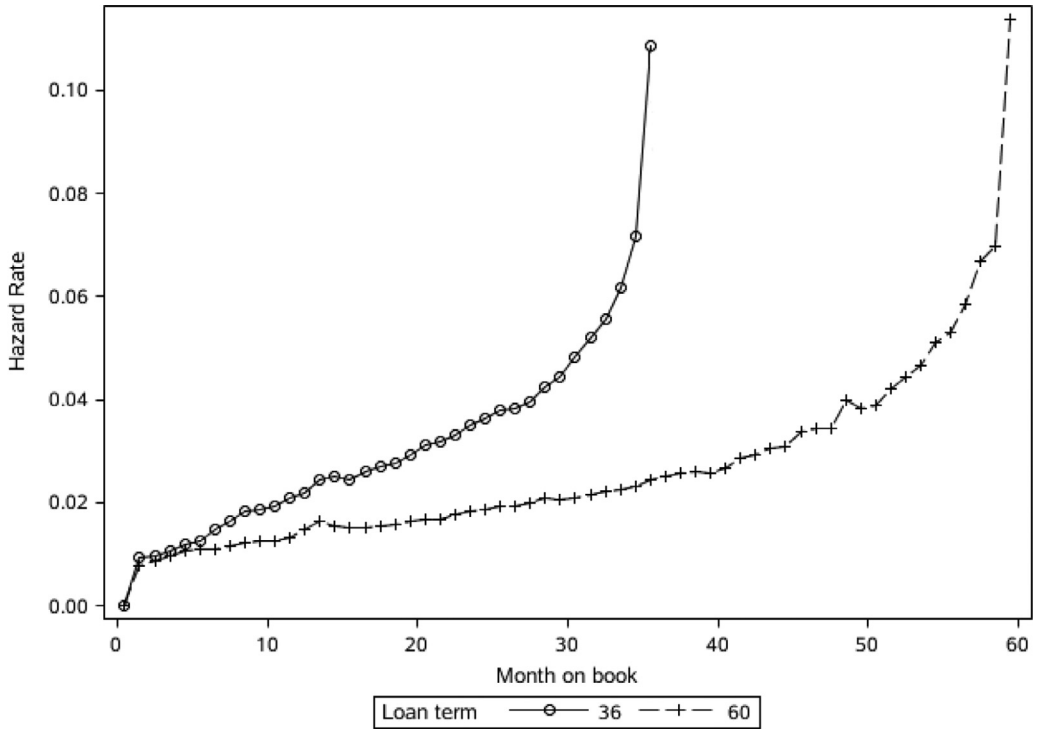
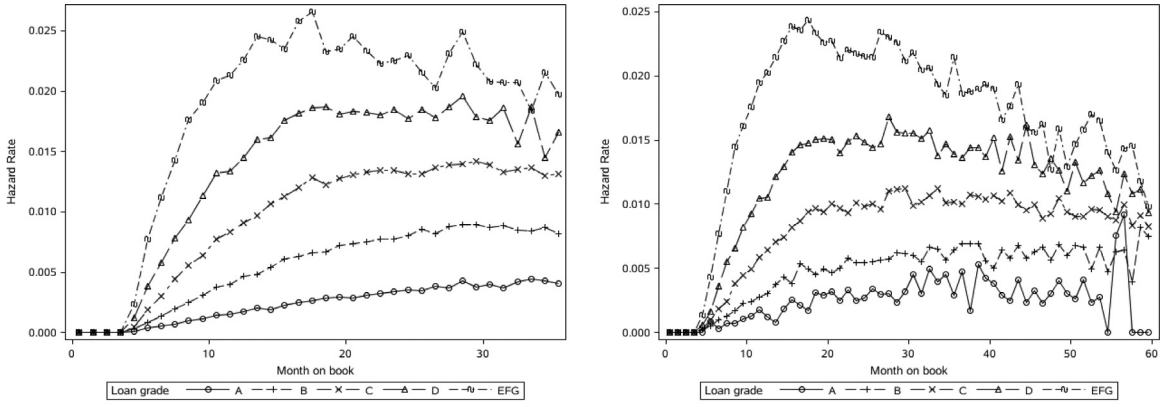
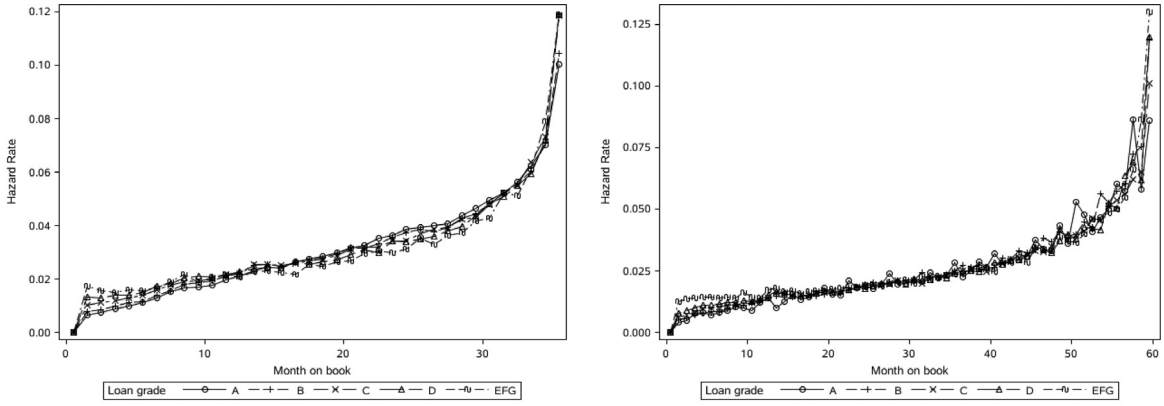


Fig. 2. Estimated hazard functions of prepayment for two loan terms.



(a) Panel A: 36-month loans (b) Panel B: 60-month loans

Fig. 3. Estimated hazard function of default for each grade.



(a) Panel A: 36-month loans (b) Panel B: 60-month loans

Fig. 4. Estimated hazard function of prepayment for each grade.

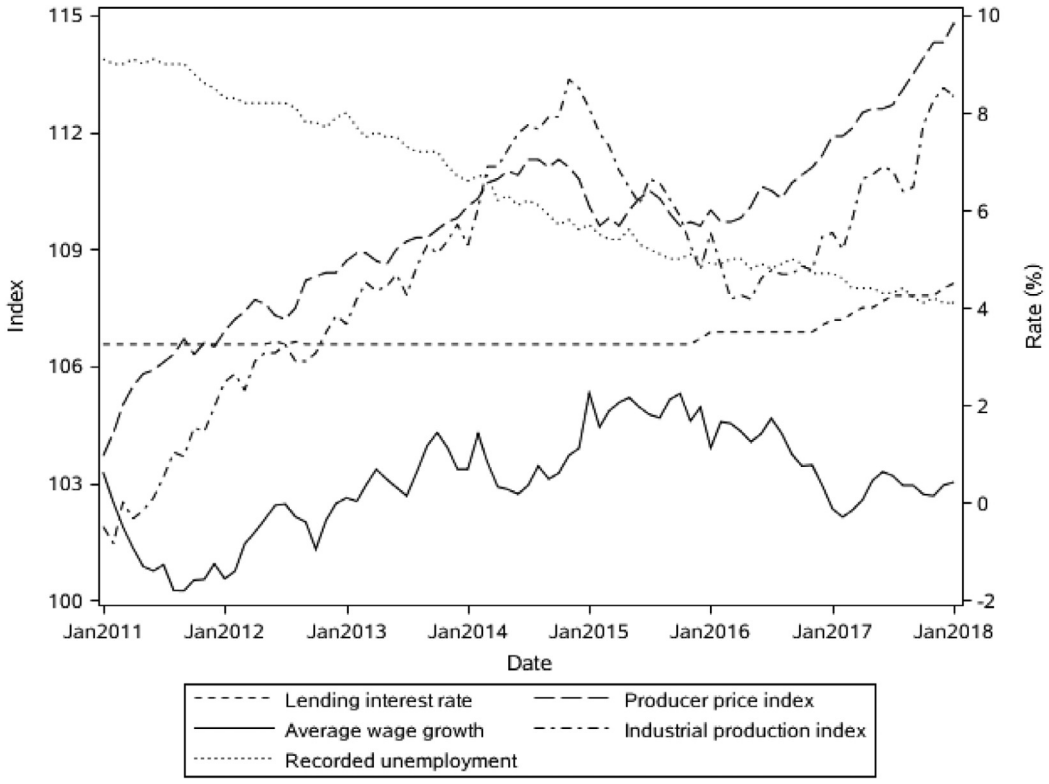


Fig. 5. Monthly macroeconomic variables of the US.

The descriptive statistics of MVs are not reported in Table A2, since these MVs are monthly time series data which have the same effect on both groups. However, graphs of the MVs time series are given in Fig. 5 to show the time trend of the MVs between January 2013 and December 2017 - the time period when loans were issued. Following Bellotti & Crook (2014), the difference in the value over a year is taken for the three MVs – the recorded unemployment, the producer price index and the industrial production index, to remove the time trend and reduce correlations in the time series. Some MVs, such as the GDP growth, may be good indicators of macroeconomic conditions. However, we have not included them in the model since they are not available on a monthly basis which is the type of granularity that lenders typically require (Bellotti & Crook, 2012).

The MVs have been incorporated as time-dependent covariates into the PH model for default prediction at the time of application (Bellotti & Crook, 2009). However, future values of MVs must be forecasted first when modelling in this way. In this paper, we will not predict future values of MVs, instead capturing the influence of economic conditions in the prediction of default, prepayment and losses at the time of application. Therefore, MVs recorded at time of application and lagged by 1 month, 3 months, 6 months are taken for analysis. The best lag structures will be chosen based on the Akaike Information Criterion (AIC) for different terms of the loan and different events, respectively.

5. Survival and competing risks analysis

5.1. Default prediction

We firstly use the CRPH model to fit to the training dataset. As shown in Fig. 1, in order to capture the difference between hazards, the model fitting will be conducted on 36-month and 60-month loans, respectively. Table 2 presents the results of the default event.

In Table 2, many covariates meet the 0.001 criterion for statistical significance. For example, an increment in loan amount to income, or debt to income, relates in a significant way to the increase of the hazard of default for both 36-month and 60-month loans. These are in accordance with Serrano-Cinca et al. (2015), who obtain positive estimates for loan amount to income and debt to income using the PH model without competing risks. The results of these grades are consistent with the evidence from Fig. 3, and show that a higher grade relates to a lower risk of default. Emekter et al. (2015) also find that the hazard rate of default decreases for every increase in credit grade by using the PH model without competing risks. In the category of credit history, having a higher value in public record bankruptcies, a delinquency of 2 years and an inquiry in the last 6 months all generally relate to a borrower's past bad credit activities, which suggests a higher risk of bad activity within the current line of credit, although a delinquency of 2 years for the 60-month loan term fails to meet the expectation of statistical significance. Serrano-Cinca et al. (2015) achieve the same result in terms of the estimates of the delinquency of 2 years and the inquiry in the last 6 months, although they do not consider cases which are censored from prepayment. Li et al. (2019) also obtain a positive estimate for both public record bankruptcies and inquiry in the last 6 months, though they model the default risk and prepayment risk with a multinomial logistic regression framework. Results of the other covariates, such as the home ownership, the borrower's employment length, the FICO score and the revolving line utilisation rate, are in accordance with the evidence from the LR model (Mo & Yae, 2022; Wang & Tong, 2020). Of those MVs, the lending interest rate, the average wage growth and the recorded unemployment are significant at the 0.001 level for both term loans. A higher lending interest rate and unemployment growth rate are expected to increase the hazard of default, since both situations represent an increased stress on the obligors (Bellotti & Crook, 2009, 2014). The average wage growth has a contrary effect on default between two panels. This is not surprising,

Table 2
Results of the CRPH model in modelling time to default.

	Panel A: 36-month loans				Panel B: 60-month loans			
Covariates	β	SE	Sig.	exp(β)	β	SE	Sig.	exp(β)
Loan features								
Loan amount to income	1.1112	0.0309	<0.0001	3.0378	0.7825	0.0420	<0.0001	2.1868
A	−1.6967	0.0181	<0.0001	0.1833	−1.7096	0.0488	<0.0001	0.1809
B	−1.0344	0.0133	<0.0001	0.3555	−1.1742	0.0167	<0.0001	0.3091
C	−0.5901	0.0122	<0.0001	0.5543	−0.7132	0.0105	<0.0001	0.4901
D	−0.2803	0.0127	<0.0001	0.7555	−0.3446	0.0100	<0.0001	0.7085
Purpose 1	0.1985	0.0191	<0.0001	1.2195	0.1397	0.0279	<0.0001	1.1500
Purpose 2	0.0348	0.0136	0.0103	1.0354	0.0320	0.0203	0.1140	1.0325
Purpose 3	0.1023	0.0083	<0.0001	1.1078	0.0512	0.0100	<0.0001	1.0525
Purpose 4	0.0917	0.0138	<0.0001	1.0960	0.0140	0.0168	0.4049	1.0141
Borrower characteristics								
Employment length	−0.0048	0.0009	<0.0001	0.9953	−0.0036	0.0011	0.0006	0.9964
Home_Mortgage	−0.1581	0.0083	<0.0001	0.8538	−0.1527	0.0095	<0.0001	0.8584
Home_Own	−0.0871	0.0107	<0.0001	0.9166	−0.0971	0.0137	<0.0001	0.9075
Debt to income	0.0079	0.0004	<0.0001	1.0079	0.0044	0.0005	<0.0001	1.0044
FICO	−0.0042	0.0002	<0.0001	0.9958	−0.0031	0.0002	<0.0001	0.9969
Tax liens	0.0374	0.0149	0.0123	1.0381	−0.0255	0.0202	0.2061	0.9748
Credit history								
Public record bankruptcies	0.0966	0.0151	<0.0001	1.1014	0.0740	0.0194	0.0001	1.0768
Credit history length	−0.0007	0.0000	<0.0001	0.9993	−0.0010	0.0001	<0.0001	0.9990
Delinquency 2y	0.0124	0.0033	0.0002	1.0124	−0.0032	0.0043	0.4539	0.9968
Inquiries last 6m	0.0698	0.0035	<0.0001	1.0723	0.0634	0.0041	<0.0001	1.0655
Open accounts	−0.0069	0.0008	<0.0001	0.9931	−0.0084	0.0009	<0.0001	0.9916
Public records	−0.0349	0.0132	0.0083	0.9657	−0.0099	0.0172	0.5639	0.9901
Revolving utilisation	−0.1929	0.0172	<0.0001	0.8246	−0.4296	0.0209	<0.0001	0.6508
Total accounts	0.0046	0.0004	<0.0001	1.0046	0.0047	0.0005	<0.0001	1.0047
Total current balance	−0.0002	0.0000	<0.0001	0.9998	−0.0003	0.0000	<0.0001	0.9997
Charge-offs within 12m	−0.0451	0.0278	0.1048	0.9559	−0.0182	0.0341	0.5931	0.9820
Number of accounts opened	0.0794	0.0019	<0.0001	1.0826	0.0848	0.0023	<0.0001	1.0885
Macroeconomic variables								
Lending interest rate	0.3285	0.0146	<0.0001	1.3889	0.3695	0.0182	<0.0001	1.4470
Average wage growth	−0.0698	0.0125	<0.0001	0.9326	0.0597	0.0148	<0.0001	1.0615
Recorded unemployment	0.0928	0.0137	<0.0001	1.0973	0.0695	0.0165	<0.0001	1.0719
Producer price index	−0.0561	0.0092	<0.0001	0.9455	0.0056	0.0106	0.5999	1.0056
Industrial production index	−0.0167	0.0026	<0.0001	0.9834	−0.0042	0.0030	0.1635	0.9958

Model fit statistics	AIC (without covariates)	AIC (with covariates)	Likelihood-ratio test	*p*-value
Panel A	2,560,568	2,515,265	45,365	<0.0001
Panel B	1,708,599	1,683,939	24,722	<0.0001

Note: In Panel A, the AIC values for the model with 1, 3 and 6 months lag on the MVs are 2,515,386, 2,515,458 and 2,515,274, respectively. However, in Panel B, the corresponding AIC values are 1,683,992, 1,684,011 and 1,684,045, respectively. Therefore, the MVs in Table 2 have not taken on a lag. The recorded unemployment, producer price index and industrial production index have already taken on twelve-month differences.

since the average wage growth have different marginal impacts on different income groups, as per Bellotti & Crook (2009).

In credit scoring, we are interested in the predictive ability of the models. The discriminant power needs to be assessed in a fixed time frame. We follow the way proposed by Banasik et al. (1999) to predict time to events with two time windows: (a) the loan will be defaulted in the first 12 months (Event A); (b) the loan which has survived 12 months, will be defaulted in the subsequent 12 months (Event B). Please be noted that both are conditional events which are not regular as simple good/bad labels in other credit scoring studies. In this paper, four classification methods - the LR model, support vector machine (SVM), artificial neural network (ANN) and random forest (RF) - are taken as benchmark models following prior studies (Dumitrescu et al., 2022). We also compare the CRPH model with the traditional PH model, i.e., the PH model without competing risks, using the above two criteria. The following are the settings and results.

The SVM is developed with a linear kernel function formulated as an optimisation problem, and the interior-point algorithm is frequently used to solve the optimal solution for SVM (Ferris & Munson, 2002). Following prior studies (Florez-Lopez & Ramon-Jeronimo (2014); Malekipirbazari & Aksakalli (2015)), the regularisation parameter C for SVM is set to be 1. Although various architectures are available for ANN, we choose the multilayer perceptron following Brown & Mues (2012). We consider one hidden layer of ANN in our classification since previous research has shown this is sufficient for credit risk models (Akkoç, 2012). We follow Florez-Lopez & Ramon-Jeronimo (2014) and test the number of hidden neurons in $[2, 10]$ based on the rule of thumb. The result shows that the best number is 5 or 6. In addition, a logistic activation function is applied to compute the output of hidden and output neurons (Brown & Mues, 2012). Following Malekipirbazari & Aksakalli (2015), we use the forest size of 200 trees since on another online loan dataset, they find that there is no further significant improvement in the accuracy with more trees beyond that. The number of randomly selected features to consider splitting on in each node is decided according to $\log_2 M + 1$, where M is the number of inputs (Breiman, 2001).

The predictive power of each model is evaluated on the test sample. More precisely, all the borrowers in the test sample are predicted as 'good' or 'bad' depending on their estimated default probability, obtained from the fitted model and a cut-off score assigned to each. Then several metrics of predictive accuracy are calculated. The Receiver Operating Characteristics (ROC) curve is a

Table 3
The AUC for predicting default in the first and second year.

Dataset	Event	LR	SVM	ANN	RF	PH	CRPH
Panel	Event A	0.7262	0.7240	0.6911	**0.7273**	0.7125	0.7207
A	Event B	**0.7025**	0.7008	0.6737	0.7021	0.6944	0.7011
Panel	Event A	**0.7210**	0.7149	0.6755	0.7194	0.6960	0.7168
B	Event B	**0.6912**	0.6867	0.6515	0.6884	0.6654	0.6902

common method used to assess the performance of a predictive model across all cut-offs, producing a measure defined as the Area Under the ROC Curve (AUC). The greater the AUC, the better the prediction is. An AUC equal to 1 indicates a perfect classification, while an AUC of 0.5 indicates no discriminant power, i.e. a purely random guess (Iyer et al., 2016). For more details about the ROC and the AUC please refer to Bauer & Agarwal (2014).

The AUC in Table 3 indicates that the CRPH model outperforms both the traditional PH model and ANN in classifying online loans as 'good' or 'bad' given the same time frames. Overall, the AUC of the CRPH model is slightly lower than that of LR and RF. The comparisons between CRPH and LR are in accordance with the results taken from data on bank loans in Banasik et al. (1999) and Stepanova & Thomas (2002). The poor performance of ANN compared to LR is not unusual, which can also be found in studies such as Brown & Mues (2012); Florez-Lopez & Ramon-Jeronimo (2014). In some literature (Dumitrescu et al., 2022; Malekipirbazari & Aksakalli, 2015), RF is found to outperform LR. But here, this only holds true for 36-month loans in predicting defaults during the first 12 months. In fact, as shown by Brown & Mues (2012), the performance of RF is affected by degrees of class imbalance (as is also the case for ANN and SVM). We have to argue that the overall AUCs of these models are not very high because we define two specific Events A and B, which are not the same as in other default prediction studies. In other studies, they generally take the final good/bad as labels but in this research, they are interim events conditional on the time frames. The results also suggest that no matter what the loan term is, the PH models and the four classification methods are better at predicting which loans will default in the first year than which loans will default in the second year. This result is reasonable, since timely information is essential for accurate prediction. All predictive features are taken at the application time in this research.

5.2. Prepayment prediction

In this section, we employ the CRPH model to predict the prepayment risk. A similar modelling process as used for default has been adopted. Specifically, the model will be fitted to the data on the training sample for each term, respectively, and make predictions on the test sample using Eq. (9). Similar to above, the PH models are compared to LR, SVM, ANN and RF by predicting the two events: (a) the loan will be prepaid in first 12 months (Event C); (b) the loan which survives over 12 months will be prepaid in the subsequent 12 months (Event D).

Table 4 presents the results of the CRPH model for prepayment. Several variables are found to be significant. For example, the loan amount to income ratio indicates that an increase in the variable relates to the reduction in the hazard of prepayment. The loan grade in particular shows some intriguing results. In Panel A, only grades A and B have any significantly different effects compared to the null category. In Panel B, higher grades are associated with a lower hazard of prepayment. Similar results have also been found in Li et al. (2019). In borrower characteristics, those whose homes are mortgaged or owned are more likely to prepay the loan than those who rent their house. Higher debt to income means a heavier burden of repayment, thus the borrower is less likely to prepay. The credit history involves many covariates that meet the 1% level of significance. These covariates show similar effect on two panels. For instance, whoever has a bad history for past credit activities, i.e., more delinquencies in the past 2 years or public records, represents a lower chance of prepayment. However, possibly out of the fear of the consequences of bankruptcy, those who have previously been bankrupted tend to repay the loan early, and thus extricate themselves from their financial obligations. Of the MVs, the lending interest rate, the average wage growth and the record of unemployment all have significant effects on the hazard of prepayment, and are consistent over both 30-month and 60-month term loans. For instance, a higher average wage growth is related to a higher hazard of prepayment, since families have more available funds to repay debts.

Table 5 shows the results of the models under the aforementioned two criteria. The process of setting parameters for SVM, ANN and RF are similar to those in Section 5.1. In general, the CRPH model generally outperforms the traditional PH model and ANN but demonstrates no advantages over the LR model, SVM and RF in predicting prepayment. Nevertheless, when comparing the AUC between Event C and Event D, respectively, it can be seen that predictions one year in advance are considerably better than those two years in advance. However, if the information generated during the first year of repayment has been collected, this situation may change.

5.3. Comparing predictive accuracy

The previous two sections examine the performance of predictive models from the perspective of the discriminant power of a binary outcome, and we find that the CRPH model displays only limited competence in predicting the occurrence of two events. However, the CRPH model can predict dynamically given the survival duration, which is not available for these widely used classification methods. In this section, we will present the performance of the CRPH model in time series, which provides us with some new insights.

Since the CRPH model can forecast survival probability for each month-on-book, a good measure of the accuracy of the probabilistic forecast is the Brier Score (BS), which is defined as

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(P_i - Y_i)^2, \tag{10}$$

where n is the number of borrowers, P_i is the predicted probability for default or prepayment, Y_i takes a value of 1 if borrower i has experienced default (or prepayment) and 0 otherwise. The BS varies from 0 to 1, with values close to 0 for well-performing models. For dynamic prediction over the time, the BS at the mth month-on-book is

$$\mathrm{BS}_m = \frac{1}{n}\sum_{i=1}^{n}(P_{im} - Y_{im})^2 = \frac{1}{n}\sum_{i=1}^{n}(1 - S_{im} - Y_{im})^2, \tag{11}$$

where S_m is the predicted probability of surviving up to the $m$thmonth-on-book, i.e., the predicted probability that they have not defaulted (or prepaid). More details about the BS in Eqs. (10) and (11) can be referred to Ateca-Amestoy & Prieto-Rodriguez (2013).

Table 4
Results of the CRPH model in modelling time to prepayment.

	Panel A: 36-month loans				Panel B: 60-month loans			
Covariate	β	SE	Sig.	exp(β)	β	SE	Sig.	exp(β)
Loan features								
Loan amount to income	−0.4050	0.0163	<0.0001	0.6670	−0.3845	0.0299	<0.0001	0.6808
A	−0.2021	0.0100	<0.0001	0.8170	−0.5249	0.0194	<0.0001	0.5916
B	−0.0711	0.0093	<0.0001	0.9314	−0.3639	0.0102	<0.0001	0.6949
C	−0.0122	0.0092	0.1859	0.9879	−0.2258	0.0079	<0.0001	0.7979
D	0.0051	0.0099	0.6089	1.0051	−0.1123	0.0083	<0.0001	0.8938
Purpose 1	−0.0111	0.0140	0.4249	0.9889	−0.0127	0.0346	0.7137	0.9874
Purpose 2	−0.0252	0.0064	<0.0001	0.9752	−0.0832	0.0116	<0.0001	0.9202
Purpose 3	−0.0504	0.0092	<0.0001	0.9509	−0.1243	0.0190	<0.0001	0.8831
Purpose 4	0.0762	0.0038	<0.0001	1.0792	0.0335	0.0067	<0.0001	1.0340
Borrower characteristics								
Employment length	−0.0076	0.0004	<0.0001	0.9924	−0.0005	0.0008	0.5151	0.9995
Home_Mortgage	0.0942	0.0039	<0.0001	1.0988	0.1486	0.0069	<0.0001	1.1602
Home_Own	0.0239	0.0055	<0.0001	1.0241	0.1011	0.0102	<0.0001	1.1064
Debt to income	−0.0076	0.0002	<0.0001	0.9924	−0.0125	0.0004	<0.0001	0.9876
FICO	0.0019	0.0001	<0.0001	1.0019	0.0024	0.0001	<0.0001	1.0024
Tax liens	−0.0258	0.0084	0.0021	0.9745	0.0024	0.0170	0.8878	1.0024
Credit history								
Public record bankruptcies	0.1252	0.0081	<0.0001	1.1334	0.1582	0.0163	<0.0001	1.1714
Credit history length	−0.0010	0.0000	<0.0001	0.9990	−0.0009	0.0000	<0.0001	0.9991
Delinquency 2y	−0.0417	0.0019	<0.0001	0.9591	−0.0550	0.0035	<0.0001	0.9464
Inquiries last 6m	−0.0008	0.0019	0.6902	0.9992	−0.0039	0.0032	0.2215	0.9962
Open accounts	−0.0306	0.0004	<0.0001	0.9699	−0.0286	0.0007	<0.0001	0.9718
Public records	−0.0261	0.0072	0.0003	0.9742	−0.0780	0.0146	<0.0001	0.9249
Revolving utilisation	−0.3693	0.0085	<0.0001	0.6912	−0.4611	0.0150	<0.0001	0.6306
Total accounts	0.0179	0.0002	<0.0001	1.0181	0.0165	0.0003	<0.0001	1.0166
Total current balance	0.0002	0.0000	<0.0001	1.0002	0.0003	0.0000	<0.0001	1.0003
Charge-offs within 12m	0.0230	0.0138	0.0958	1.0233	0.0320	0.0246	0.1931	1.0325
Number of accounts opened	0.0347	0.0010	<0.0001	1.0354	0.0278	0.0018	<0.0001	1.0282
Macroeconomic variables								
Lending interest rate	0.2096	0.0105	<0.0001	1.2331	0.1109	0.0131	<0.0001	1.1173
Average wage growth	0.1021	0.0048	<0.0001	1.1075	0.0283	0.0105	0.0068	1.0287
Recorded unemployment	0.0144	0.0065	0.0274	1.0145	0.1306	0.0116	<0.0001	1.1395
Producer price index	0.0660	0.0039	<0.0001	1.0682	−0.0121	0.0075	0.1053	0.9879
Industrial production index	−0.0188	0.0012	<0.0001	0.9814	0.0192	0.0021	<0.0001	1.0193

Model fit statistics	AIC (without covariates)	AIC (with covariates)	Likelihood-ratio test	*p*-value
Panel A	10,634,927	10,605,778	29,210	<0.0001
Panel B	3,356,621	3,344,388	12,295	<0.0001

Note: In Panel A, the MVs have taken on a lag of 6 months, as the AIC values for the model without lag and with the 1 or 3 month lag on the MVs are 10,605,912, 10,606,308 and 10,606,438, respectively. However, in Panel B, MVs in Table 4 have not taken on a lag, the corresponding AIC values for model with the 1, 3 and 6 months lag on the MVs are thus 3,344,475, 3,344,536 and 3,344,472, respectively. The recorded unemployment, the producer price index and the industrial production index have already taken on twelve-month differences.

Table 5
The AUC for predicting prepayment in the first and second year.

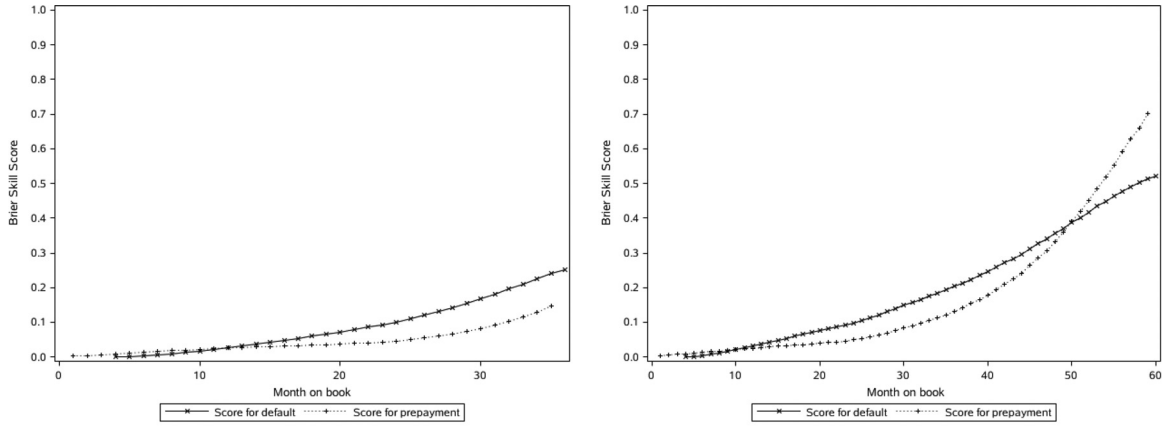
Dataset	Event	LR	SVM	ANN	RF	PH	CRPH
Panel	Event C	**0.6305**	0.6301	0.5894	0.6248	0.6233	0.6230
A	Event D	**0.5915**	0.5910	0.5583	0.5901	0.5849	0.5880
Panel	Event C	**0.6440**	0.6406	0.5889	0.6322	0.6334	0.6376
B	Event D	**0.6047**	0.6038	0.5576	0.6008	0.5950	0.6040

When using the BS to compare probabilistic forecasts for distinct sets of situations or events, attention must be paid to separate the confounding effects of intrinsic predictability and predictive performance (Gneiting & Raftery, 2007). Confounding effects tend to be caused by the extraneous factors when comparing the predictive performance of the model for different events, leading to biased results. For example, the class distributions of default and prepayment are significantly different (see Table 1), and the predictive performance is inevitably influenced due to the imbalanced sample (Lin et al., 2017). Thus, the superior performance of BS for default prediction might be associated with the inferior performance for prepayment prediction. To address this issue, we standardise the BS by the score under a noninformative model, according to Gneiting & Raftery (2007), which is called the Brier Skill Score (BSS)

$$\mathrm{BSS}_m = 1 - \frac{\mathrm{BS}_m}{\mathrm{BS}_m^{ref}}, \tag{12}$$

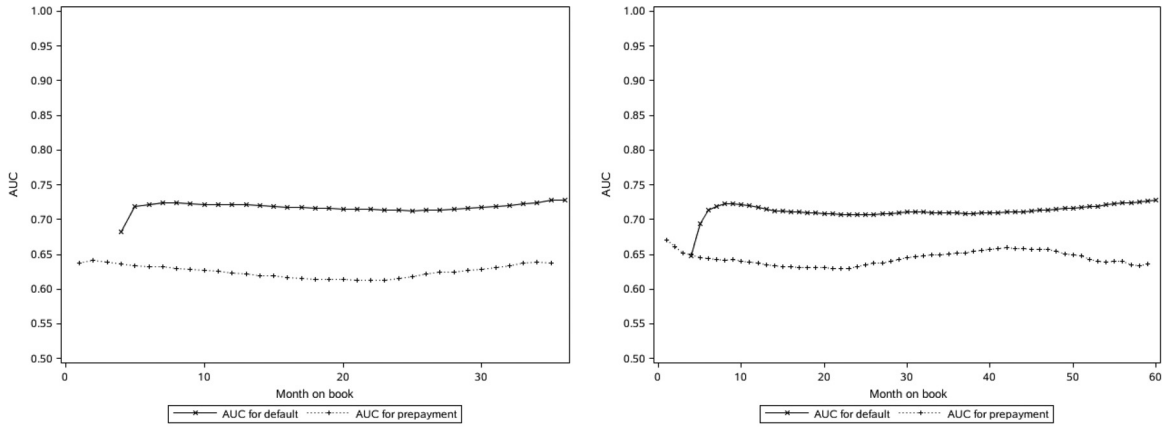
where BS_m^{ref} is the BS for a reference model, i.e., a noninformative model, calculated based on the observed unconditional probability of the event in each of those months-on-book. The higher value of BSS, the better the probabilistic forecast. A value of 1 indicates a perfect forecast, whereas a value of 0 represents a forecast from the null model.

Fig. 6 shows the BSS at each month-on-book for both default and prepayment. Fig. 6 shows that the BSS for default is higher than the BSS for prepayment most of the time. It means that, overall, the CRPH model performs better in predicting the probabil-



(a) Panel A: 36-month loans (b) Panel B: 60-month loans

Fig. 6. Brier Skill Score at each month-on-book.



(a) Panel A: 36-month loans (b) Panel B: 60-month loans

Fig. 7. AUC at each month-on-book.

ity of default than the probability of prepayment. This is due to the fact that much of the application information, especially the credit history, is provided for the purpose of estimating default risk rather than prepayment risk. However, if the period is longer than 50 months-on-book for the 60-month loans, the BSS for prepayment prediction is significantly higher than that for default prediction, thus the predicted probability of prepayment is much more accurate. This indicates that the loan term is a significant factor influencing the probability forecast for default and prepayment. The BSS in Fig. 6 presents an intuitive illustration on the predictive power of the model, in terms of how well the probabilistic forecasts match the outcomes.

Apart from the BSS, we also analyse the AUC dynamically, depending on the borrowers surviving through the previous months-on-book. These borrowers are first ranked at each month-on-book according to their monthly estimated survival probability of both default and prepayment. We then pin the true labels (defaulted 0/1, prepaid 0/1) to each loan instance. Last, an AUC from the ROC curve at each month-on-book for both events can thus be obtained. We plot the AUC values over time in Fig. 7, where the AUC for default is always higher than that for prepayment, for both terms of loans. The gap of AUC between the default and prepayment is consistently larger than 0.05. These demonstrate that it is not easy to distinguish prepaid and fully-paid loans on maturity using the application information. The performance is weaker than classification of defaulted and fully-paid loans, as the former task is to distinguish 'good' borrowers from 'good but impatient' ones.

Here these measures of predictive power (the BSS and AUC) both suggest that the CRPH models have a better ability in predicting default than predicting prepayment. This is not surprising, since most application information and bureau data refers to default risk rather than the prepayment risk. If more relevant information regarding the prepayment behaviour has been collected, the risk of prepayment could be better forecasted.

6. Risk-based profitability

In this section, we propose a method for forecasting the profitability of a loan based on the dynamic estimates obtained in the previous sections. Profitability is related to when the event occurs, thus it cannot be estimated using classification methods, but can be done much more intuitively by using survival models. In concrete terms, we firstly derive an equation to estimate the expected profit percentage as a base of profitability forecast of each loan only taking default into account, which is the traditional way of credit scoring, as delineated by Guo et al. (2016). And then, losses resulting from early repayment are evaluated, and a feasible way to charge prepayment penalty is thus framed. Finally, we predict profitability in the light of both default and prepayment risk and evaluate the profitability of loan portfolios to meet a lender's profit-focused investment decisions.

6.1. Expected returns with default

Let $\hat{S}_{t,j}$ denote the credit score developed based on the survival probability at time t for loan j. Note that $\hat{S}_{t,j}$ is also a product-limit estimator of the survival function in the event of default. Thus, the expected return R_j of loan j can be calculated using the equation in Stepanova & Thomas (2001):

$$R_j = \sum_{t=1}^{N_j} \frac{\hat{S}_{t+3,j} I_{t,j}}{(1+r_t^0)^t} - K_j, \tag{13}$$

where N_j is the term of the jthloan: 36 or 60 months in this paper; $I_{t,j}$ is the monthly instalment; K_j is the loan amount of jth loan, and r_t^0 is the monthly discount rate obtained from the 3-year and 5-year treasury yields and matched with the terms of the loans. The subscript of $\hat{S}$ is $t+3$ because default is defined as four consecutive missed payments. Thus, survival up to time $t+3$ is equivalent to default at time $t+4$, and the last instalment paid is four months prior to default, i.e., at time t. The expected profit in Eq. (13) is computed by the sum of the present values of expected monthly instalments minus the total amount of the investment.

One might consider that Eq. (13) underestimates the return, since the recoveries from defaulted loans have not been taken into account. These are neglected for several reasons. The Loss Given Default (LGD) of an online loan, a measure of the percentage of exposure that would not be collected after default occurs, is generally high (Xia et al., 2021). Any potential recoveries are therefore minor and have limited impact on overall profitability. Also, recoveries are very difficult to predict at the time of application. Even if there is post-default information available (Li et al., 2021), such as information on the collection process, LGD remains difficult to estimate (Zhang & Thomas, 2012). Furthermore, the dates of collections, which are used to discount the future value of the recoveries to the present value, are not available in our dataset. Predicting recoveries and LGD is in fact another topic in credit risk research. Extant studies generally predict LGD at the time of default rather than at the time of application (Bastos, 2010; Bellotti & Crook, 2012; Bellotti et al., 2021; Do et al., 2018; Tobback et al., 2014; Yao et al., 2017). We can collect as much information facilitating LGD prediction as possible until final default occurs. Therefore, at the time of application, it is probably better to concentrate on profit from the loans themselves, without taking recoveries into consideration.

Since the average-capital-plus-interest repayment method is applied by the lending company, the instalment $I_{t,j}$ is fixed for each month-on-book and can be obtained by using the following equation:

$$I_{t,j} = K_j \cdot \frac{r_j}{1-(1+r_j)^{-N_j}} \quad \text{for all } t \in \{1, 2, \cdots, N_j\}, \tag{14}$$

where r_j denotes the monthly interest charged. Therefore, the profit percentage is computed as

$$\gamma_j = \frac{R_j}{K_j} = \frac{r_j}{1-(1+r_j)^{-N_j}} \cdot \sum_{t=1}^{N_j} \frac{\hat{S}_{t+3,j}}{(1+r_t^0)^t} - 1. \tag{15}$$

In Eq. (15), the expected profit is standardised by the amount invested, i.e., the return for each unit of invested money. Therefore γ_j captures the percentage of expected profit in terms of loan amount.

It appears in Eq. (15) that the expected profit percentage is positively affected by the interest charged. In fact, raising the current interest rate may not actually improve the profit percentage. As emphasised by Serrano-Cinca et al. (2015), the higher the interest rate in lending, the higher the default risk. Hence, the survival probability and possible profit percentage of a loan declines as the interest rate increases. A discussion on the pricing of each online loan to maximise profit is certainly of value, though beyond the scope of this study. Nevertheless, we evaluate the profit percentages on each credit grade by plotting the violin plots in Fig. 8.

As illustrated in Fig. 8, each violin plot represents the density trace by demonstrating the distributional characteristics, symmetrically on the right and the left side, and quantiles as line boundaries of horizontal stripes. From the bottom to the top, the horizontal lines indicate the 10% quantile, first quartile, median, third quartile and 90% quantile values. The black line in each violin plot is effectively a small neighbourhood of medians. The 'thicker' the black line looks, the more scattered the profit percentages around the median are. For more details on violin plots, see Hintze & Nelson (1998).

In Panel A, as depicted in Fig. 8(a), over 90 percent of Grade A loans yield a positive profit. This number drops as the grade becomes lower. For Grade C, just 75 percent of loans would generate positive profits. And for those lower than Grade D, only about 50 percent would yield positive profits. Interestingly, in these subprime ratings of EFG, nearly 10 percent of loans will generate at least 10% profits. In contrast, in the safest category A, few loans produce such a high profit percentage. This shows that some of the riskier loans may in fact be profitable while some of the prime loans may bring about losses. The violin plots for Panel B have similar patterns but are spread more widely. The credit grades provided by the platform may fail to fully recognise their potential profits and losses, while the lower the credit grade, the larger the deviation of the profits are. Since a higher grade does not necessarily lead to higher profits, the institutional investors who build their portfolios on the loans with higher grades may overestimate their potential returns, plus potential losses may occur due to prepayment, as illustrated in the next section.

In addition, the distribution of the profit percentage for each grade is unimodal and unsymmetrical. For both loan terms, a lower grade is represented by a distribution with a lower peak, and fatter and longer tails.

6.2. Losses resulting from prepayment

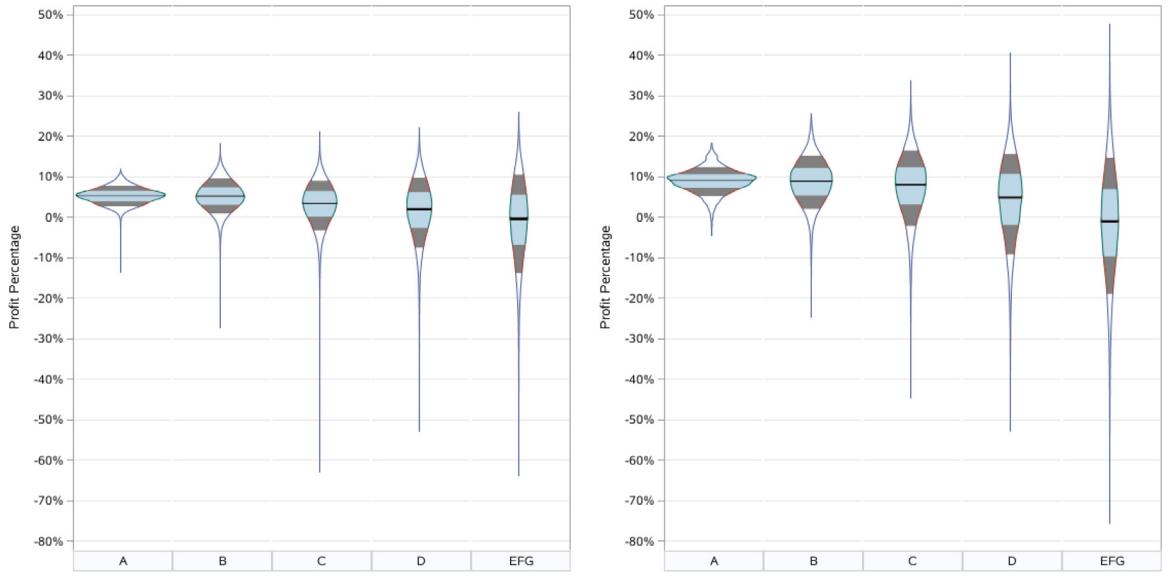
In the previous section, we estimated expected return by considering default event only. In this section, losses resulting from prepayment are estimated and incorporated into profit percentage estimates. Once prepayment occurs, not only the time of the event, but also the revenue can be easily assessed and calculated *ex post*. Assuming that the prepayment occurs at n_jth ($n_j < N_j$) month-on-book for loan j, the net revenue is

$$R'_{n_j,j} = \sum_{t=1}^{n_j} \frac{I_{t,j}}{(1+r_t^0)^t} + \frac{O_{n_j,j}}{(1+r_t^0)^{n_j}} - K_j + \varepsilon_j, \tag{16}$$

where $O_{n_j,j}$ is the outstanding balance or the remaining principal that is prepaid. For the average-capital-plus-interest repayment method, this is given by:

$$O_{n_j,j} = K_j \cdot \left[\frac{1-(1+r_j)^{n_j-N_j}}{1-(1+r_j)^{-N_j}} \right]. \tag{17}$$

The term ε_j represents an opportunity to gain profits by reinvesting the prepaid money $O_{n_j,j}$in new assets. There are generally two reinvestment options available to the investor. For retail investors, they can reinvest the money in other available assets on the lending platform, depending on their personal preferences. For institutional investors, they can only hold the funds as cash or deposits until the portfolio is fully terminated. If they decide to reinvest, the new assets are considered as a separate portfolio unrelated to the current one, as wholesale loans described on the



(a) Panel A: 36-month loans (b) Panel B: 60-month loans

Fig. 8. Violin plots of estimated profit percentages across loan grade.

platform. According to LendingClub's official report[3], in recent 10 years, over 90% of the loans are wholesale loans managed by institutional investors such as banks, insurance companies and asset managers, but not by retail investors. Therefore, we only consider the prepaid cash as risk-free assets. At the time of maturity, they will obtain an amount of $O_{n_j,j} \cdot (1 + r_t^0)^{N_j - n_j}$. We take the 3-year and 5-year treasury yields as risk-free returns. The present value of the revenue is equal to $O_{n_j,j} \cdot (1 + r_t^0)^{-n_j}$, thus the net revenue from this additional opportunity is actually equal to 0, i.e., $\varepsilon_j = 0$. Since most loans are funded by institutional investors at LendingClub, reinvesting the money $O_{n_j,j}$ in risk-free assets is assumed to be the optimal choice, to be hereafter taken as $\varepsilon_j = 0$.

If the revenue $R'_{n_j,j}$ is less than the *ex-ante* expected profit R_j in formula (13) without the consideration of prepayment, it means that prepayment may result in profit reduction, and a penalty charge for this becomes increasingly necessary. On the contrary, between two events, whichever happens first, then the other fails. In the circumstance that prepayment occurs, it can be regarded as an inherent mechanism to lower the losses resulting from possible default, and thus should not be punished. Taken together, the loss resulting from prepayment is defined by the equation:

$$L_{n_j,j} = \max\left(0, R_j - R'_{n_j,j}\right). \tag{18}$$

When a difference exists between R_j and $R'_{n_j,j}$, i.e., $L_{n_j,j} > 0$, a possible way is recommended to prevent and compensate the loss from prepayment. For mortgages, lenders often charge *ex ante* higher mortgage premiums to compensate for prepayment risk (Mayer et al., 2013). However, in online lending one unfortunately cannot easily follow this way. Once the lender charges additional premiums for prepayment, the increasing interest rates could undermine the affordability of the loan to the borrower, leading to the adverse selection since borrowers can easily compare prices of loans online.

A plausible alternative way to compensate the loss from prepayment is to offer an *ex ante* contract or provision, to charge a penalty conditional on *ex post* losses $L_{n_j,j}$ if prepayment occurs. This approach does not change the interest rates and the cost of the loan would not increase even when prepayment occurs, as long as the prepayment does not result in profit losses. Specifically, anyone considering offering an *ex-ante* contract or provision would need to initially evaluate the potential losses $L_{n_j,j}$, given that prepayment occurs at the n_jth month-on-book. At the time of offering an *ex-ante* contract, the value of n_j is virtually unknown. Therefore, the *ex-ante* contract may stipulate the penalties at every possible month-on-book (from the 1st to 35th month-on-book for 36-month loans and from the 1st to 59th month-on-book for 60-month loans) according to $L_{n_j,j}$, which would be charged once prepayment occurs. While this contract might seem cumbersome, it can be simplified by focusing only on the starting n_p months-on-book at which $L_{n_p,j}$ is large enough, for example, $n_p = 6$. Although such a conditional penalty has an advantage over premiums in that the penalty does not increase the cost of the loan, we have no source to test the effectiveness and efficiency of this proposition in this paper. Therefore, we argue that the actual decision to use interest premiums or penalties to compensate for the loss from prepayment is ultimately an operational decision of the lending management team.

Finally, we note that for those with $R_j \leq 0$, attention should be given to more severe default risk, which represents a much lower value of $\hat{S}_{t,j}$, rather than the prepayment risk, thus the penalty term $L_{n_j,j}$ is always equal to 0.

6.3. Profitability estimates

One may consider utilising the profit percentage γ_j in Eq. (15) as a possible solution to profitability forecasting for each loan. However, the primary issue is that γ_j fails to account for the potential loss caused by prepayment. The event of prepayment should not be neglected, since it sometimes indeed results in losses ($L_{n_j,j} > 0$).

When default and prepayment are both taken into account, Eq. (19) provides a solution to forecast the profit percentage over the

[3] The Investor Roadshow Presentation 2019 by LendingClub can be available at https://ir.lendingclub.com/financials/QuarterlyResults/

Table 6
Out-of-time predictive performance on different profit buckets.

	Bucket	N	Mean profitability	$\hat{p}_{OLS}$		$\hat{p}_{PH}$		$\hat{p}^{D}_{CRPH}$		$\hat{p}^{D+P}_{CRPH}$	
				RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Panel A	P1	8147	−0.0991	0.1638	0.1222	**0.1380**	**0.1009**	0.1669	0.1250	0.1661	0.1237
	P2	8147	0.0328	0.0117	**0.0094**	0.0366	0.0320	**0.0114**	0.0096	0.0123	0.0107
	P3	8147	0.0487	0.0247	0.0229	0.0568	0.0526	**0.0226**	**0.0203**	0.0238	0.0218
	P4	8147	0.0658	0.0417	0.0406	0.0757	0.0725	**0.0353**	**0.0336**	0.0365	0.0352
	P5	8147	0.0877	0.0698	0.0679	0.1201	0.1157	**0.0652**	**0.0618**	0.0659	0.0627
	All	40,735	0.0272	0.0827	0.0526	0.0936	0.0747	0.0825	**0.0501**	**0.0824**	0.0508
Panel B	P1	1898	−0.1224	0.1413	0.1354	**0.1110**	**0.1017**	0.1519	0.1460	0.1503	0.1445
	P2	1898	0.0104	**0.0251**	**0.0189**	0.0398	0.0321	0.0317	0.0239	0.0299	0.0222
	P3	1898	0.0517	0.0335	0.0303	0.0622	0.0576	**0.0258**	**0.0217**	0.0275	0.0240
	P4	1898	0.0773	0.0605	0.0585	0.0896	0.0859	**0.0487**	**0.0455**	0.0505	0.0479
	P5	1897	0.1114	0.1002	0.0986	0.1356	0.1329	**0.0879**	**0.0850**	0.0890	0.0864
	All	9489	0.0257	0.0842	0.0683	0.0940	0.0820	0.0835	**0.0644**	**0.0833**	0.0650

repayment period:

$$\tilde{\gamma}_j = \gamma_j - \sum_{n_j=1}^{N_j} \frac{L_{n_j,j} \cdot \left(\hat{S}'_{n_j-1,j} - \hat{S}'_{n_j,j}\right)}{K_j}$$
$$= \gamma_j - \sum_{n_j=1}^{N_j} \frac{\max\left[\left(\hat{S}'_{n_j-1,j} - \hat{S}'_{n_j,j}\right) \cdot \left(R_j - R'_{n_j,j}\right), 0\right]}{K_j}, \quad (19)$$

where $\hat{S}'_{n_j,j}$ denotes the survival probability for an event of prepayment at time n_j for loan j, $\hat{S}'_{n_j-1,j} - \hat{S}'_{n_j,j}$ thus gives the probability of prepayment occurring in the n_jth month-on-book.

There are some clarifications which need to be set out for Eq. (19). First, the independence of competing risks is assumed. Otherwise, the term $\hat{S}'_{n_j-1,j} - \hat{S}'_{n_j,j}$ needs to be improved in order to incorporate the influence of default risk. Second, our main focus is still the profit percentage γ_j, and the higher this is, the more profitable the loan. Since the occurrence of prepayment at the n_jth month-on-book would result in losses $L_{n_j,j}$, the expected losses caused by the prepayment are then calculated in sums and standardised by K_j. Finally, the adjusted profit percentage $\tilde{\gamma}_j$ contains a penalty term for those loans whose expected profits could have been large but are potentially limited by the event of early repayment. For those loans with a lower or negative expected profit, our principal target is still the default risk rather than the prepayment risk, i.e., the profit percentage γ_j does not need to be adjusted on the basis of prepayment risk, thus the penalty term is equal to 0. In Appendix B, we provide another intuitive approach to building the profit percentage, one which takes both default and prepayment into consideration. The resultant penalty term is very similar in structure to Eq. (19), but does not adequately take account of the severity of the loss. Therefore, we use Eq. (19) in our experiments. Please refer to Appendix B for further discussions.

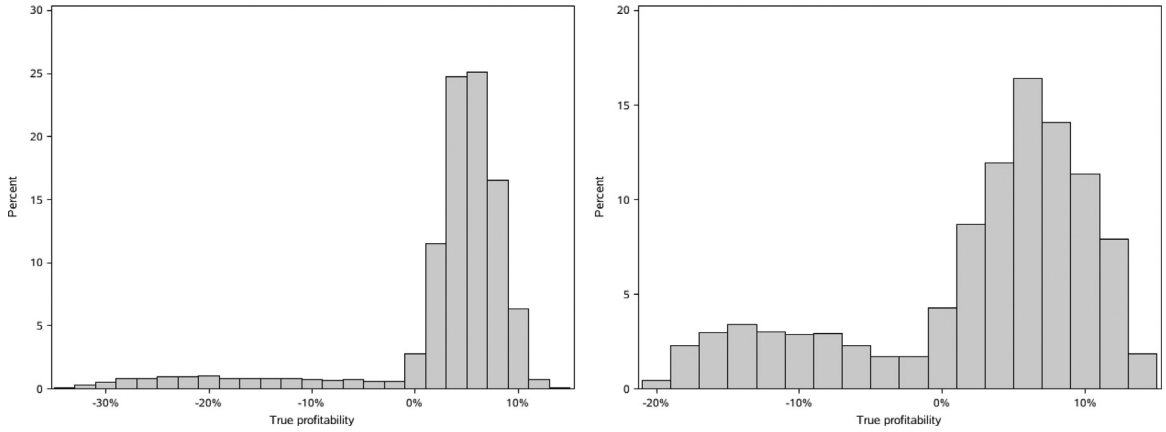
Rather than out-of-sample in Section 5, here we use out-of-time prediction since the true profitability requires the loans to be fully matured. Facing a similar situation as Andreeva et al. (2007), we cannot go beyond the end of the observation period, in our case December 2019. We follow their example, by using earlier accounts opened in the previous year as a holdout out-of-time sample. In total, 40,735 individual 36-month loans and 9489 60-month loans issued in 2012 are taken as the out-of-time sample, for Panels A and B, respectively. Two performance measures, the root mean squared error (RMSE) and the mean absolute error (MAE), are adopted. A lower value of these two measures indicates a superior predictive accuracy.

The true profitability p is the difference between the present value of payments and investments as a yearly average percentage of the total investments, which is also known as a lender's effective interest rate or internal rate of return in the loans market (Serrano-Cinca & Gutierrez-Nieto, 2016). Fig. 9 shows the distributions of the true profitability on the out-of-time sample. It can be seen that the true profitability for Panel A is distributed between −0.34 and 0.14, and the true profitability of around 80% observations is in the $(0, 0.1]$ range, whilst for Panel B, the true profitability is distributed from −0.2 to 0.15 and around 60% of observations are located in the $(0, 0.1]$ range. Please note that the profitability can go infinity if the interest rate increases and its lower bound is $-12/N_j$ if the loan is totally lost.

Profitability, defined as the annualised values of profit percentages is taken for comparison between loans of two terms. For simplicity and consistency, we denote the profitability estimated from the CRPH model with default as $\hat{p}^{D}_{CRPH} = 12\gamma_j/N_j$ and the profitability estimated from the CRPH model with default and prepayment as $\hat{p}^{D+P}_{CRPH} = 12\tilde{\gamma}_j/N_j$, respectively. Profitability forecasts $\hat{p}^{D}_{CRPH}$ and $\hat{p}^{D+P}_{CRPH}$ are calculated on the out-of-time sample with the parameters in the CRPH model which fits the time to events.

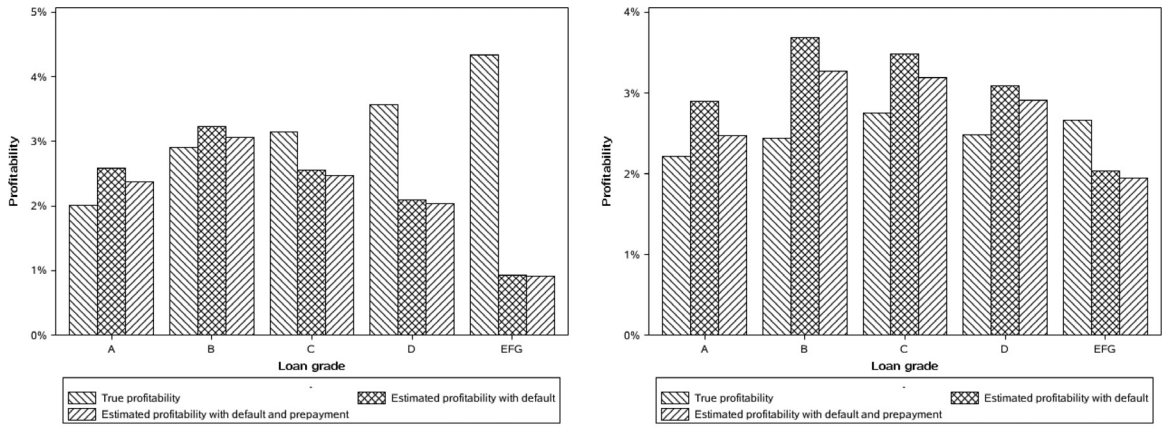
Linear regression is often used in modelling profitability (Fitzpatrick & Mues, 2021; Serrano-Cinca & Gutierrez-Nieto, 2016), and thus the linear profitability forecast $\hat{p}_{OLS}$ is employed as a benchmark for comparison. We understand that if it is used for a bounded dependent variable such as LGD (Nazemi et al., 2017) and the profitability in our case, it is flawed in that it may produce estimates out of the range of the bounded dependent variable. We still use it in analysis because it has been widely applied to profitability prediction (Andreeva et al., 2007; Finlay, 2010) and performs well (Fitzpatrick & Mues, 2021). Linear regression fits the true profitability on the training sample using the variables in Table A1. The predicted $\hat{p}_{OLS}$ is then calculated on the out-of-time sample using estimated coefficients on the training sample. To save space, the results of linear regression fitting on the training sample have not been reported. It should be noted that the adjusted R^2 of linear regression is 0.0297 for 36-month loans and 0.0701 for 60-month loans, which are comparable with literature (Serrano-Cinca & Gutierrez-Nieto, 2016). The forecast $\hat{p}_{PH}$ comes from the traditional PH model, i.e., the PH model without competing risks. The process of estimating $\hat{p}_{PH}$ is similar to those for $\hat{p}^{D}_{CRPH}$, whereas the difference lies in the fact that prepayment serving as a competing risk for default is not considered. The prepaid samples are left out when we fit the PH model and estimate survival probability of default. We report these results in Table 6 and split the whole validation sample into five buckets based on ascending true profitability and the approximately equal number of observations in each bucket.

The first bucket 'P1' includes 20 percent of loans with the lowest profitability, while the fifth bucket 'P5' contains the most profitable 20 percent of loans. In the first bucket, where most loans generated negative profits, the traditional PH model has better ac-



(a) Panel A: 36-month loans (b) Panel B: 60-month loans

Fig. 9. Distributions of true profitability on the out-of-time sample.



(a) Panel A: 36-month loans (b) Panel B: 60-month loans

Fig. 10. The out-of-time profitability forecasts of different loan portfolios.

curacy on profitability forecast than the CRPH model. In the second bucket, linear regression shows competitive power in Panel A and Panel B. For buckets P3 to P5, namely highly profitable loans, $\hat{p}^{D}_{CRPH}$ and $\hat{p}^{D+P}_{CRPH}$ consistently outperform $\hat{p}_{OLS}$ and $\hat{p}_{PH}$ in terms of RMSE and MAE. These results show that the predictive performance of the traditional PH model and linear regression are superior for the lower profit buckets, and the CRPH model gives better results for the higher profit buckets. Overall, across the out-of-time validation sample, the CRPH model show robust and consistent power in estimating the profitability.

We further evaluate the profitability of loan portfolios based on the loan grade. It is the practice that banks, insurance companies and fund managers all have their portfolios built on the grade. We reinvestigate performance from the perspective of such portfolios. Each portfolio is a pool of loans with the same grade and the profitability is estimated on the individual loan level. The out-of-time profitability estimates are shown in Fig. 10. The estimated profitability of each grade portfolio is obtained by taking a weighted average of each loan's profitability forecast, and the weight is defined as the individual loan amount to the size of portfolio. The details of comparison are given in Table 7.

In Fig. 10, it can be seen that for portfolios A and B, profitability tends to be overestimated, while in portfolio EFG, profitability tends to be underestimated. The big gap shown on portfolio EFG for 36-month loans may be caused by the large standard deviation of profitability in these high-risk categories, along with the relatively low sample size. We notice that the average profitability increases along with a lower grade, although lower grades do imply a larger default risk. It makes sense that if we look at the distributions of profitability in each grade (Fig. 8), the variance in lower grades is considerably larger than that observed in higher grades. According to the portfolio theory, investors make trade-offs between the expected return and the risk indicated by the variance. A commonly used measure which assesses the trade-offs is the Sharpe ratio, defined as the difference between a portfolio's average return and its risk-free return, divided by the standard deviation of the portfolio's returns (Kirkby et al., 2020). The larger the Sharpe ratio, the better the performance of the portfolio from the risk/return perspective. It is used to select the optimal portfolio of online loans in Guo et al. (2016). We also report the Sharpe ratios in Table 7. In fact, the relationship between grades and individual loan's profitability is also not linear but complex (Serrano-Cinca & Gutierrez-Nieto, 2016). Thus, a grade cannot indicate a loan's profitability or be used as a reference for loan portfolio selection. The CRPH model can capture the individual loan's profitability well. We find that an accurate point estimation on an individual loan is not easy, particularly when there is a large variance. It is also suggested that a tailed segment model may be necessary for better forecasting. The results in Table 7 not only show the gap between true profitability and estimated profitability on each portfolio, but also quantify the improvement of predictions when prepayment is taken into account. The estimated profitability with default and

Table 7
True and estimated profitability of loan portfolios.

	Loan grade	N	Size of portfolios (m$)	p		$\hat{P}^{D}_{CRPH}$			$\hat{P}^{D+P}_{CRPH}$	
				Mean	St.D	Sharpe ratio	Mean	Diff. with p	Mean	Diff. with p
Panel A	A	10,133	113.760	0.0200	0.0569	0.2821	0.0258	−0.0058	0.0237	−0.0037
	B	15,795	175.354	0.0291	0.0772	0.3252	0.0324	−0.0032	0.0307	−0.0015
	C	9235	106.924	0.0315	0.0956	0.2878	0.0256	0.0059	0.0248	0.0067
	D	4724	65.635	0.0357	0.1077	0.2946	0.0209	0.0148	0.0204	0.0153
	EFG	848	17.465	0.0434	0.1164	0.3386	0.0093	0.0341	0.0091	0.0343
Panel B	A	140	1.824	0.0221	0.0375	0.3527	0.0290	−0.0068	0.0247	−0.0026
	B	1637	33.493	0.0243	0.0616	0.2508	0.0369	−0.0125	0.0327	−0.0083
	C	1901	38.047	0.0275	0.0728	0.2555	0.0348	−0.0073	0.0319	−0.0044
	D	2151	41.875	0.0249	0.0872	0.1831	0.0309	−0.0060	0.0291	−0.0042
	EFG	3660	86.963	0.0267	0.0963	0.1846	0.0203	0.0063	0.0194	0.0073

prepayment is generally closer to the true profitability than the estimated profitability with default only. Though the distributions for each grade between the two terms are different, the CRPH model is preferred for estimation in both cases.

7. Conclusions

In online lending, the credit grades provided by the lending platform mainly capture the risk of default, indicating the borrowers are 'good' or 'bad' in repayment. As early repayment before maturity is common in online lending and brings losses to investors, it is necessary for credit scoring models to capture the risk of both default and prepayment in order to support investor's profit-focused decisions, and estimate the profitability of online loan portfolios.

In this paper, we make several contributions to the larger corpus of credit risk and profitability research. First, the impact of the loan term on the distribution patterns of default and prepayment is shown by the life-table method in survival analysis. Second, we apply a competing risks analysis to the PH model to predict time to default and prepayment, and compare its predictive performance with other state-of-the-art classification methods. The main contribution set out by this paper is that we develop a novel model with the survival estimation of default and prepayment to forecast the profitability of online loans, based on a large volume of real-life lending and repayment data. This will lead to smart and profitable investment strategy for institutional investors such as banks when they make decisions and build up portfolios on the online platform.

In the competing risks analysis, we first examine how the application information provided by the lending company and various other MVs affect the potential hazards of default and prepayment. Then, two measures of the predictive performance are assessed. One is to compare the PH models to the widely-used classification methods, while the other is to dynamically evaluate the performance of the CRPH model over time across each month-on-book. It is found that the CRPH model is competitive with LR, SVM and RF and outperforms the traditional PH model and ANN in identifying those who will default or prepay in the first and second year. What is more, the risk of prepayment cannot be as accurately predicted as the default risk, due to the fact that most application information and bureau data pertains to the default risk rather than the prepayment risk.

We also propose a method to estimate profitability with both default and prepayment risks integrated. The expected profit percentage and its annualised profitability are both analysed in this research. It is found that the loan grade cannot be an indicator of profitability. We present a framework to measure the individual loan's profitability and further apply it to the prediction of an entire portfolio's profitability. In addition, to compensate for the *ex-post* losses arising from early repayment, we suggest the inclusion of prepayment penalties in an *ex-ante* contract to prevent and reduce the losses. This is also the common practice in most mortgage contracts. The risk-adjusted profitability forecast can be realised by competing risks survival analysis as dynamic estimates over the repayment period can be generated. Supplementary information from discounted recoveries could also be added to the model, but it would need LGD estimates, which is another question in the Basel Accord. Further work could be done to redesign the prepayment penalty contract and make it practicable for business.

Declaration of Competing Interest

None.

Acknowledgement

The authors would like to thank for the support from the funding of National Natural Science Foundation in China [No. 71901230] and the Fundamental Research Funds for the Central Universities, China [JBK2103013]. This work was also supported by Financial Innovation Centre (Project No. 2022A0011) at Southwestern University of Finance and Economics, China. We thank the editors and the anonymous reviewers for their comments to improve this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2022.08.013.

References

Agarwal, S., Ambrose, B. W., & Liu, C. L. (2006). Credit lines and credit utilization. *Journal of Money Credit and Banking, 38*(1), 1–22.
Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research, 222*(1), 168–178.
Allison, P. D. (2010). *Survival analysis using SAS: A practical guide* (2nd ed.). Sas Institute.
Anderson, R. (2007). *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*. Oxford University Press.
Andreeva, G., Ansell, J., & Crook, J. (2007). Modelling profitability using survival combination scores. *European Journal of Operational Research, 183*(3), 1537–1549.
Ateca-Amestoy, V., & Prieto-Rodriguez, J. (2013). Forecasting accuracy of behavioural models for participation in the arts. *European Journal of Operational Research, 229*(1), 124–131.
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.
Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance, 34*(10), 2510–2517.
Bauer, J., & Agarwal, V. (2014). Are hazard models superior to traditional bankruptcy prediction approaches? A comprehensive test. *Journal of Banking & Finance, 40*, 432–442.

Bellotti, A., Brigo, D., Gambetti, P., & Vrins, F. (2021). Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting, 37*(1), 428–444.
Bellotti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society, 60*(12), 1699–1707.
Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting, 28*(1), 171–182.
Bellotti, T., & Crook, J. (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society, 65*(3), 340–350.
Beltratti, A., Benetton, M., & Gavazza, A. (2017). The role of prepayment penalties in mortgage loans. *Journal of Banking & Finance, 82*, 165–179.
Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32.
Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446–3453.
Brown, S. R. (2016). The influence of homebuyer education on default and foreclosure risk: A natural experiment. *Journal of Policy Analysis and Management, 35*(1), 145–172.
Byanjankar, A., & Viljanen, M. (2020). Predicting expected profit in ongoing peer–to-peer loans with survival analysis-based profit scoring. *Intelligent decision technologies 2019* (pp. 15–26). Springer.
Cole, R. A., & Gunther, J. W. (1995). Separating the likelihood and timing of bank failure. *Journal of Banking & Finance, 19*(6), 1073–1089.
Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B, 34*(2), 187–202.
Deng, Y. H., & Gabriel, S. (2006). Risk-based pricing and the enhancement of mortgage credit availability among underserved and higher credit-risk populations. *Journal of Money Credit and Banking, 38*(6), 1431–1460.
Deng, Y. H., Quigley, J. M., & Van Order, R. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica : Journal of the Econometric Society, 68*(2), 275–307.
Do, H. X., Rösch, D., & Scheule, H. (2018). Predicting loss severities for residential mortgage loans: A three-step selection approach. *European Journal of Operational Research, 270*(1), 246–259.
Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., et al., (2016). Description-text related soft information in peer-to-peer lending - evidence from two leading European platforms. *Journal of Banking & Finance, 64*, 169–187.
Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *Review of Financial Studies, 25*(8), 2455–2483.
Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research, 297*(3), 1178–1192.
Emekter, R., Tu, Y. B., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics, 47*(1), 54–70.
Ergungor, O. E., & Moulton, S. (2014). Beyond the transaction: banks and mortgage default of low-Income homebuyers. *Journal of Money Credit and Banking, 46*(8), 1721–1752.
Ferris, M. C., & Munson, T. S. (2002). Interior-point methods for massive support vector machines. *SIAM Journal on Optimization, 13*(3), 783–804.
Finlay, S. (2010). Credit scoring for profitability objectives. *European Journal of Operational Research, 202*(2), 528–537.
Fitzpatrick, T., & Mues, C. (2021). How can lenders prosper? Comparing machine learning approaches to identify profitable peer-to-peer loan investments. *European Journal of Operational Research, 294*(2), 711–722.
Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2014). Modelling credit risk with scarce default data: On the suitability of cooperative bootstrapped strategies for small low-default portfolios. *Journal of the Operational Research Society, 65*(3), 416–434.
Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378.
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.
Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *American Statistician, 52*(2), 181–184.
Hsieh, F. Y. (1995). A cautionary note on the analysis of extreme data with Cox regression. *The American Statistician, 49*(2), 226–228.
Hu, R. C., Liu, M., He, P. P., & Ma, Y. (2019). Can investors on P2P lending platforms identify default risk? *International Journal of Electronic Commerce, 23*(1), 63–84.
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.
Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science, 62*(6), 1554–1577.
Jiang, C. Q., Wang, Z., Wang, R. Y., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research, 266*(1–2), 511–529.
Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data* (2nd ed.). John Wiley & Sons.
Kirkby, J. L., Mitra, S., & Nguyen, D. (2020). An analysis of dollar cost averaging and market timing investment strategies. *European Journal of Operational Research, 286*(3), 1168–1186.
Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: A self-learning text* (3rd ed.). Springer.
Kumar, D., & Klefsjö, B. (1994). Proportional hazards model: A review. *Reliability Engineering & System Safety, 44*(2), 177–188.
Lane, W. R., Looney, S. W., & Wansley, J. W. (1986). An application of the cox proportional hazards model to bank failure. *Journal of Banking & Finance, 10*(4), 511–531.
Li, K., Zhou, F., Li, Z., Li, W., & Shen, F. (2021). A semi-parametric ensemble model for profit evaluation and investment decisions in online consumer loans with prepayments. *Applied Soft Computing, 107*, Article 107485.
Li, K., Zhou, F., Li, Z., Yao, X., & Zhang, Y. (2021). Predicting loss given default using post-default information. *Knowledge-Based Systems, 224*, Article 107068.
Li, Z., Li, K., Yao, X., & Wen, Q. (2019). Predicting prepayment and default risks of unsecured consumer loans in online lending. *Emerging Markets Finance and Trade, 55*(1), 118–132.
Lin, M. F., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science, 59*(1), 17–35.
Lin, W. C., Tsai, C. F., Hu, Y. H., & Jhang, J. S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences, 409-410*, 17–26.
Ma, L., Zhao, X., Zhou, Z. L., & Liu, Y. Y. (2018). A new aspect on P2P online lending default prediction using meta-level phone usage data in China. *Decision Support Systems, 111*, 60–71.
Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications, 42*(10), 4621–4631.
Malik, M., & Thomas, L. C. (2010). Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society, 61*(3), 411–420.
Mayer, C., Piskorski, T., & Tchistyi, A. (2013). The inefficiency of refinancing: Why prepayment penalties are good for risky borrowers. *Journal of Financial Economics, 107*(3), 694–714.
Michels, J. (2012). Do unverifiable disclosures matter? Evidence from peer-to-peer lending. *Accounting Review, 87*(4), 1385–1413.
Mild, A., Waitz, M., & Wockl, J. (2015). How low can you go? - Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. *Journal of Business Research, 68*(6), 1291–1305.
Mo, L., & Yae, J. (2022). Lending Club meets Zillow: local housing prices and default risk of peer-to-peer loans. *Applied Economics, 54*(35), 4101–4112.
Nazemi, A., Fatemi Pour, F., Heidenreich, K., & Fabozzi, F. J. (2017). Fuzzy decision fusion approach for loss-given-default modeling. *European Journal of Operational Research, 262*(2), 780–791.
Quercia, R., & Spader, J. (2008). Does homeownership counseling affect the prepayment and default behavior of affordable mortgage borrowers? *Journal of Policy Analysis and Management, 27*(2), 304–325.
Rose, M. J. (2013). Geographic variation in subprime loan features, foreclosures, and prepayments. *Review of Economics and Statistics, 95*(2), 563–590.
Schmeiser, M. D., & Gross, M. B. (2016). The determinants of subprime mortgage performance following a loan modification. *Journal of Real Estate Finance and Economics, 52*(1), 1–27.
Serrano-Cinca, C., & Gutierrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems, 89*, 113–122.
Serrano-Cinca, C., Gutierrez-Nieto, B., & Lopez-Palacios, L. (2015). Determinants of default in P2P lending. *PloS One, 10*(10), Article e0139427.
Steinbuks, J. (2015). Effects of prepayment regulations on termination of subprime mortgages. *Journal of Banking & Finance, 59*, 445–456.
Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research, 50*(2), 277–289.
Stepanova, M., & Thomas, L. C. (2001). PHAB scores: Proportional hazards analysis behavioural scores. *Journal of the Operational Research Society, 52*(9), 1007–1016.
Tan, F., Hou, X. R., Zhang, J., Wei, Z., & Yan, Z. Y. (2019). A deep learning approach to competing risks representation in peer-to-peer lending. *IEEE Transactions on Neural Networks and Learning Systems, 30*(5), 1565–1574.
Thackham, M., & Ma, J. (2022). On maximum likelihood estimation of competing risks using the cause-specific semi-parametric Cox model with time-varying covariates – An application to credit risk. *Journal of the Operational Research Society, 73*(1), 5–14.
Tobback, E., Martens, D., Van Gestel, T., & Baesens, B. (2014). Forecasting Loss Given Default models: Impact of account characteristics and the macroeconomic state. *Journal of the Operational Research Society, 65*(3), 376–392.
Varli, Y., & Yildirim, Y. (2015). Default and prepayment modelling in participating mortgages. *Journal of Banking & Finance, 61*, 81–88.
Wang, C., & Tong, L. (2020). Lender rationality and trade-off behavior: Evidence from Lending Club and Renrendai. *International Review of Economics & Finance, 70*, 55–66.
Wang, Z., Jiang, C. Q., Ding, Y., Lyu, X. Z., & Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications, 27*, 74–82.
Xia, Y., Zhao, J., He, L., Li, Y., & Yang, X. (2021). Forecasting loss given default for peer-to-peer loans via heterogeneous stacking ensemble approach. *International Journal of Forecasting, 37*(4), 1590–1613.
Xu, J. J., & Chau, M. (2018). The impact of lender-borrower communication on peer–to-peer lending outcomes. *Journal of Management Information Systems, 35*(1), 53–85.

Yang, S. (1997). A generalization of the product-limit estimator with an application to censored regression. *The Annals of Statistics, 25*(3), 1088–1108 1021.
Yao, X., Crook, J., & Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research, 263*(2), 679–689.
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.
Zhang, N. L., Yang, Q. Y., Kelleher, A., & Si, W. J. (2019). A new mixture cure model under competing risks to score online consumer loans. *Quantitative Finance, 19*(7), 1243–1253.