Applied Soft Computing 107 (2021) 107485

Contents lists available at ScienceDirect

Applied Soft Computing

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

A semi-parametric ensemble model for profit evaluation and investment decisions in online consumer loans with prepayments

Ke Li [a,b], Fanyin Zhou [a,*], Zhiyong Li [c,b], Wanqing Li [a,d], Feng Shen [c]

[a] Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, 555 Liutai Avenue, Chengdu 611130, China
[b] Collaborative Innovation Center of Financial Security, Southwestern University of Finance and Economics, 555 Liutai Avenue, Chengdu 611130, China
[c] School of Finance and Fintech Innovation Center, Southwestern University of Finance and Economics, 555 Liutai Avenue, Chengdu 611130, China
[d] Department of Statistics, Rutgers University, 501 Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA

ARTICLE INFO

Article history:
Received 12 May 2020
Received in revised form 23 April 2021
Accepted 3 May 2021
Available online xxxx

Keywords:
Online consumer loans
Prepayment
Profit evaluation
Portfolio selection

ABSTRACT

In recent years, online consumer loans have advanced rapidly. Prepayment, which is the practice of repaying the loan before maturity, is increasingly emerging as a feature of such loans. Unlike for traditional loans, the effect of prepayment on online loans has not been well addressed by either industry or academia. In this study, we took Peer-to-Peer (P2P) online lending as an example in order to provide a new perspective for evaluating the profitability of online loans where there are also prepayments present. Firstly, we defined a future-value-based return rate to measure the profits of a loan. Then, we proposed an ensemble model based on a semi-parametric mixture distribution to predict the expected return rate. In addition, we formulated a kernel method by borrowing information from similarly profitable loans to assess the return risk. Finally, we established a method to optimise the portfolio selection while taking prepayments into account.

1. Introduction

When a borrower submits a loan request, the investor usually applies credit scoring models to the loan application and then decides whether or not to issue the loan. As [1] summarised, credit scoring is functional in four scenarios denoted by the acronym 4R, namely Risk, Response, Revenue and Retention. In practice, risk is the main focus, whereas response matters in terms of customer acquisition, which takes place before the application process itself. Customer retention is also important to investors, as customer acquisition is costly. Investors therefore want to keep their customer relationships for as long as possible, in order to cross-sell other services and products for additional profits beyond the single credit line. However, in some circumstances, customers may choose to close their accounts and refinance with other investors (referred to as attrition or churn) due to attractive initial promotions, such as an interest discount, offered by a competitor. The prepayment behaviour of fixed loans and closure of revolving accounts are not situations which favour investors.

In recent years, online lending represented by Peer-to-Peer (P2P) lending has provided fast funding for online users. Borrowers can easily access credit supplied by a group of individuals or institutional investors. In traditional lending, if a borrower decides to terminate the loan earlier than its maturity, he or she has to pay an extra penalty fee, which brings extra income to the investor to fully or partially recover opportunity cost due to the unexpected change of investing plan. However, ending an immature loan in online lending, the borrower usually pays no prepayment penalty. Additionally, making a prepayment with online lending is very convenient, while by contrast the borrower needs to sign an array of documents on-site with traditional lending. Thus, online borrowers are more willing to prepay than those in traditional lending. Based on the data from the largest P2P lending platform LendingClub, the prepaid percentage is as high as 55% for 36-month loans, while the corresponding percentage in traditional lending is much lower as [2] and [3] reported.

While the prepayment behaviour saves partial interests for the borrowers, it cuts the investors' revenue that would otherwise have been generated from a matured loan. Due to the frequent prepayments, investors are exposed to additional uncertainty in returns, aside from default risk, where this uncertainty can cause the expected return of P2P loan portfolios to decrease and thus force the investment plan to change. For individual investors this may have no serious impact on their arrangements as they can easily and immediately find alternatives to the paid-off loans on the platform. However, for institutional investors, monthly repayments not only act as guarantees for future income flows, but also act as sources of the principal and any coupons needed for securitised assets. According to LendingClub's official report, about

* Corresponding author.
E-mail addresses: zfy@swufe.edu.cn (F. Zhou), liz@swufe.edu.cn (Z. Li).

https://doi.org/10.1016/j.asoc.2021.107485
1568-4946/

90% of their loans have been funded by institutional investors, and this proportion is still increasing. For these investors, it is not enough to make investing decisions solely based on a default-risk evaluation. It is necessary to have effective management and pricing of their online portfolios based on the predicted risk from both default and prepayment.

As early as 2009, [4] recommended incorporating prepayment into portfolio pricing. However, most existing studies on online lending still focused on default risk. A few such as [5] discussed the prediction of prepayment probability, while the rest do not take the activities of prepayments into account at all. Motivated by the above, our study aims to fill in the gap by proposing an appropriate profit measure and its evaluation tools for fixed-term online loans with prepayments. We define a new type of future-value-based return rate to evaluate the loan profit. A semi-parametric ensemble model is designed to provide the pointwise prediction of this new return rate. In order to help the investors with appropriate portfolio selection, we also invoke a kernel method to evaluate the mean and variance of the return rate. Data analysis and experiments are performed based on a LendingClub dataset.

Our contribution comprises the following three aspects. Firstly, while the traditional profit measure ignores the effect of prepayment, the proposed profit measure considers the influences of both default and prepayment behaviours in the online loans. The analysis based on real data shows that this new measure indeed evaluates the profit of P2P loans more appropriately. Secondly, the proposed measure exhibits a characteristic of mixture distribution. Hence, we design a semi-parametric ensemble model to properly describe the relationship between the proposed measure and the loan features. Experimental results show that this new model provides higher predictive accuracy than the traditional regression models. Thirdly, the suggested kernel method allows the investors to evaluate the return and risk of a newly-issued loan so they can make investment strategies accordingly. In the experiments, the portfolio selection based on the proposed method presents better performance than the existing strategies.

The remainder of this paper is organised as follows. In Section 2, the relevant literatures are briefly reviewed. In Section 3, we explain the limitation of traditional measure for profit evaluation. Then we propose the new profit measure, the predictive model, and the portfolio selection method. Section 4 presents the data description, experiments and results. Finally, our contributions are summarised in Section 5. Our profit-oriented method is recommended to the investors, so that their investing decisions can be properly managed.

2. Literature review

Scholars have developed many methods for credit scoring and risk evaluation for both traditional lending and online lending. The traditional approaches of credit risk modelling are statistical methodology based on structural financial data. As [6] summarised, a wide range of statistical methods have been applied for credit scoring from different aspects. [7] introduced the application of survival analysis, and many others such as [8] followed this path and proposed more survival models into the evaluation of credit risk. On the other hand, machine learning techniques become more and more popular in credit risk modelling. [9] conducted the modelling based on neural network and illustrated its effectiveness for credit risk evaluation. [10] applied a random-forest-based method to predict the probability of default in P2P lending. [11] considered an approach with feature selection based on linear support vector machine. In order to provide consistent feature selection, [12] proposed a Bootstrap-Lasso method for the classification algorithms and applied in credit scoring. [13] introduced a multi-view learning and an adaptive clustering method to predict the probability of default in P2P lending. To summarise the existing contributions including the above, [14] and [15] separately conducted the systematic literature review about the statistical and machine learning models in credit scoring.

Besides the methodology studies, the use of alternative data has been discussed by the researchers as well. [16] examined the borrower screening based on the soft information and suggested to leverage nonstandard information to enhance the lending efficiency. Both [17] and [18] investigated the effect of borrowers' online friendship on their creditworthiness. In addition to the general cases, researchers also conducted studies in sub-area of credit modelling. For example, [19] proposed a two-stage model using heterogeneous ensemble learning to predict the expected loss. [20] introduced a semi-supervised learning method to provide a solution to the reject reference. [21] investigated the classification between revolvers and transactors in addition to the evaluation of default risk.

In traditional lending, the prepayment behaviours have been investigated in several empirical studies. [2] showed a cumulative 10% prepaid rate in the first 36 months for house mortgages, and [3] reported a 23% prepaid rate for auto loans, both of which are much lower than the 55% prepaid rate in P2P lending. [22] presented an empirical analysis to evaluate the importance of the mortgage ownership upon the prepayment and the homeowner mobility. [23] discussed how a household may decide whether to use limited savings to prepay a mortgage or contribute to a tax-deferred retirement account. [2] proposed a unified model of the competing risks of mortgage termination by prepayment and default. [24] employed a competing risks model to examine the default and prepayment behaviour of commercial loans and concluded that changes in the yield curve have a direct impact on the probability of mortgage termination.

The prepayment penalty provides partial protection for the loan investors and allows them to recover opportunity cost due to the early termination. A few studies paid attention to the relationship between the penalty and the prepayment behaviour of the borrowers in different types of mortgages. [25] analysed the bank's policy of prepayment penalty and due-on-sale clause in conventional home mortgage. [26] established a financing model for participating mortgages to incorporate early termination, defeasance and prepayment penalty. [27] investigated the effect of prepayment penalty restrictions on the performance of endorsed subprime mortgages and concluded that the restrictions should raise prepayment and lower default.

In scenarios where prepayments are allowed, the method to build up portfolios solely based on default risk no longer fully reflect the return and risk. The scholars suggested that a profit-based perspective for evaluating loans and forming portfolios might be more appropriate. As early as [28], it has been commented that, at a relatively early stage in lending, it is better to integrate credit scoring with the earning process. Then, the model can be more meaningful and act as a basis for an investor to make optimal and profitable decisions. [29] also commented that the performance of loan asset portfolios depends on the profits accrued in each loan, rather than the performance of risk-based scoring models. The objective of investors has been shifted from minimising risks to maximising profits in the past few years. Profit-based scoring, which takes both revenue and retention issues into consideration, is a relatively new trend in both industry and academia. [30] showed that the use of the expected maximum profit measure for model selection leads to more profitable credit scoring models. [31] used the relative profit measure defined as the customer lifetime value divided by the outstanding debt, and then [32] introduced the time-to-profit scorecards. [33] calculated the internal rate of return (IRR) to estimate profit as a function of the independent variables

Table 1
Summary of literatures in risk and profit evaluation.

Considering prepayment	Lending type	Evaluation target	Methodology	Exemplary literature
No	Traditional, Online	Risk	Multiple statistical and machine learning methods	[6–15]
No	Traditional, Online	Profit	Multiple statistical and machine learning methods	[28–35]
Yes	Traditional	Risk	Multiple statistical methods	[2,3,22–27]
Yes	Online	Risk	Logistic regression	[5]

in regression. [34] proposed a kernel method to evaluate the mean and variance of loan return rate, so the investors can make portfolio selection accordingly. [35] suggested to use a robust optimisation with feature selection to perform the profit-based credit modelling.

The literatures related to the evaluation of loan risk or profit are partially summarised in Table 1. It is found that most of the literatures did not consider the prepayment behaviours. Some discussed about the prepayment but only focused on the risk evaluation. Very few have considered the effect of prepayment in online lending. Accordingly, our study starts from the perspective of prepayment and defines a new profit measure to involve the prepayment's influence in online lending. The details of our design are discussed in next section.

3. Methodology

3.1. Internal rate of return

Traditionally, the profit measurement is the internal rate of return (IRR), which can be calculated by solving the following equation:

$$P = \sum_{i=1}^{T} \frac{CF_i}{(1 + IRR/12)^i} + \frac{E}{(1 + IRR/12)^T} \tag{1}$$

where T represents the predetermined terms, P is the total amount of an issued loan, CF_i is the amount of repayment on the ith MoB, and E is the eventually collected money after a loan was defaulted, which is positive only if an effective recovery happened, zero otherwise.

In practice, the principal amount and the interest rate are determined when the loan is issued, and the repayments are scheduled consequently. If a borrower decides to pay off the loan earlier than maturity, the amount and the schedule of the repayments will be altered so that the IRR remains the same as the interest rate (the APR). In this way, the IRR drops only if default occurs. It cannot identify the difference between a prepaid loan and a fully-paid loan under the same APR.

In investments with repayments, there is an important factor affecting the overall profit to be taken into account: return by reinvesting the repayments. If the investor plans to make full use of the investing capital only over the predetermined terms (e.g., 36 months), the reinvestment of repayments will last only until the end of 36 months, and these re-investments typically generate profit with a return rate less than the APR of the loan. Therefore, the prepayment reduces the overall profitability of a loan due to its longer reinvestment period.

As mentioned above, the IRR is only evaluated over the actual period of loan instead of the predetermined terms, so it fails to be a good measure of overall profit where there are prepayments. Meanwhile, it is obvious that the traditional credit risk measures (probability of default, loss given default and exposure at default) cannot play an appropriate role due to the negligence of prepayments. Therefore, we will propose another approach to properly evaluate the profit of such online lending in the next section.

3.2. End-of-term rate of return

To make all repayment procedures comparable regardless of their outcomes, which means whether they are paid-off as scheduled, prepaid ahead of time or defaulted, we conjecture a new profit measure for P2P lending. For each repayment collected by the investor, we assume it is reinvested immediately on a risk-free asset until the end of the last MoB. Although this act of reinvestment seems to be hypothetical, it actually represents the minimum return that an investor would obtain if no idle fund from the investor's side is allowed during the predetermined term. Therefore, upon the term date, the total profit received would be made up of two parts. One is the return over the actual duration of the loan with the APR, and the other is the return that would accumulate from reinvesting each repayment into a risk-free asset. Since the risk-free return rate is almost certainly smaller than the interest rate of the loan, prepayment accelerates the repayment procedure and causes the overall profit to be less than expected.

In contrast to IRR, we denote the annual percentage rate of return as the **End-of-term Rate of Return (ERR)** and its formula is given by Eqs. (2) and (3):

$$FV = \sum_{i=1}^{T} \left[CF_i \prod_{j=i+1}^{T} \left(1 + \frac{RF_j}{12}\right) \right] + E \cdot I\,(D = 1) \tag{2}$$

$$ERR = \left(\frac{FV}{P}\right)^{12/T} - 1 \tag{3}$$

where T represents the predetermined terms, P is the principal, FV is the future value of all cash flows at the end of the term, CF_i is the amount of repayment at the ith MoB, RF_j is the risk-free annual percentage rate of return at the jth MoB, $I\,(D = 1)$ is the default indicator, and E is the amount of recovery given default. Unlike the IRR, the ERR represents the minimum annual percentage return rate if money was continuously invested over the predetermined term, regardless of the possible outcomes of the loan. From another perspective, the ERR is a kind of effective rate of return, which has been widely used when the generated earnings are reinvested to produce earnings of their own.

3.3. Return prediction model

A loan's ERR is unknown at the issued time, as the repayments have not yet been made available. Therefore, it is necessary to predict the ERR given the loan's features and borrowers' characteristics. As a continuous variable, the ERR can be predicted with multiple choices of statistical models or machine learning techniques. The classical models usually prefer or indeed require certain distribution assumptions. For example, linear regression relies on the assumption of a Gaussian distribution, which is a unimodal distribution.

To explore the distribution of ERR, we examined the data of 36-month loans issued by LendingClub between January 2012 and December 2013 (more details are described in the experimental section). The ERR histogram as shown in Fig. 1 looks like a unimodal distribution with a long negative tail. However, when the histogram is divided into the negative part and the non-negative part (Fig. 2), two distinct peaks emerge when we 'zoom in'. This bimodal pattern implies that the ERR may follow a mixture of two independent distributions. A classical model based on the unimodal-distribution assumption may not provide

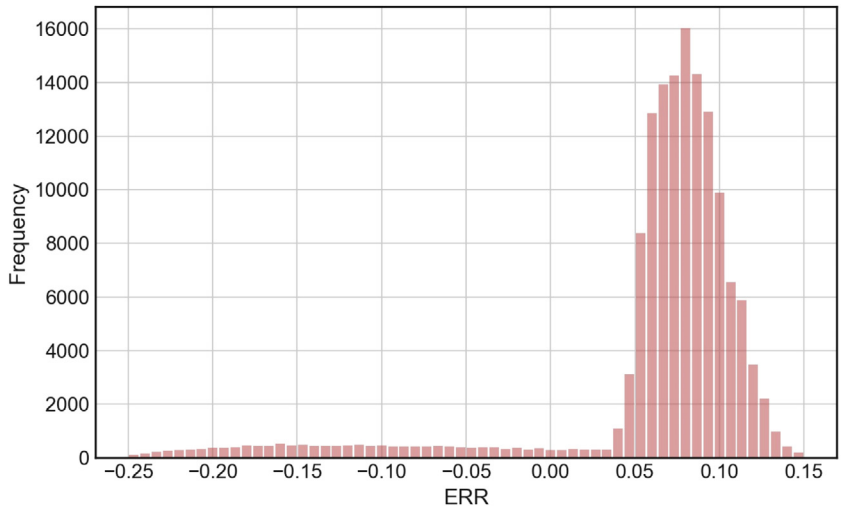

Fig. 1. Histogram for ERR.

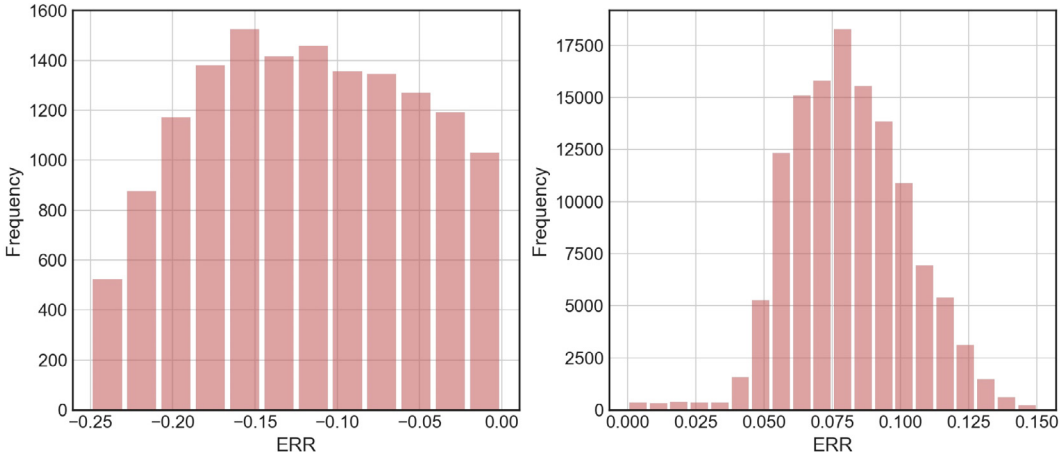

Fig. 2. Histograms for negative ERR and non-negative ERR.

satisfying predictions for a variable with a mixed distribution. Taking the financial interpretation into consideration, we divide the loans into two groups: the profitable (ERR≥0) and nonprofitable (ERR<0). We then propose a model based on the following mixture distribution:

$$f_{R|X}(r|x) = \pi f_{R|Z=0,X}(r|z=1,x) + (1-\pi) f_{R|Z=1,X}(r|z=0,x) \quad (4)$$

where R is ERR, X is the vector of the loan features that can be observed at issue time, Z is an indicator variable equal to 1 if R is non-negative or 0 otherwise, $\pi = P(Z=1|X)$ is the probability of non-negative return, and $f_{R|Z=0,X}$ and $f_{R|Z=1,X}$ are density functions without specified form. This implies that

$$E(R|X) = \pi E(R|Z=1,X) + (1-\pi) E(R|Z=0,X) \quad (5)$$

Thus, we could first estimate $E(R|Z=1,X)$, $E(R|Z=0,X)$ and $\pi = P(Z=1|X)$, and then integrate the results to achieve the estimation of $E(R|X)$ for a given loan. The procedure is given in the following steps and shown graphically in Fig. 3.

Step1: Determine a training dataset $DATA_{train}$. Divide $DATA_{train}$ into $DATA^{+}_{train}$ containing loans with non-negative ERR and $DATA^{-}_{train}$ otherwise.

Step2: On $DATA_{train}$, build a model $\hat{m}$ to estimate $\pi_i = P(Z_i = 1|X_i)$ for each loan i:

$$\hat{\pi}_i = \hat{m}_i(x_i) \quad (6)$$

Simultaneously, build a model $\hat{f}_+$ to estimate $E(R_i|Z_i=1, X_i)$ on $DATA^{+}_{train}$ and build a model $\hat{f}_-$ to estimate $E(R_i|Z_i=0, X_i)$ on $DATA^{-}_{train}$ by

$$\hat{R}^{+}_i = \hat{f}_+(x_i) \quad (7)$$

$$\hat{R}^{-}_i = \hat{f}_-(x_i) \quad (8)$$

Step3: Integrate the results to obtain a final model for prediction

$$\hat{R}_i = \hat{\pi}_i \hat{R}^{+}_i + \left(1-\hat{\pi}_i\right) \hat{R}^{-}_i \quad (9)$$

For a new loan given X_{new}, both the estimation of $E(R_{new}|X_{new})$ and the prediction of R_{new} are given by

$$\hat{R}_{new} = \hat{\pi}_{new} \hat{R}^{+}_{new} + \left(1-\hat{\pi}_{new}\right) \hat{R}^{-}_{new} \quad (10)$$

Note that Eq. (4) assumes the underlying distribution of ERR. π is a parameter, while $f_{R|Z=0,X}$ and $f_{R|Z=1,X}$ are probability density functions with no parametric assumption on their form. Thus, the entire distribution is semi-parametric in the context of statistics. The final prediction (9) actually gives a weighted average of outputs from two nonparametric sub-models, where the estimated parameter $\hat{\pi}$ and $1-\hat{\pi}$ are the weights. From the terminology of machine learning, such a model averaging technique provides an ensemble model. Therefore, this proposed method is called the **semi-parametric ensemble model (SPEM)**.

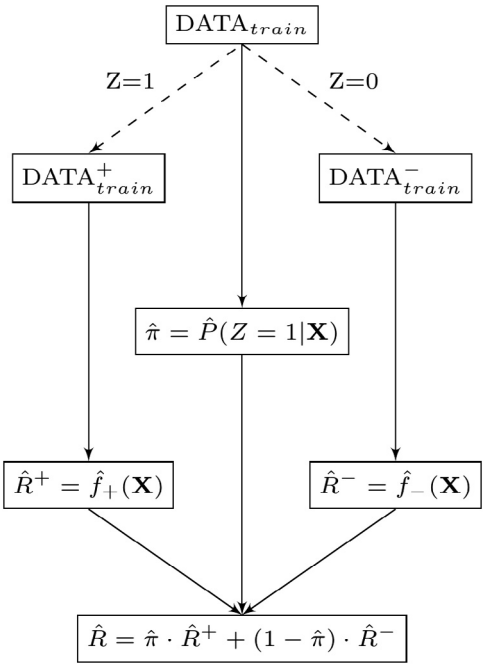

Fig. 3. SPEM structure.

The predicted ERR allows us to re-segment the vast number of loan applications on the P2P lending platform according to its predicted profit, rather than the current credit grades. This might provide a more useful reference for investors. As discussed earlier, the majority of the investors in P2P lending are institutions, whose interests might be more focused on the return of loan portfolios rather than individual performance. Motivated by this, we aim to extend our work further into portfolio selection in the next section.

3.4. Portfolio selection model

P2P lending platforms allow investors to spread their investment over a set of loans. Searching for an optimal investment, the investors will not only distinguish good loans from bad ones, but also try to figure out how to allocate the available capital across different loans.

This scenario falls into a typical portfolio selection framework, referring to the modern portfolio theory [36]. We assume that the investor considers the expected return a desirable thing and the risk (measured by the variance of return) an undesirable thing. One type of portfolio choice is that the investor maximises the expected return with an allowance of risk. The other type of portfolio selection is that the investor minimises the risk given that the expected return at least reaches a desirable level. Without loss of generality, we choose the second type of optimal selection for the remaining of this study.

Since we have taken account of macroeconomic factors in our SPEM model to capture the systematic risk, the correlation between loans is negligible. Therefore, the objective of portfolio selection is simplified to minimise the risk of a portfolio, which is given by the weighted sum of variances,

$$\sum_j \alpha_j^2 Var\left(R_j|X_j\right) \tag{11}$$

subject to

$$\alpha_j \geq 0, \forall j$$

$$\sum_j \alpha_j = 1$$

$$\alpha_j M \leq m_j, \forall j$$

$$\sum_j \alpha_j E\left(R_j|X_j\right) \geq r_0$$

where R_j is the ERR of the jth candidate loan, r_0 is a given level of expected return of the portfolio, M is the total capital the investor will invest into the portfolio, and m_j is the total requested amount of jth loan (available for investing) on the platform.

To solve this optimisation problem, we need to know the expected return rate $E\left(R_j|X_j\right)$ and the risk $Var\left(R_j|X_j\right)$ based on the past performance of loan j. The $E\left(R_j|X_j\right)$ has been estimated by the predicted ERR given by SPEM. However, unlike assets in the stock market, each credit-based loan has only one observable outcome, so $Var\left(R_j|X_j\right)$ cannot be directly estimated by the jth loan's historical outcomes. [34] suggested an instance-based model to estimate the loan risk. They assumed that a loan performs similarly to its 'neighbour' loans, where the return rates of the neighbours were already observed. The return rates of the neighbours can thus be used to 'simulate' or 'approximate' the outcomes of the target loan. Therefore, the risk of a loan can be estimated by a weighted average of the squared difference of return rates.

[34] defined a metric, the absolute difference of the predicted probabilities of default, to evaluate the similarity between two loans and find neighbours based on their similarity. However, we are discussing a scenario where there are prepayments. The predicted probability of default does not provide us with any information about prepayments. Thus, the metric defined by [34] is not an appropriate definition for loan similarity in our scenario. The ERR is designed to evaluate the profit of a loan with prepayments, and can be predicted given the loan's characteristics. Therefore, a natural choice of similarity metric is the absolute difference of the predicted ERRs.

In particular, the risk of the jth loan can be predicted by

$$\hat{\sigma}_j^2 = \frac{\sum_{k \in DATA_{train}} \omega_{jk} \left(R_k - \hat{R}_j\right)^2}{\sum_{k \in DATA_{train}} \omega_{jk}} \tag{12}$$

where

$$\omega_{jk} = K\left(\frac{\left|\hat{R}_j - \hat{R}_k\right|}{h}\right) \tag{13}$$

Note that R_k is the ERR of the kth loan, $\hat{R}_k$ is the corresponding predicted ERR, $K(\cdot)$ is a bounded and symmetric (e.g. Gaussian) kernel function with support $[-1, 1]$, $\left|\hat{R}_j - \hat{R}_k\right|$ which gives a similarity measure between loan j and loan k, and h is a hyper-parameter representing the bandwidth.

The weight ω_{jk} is calculated by the kernel function $K(\cdot)$ involving the bandwidth h. Suggested by [37], the h is obtained by performing leave-one-out cross-validation (LOOCV) to minimise the error function given by

$$CV(h) = \frac{1}{n} \sum_{j \in DATA_{train}} \left(R_j - \hat{\mu}_j\right)^2 \tag{14}$$

where the $\hat{\mu}_j$ is given by

$$\hat{\mu}_j = \frac{\sum_{k \in DATA_{LOOCV-train-j}} \omega_{jk} R_k}{\sum_{k \in DATA_{LOOCV-train-j}} \omega_{jk}}. \tag{15}$$

Once the expected return and risk of each loan has been estimated, the portfolio selection can be performed accordingly.

Table 2
Description of main variables.

Symbol	Variables	Definitions
	Borrower characteristics	
X_1	Oldest revolving account	Months since oldest revolving account opened.
X_2	Oldest bank instalment account	Months since oldest bank instalment account opened.
X_3	Recent account	Months since most recent account opened.
X_4	Recent revolving account	Months since most recent revolving account opened.
X_5	Trades within 24 months	Number of trades opened in past 24 months.
X_6	Mortgage	Number of mortgage accounts.
X_7	Enquiries within 6 months	Number of enquiries in past 6 months.
X_8	Revolving trades	Number of currently active revolving trades.
X_9	Positive revolving trades	Number of revolving trades with balance > 0.
X_{10}	Revolving account	Number of revolving accounts.
X_{11}	Credit lines	The total number of credit lines currently in the borrower's credit file.
X_{12}	Public record	Number of derogatory public records.
X_{13}	Average current balance	Average current balance of all accounts.
X_{14}	Revolving credit balance	Total credit revolving balance.
X_{15}	Debt to income	A ratio calculated using the borrower total monthly debt payments on the total debt obligations.
X_{16}	Monthly income	Monthly income of the borrower.
X_{17}	Earliest credit line	The month the borrower's earliest reported credit line was opened.
X_{18}	Employment length	Employment length in years.
X_{19}	FICO range (low)	The lower boundary the borrower's FICO at loan origination belongs to.
X_{20}	Home ownership	The home ownership status
X_{21}	Revolving line utilisation	Amount of credit the borrower is using relative to all available revolving credit.
X_{22}	Open to buy	Total open to buy on revolving bankcards.
X_{23}	Residency	The state of residence.
	Loan information	
X_{24}	Amount	The total amount of the loan.
X_{25}	Interest rate	Interest rate on the loan.
X_{26}	Purpose	Purpose of the loan request.
X_{27}	Subgrade	LendingClub assigned loan subgrade.
	Macroeconomic factors	
X_{28}	Unemployment rate	Seasonal unemployment rate of the US.
X_{29}	Inflation	Seasonal inflation rate of the US.
X_{30}	SP500	Monthly average close price of SP500.
X_{31}	CPI	Seasonal consumer price index of the US.
X_{32}	GDP growth rate	Seasonal GDP growth rate of the US.

4. Experiments and results

The data processing and consequent experiments were performed on a PC with Intel Core i5 and 8 GB RAM. The programming environment is Python 3.7.6 on Windows 10, with installed packages of NumPy, Pandas, Sklearn, Matplotlib, Random, XGBoost, Category_Encoders, CvxOpt, etc. The code and example data are available at the website doi.org/10.5281/zenodo.4705439.

4.1. Data description

P2P lending is a typical format of online consumer loans and provides an access point to big data for research purposes. In this paper, LendingClub in the United States, the world's largest online credit marketplace offering P2P lending, is taken as an example. We investigated a dataset publicly available on LendingClub's website. The data consists of the 36-month loans issued between January 2012 and December 2013 with final status (fully paid, prepaid or defaulted) observed by January 2017. Excluding those with obvious errors, this subset consists of 142,992 issued loans, in which there are 19,348 defaults, 78,022 prepays, and 45,622 fully paid. In this dataset, the characteristics of borrowers and the loan information, including the repayments and the recovery, are recorded. With the addition of a few macroeconomic factors, the main variables are described in Table 2.

Fig. 4 shows the accumulated proportion of prepaid loans and defaulted loans on different months-on-book (MoBs). About 55% of the loans were prepaid, whereas only 13.5% of the loans were defaulted by the end of 36 months. Meanwhile, Table 3 shows the cross-tabulation between the loan status and their official LendingClub grades (from A to E&E-, corresponding to from "low default risk" to "high default risk"). It is observed that the default rate increases rapidly as the loan grade deteriorates. On the contrary, the prepayment rate shows an opposite pattern, that the loans with better grades are more likely to be prepaid. The proportion of fully-paid loans is consistently less than that of the early-settled, across all grades.

Table 3
Cross table between loan status and grade.

Grade	A	B	C	D	E&E-	Total
Fully paid	36.3%	33.1%	29.4%	27.8%	26.6%	45,622 (31.9%)
Prepaid	57.7%	55.6%	53.6%	50.4%	48.7%	78,022 (54.6%)
Defaulted	6.0%	11.3%	17.1%	21.9%	24.8%	19,348 (13.5%)
Total	27,687	56,814	34,329	19,454	4708	142,992 (100%)

Table 4 shows the median statistics for IRR and ERR of the loans within each grade. For the fully-paid loans and the prepaid loans, both the median IRR and the median ERR increase as the APR increases. The IRR remains the same for either fully-paid loans or prepaid, but the ERR shows us that the prepaid loans have a slightly lower return rate than the fully-paid ones. For the defaulted loans, both IRR and ERR decrease as the APR increases (except for Grade A). The absolute values of IRR and ERR are quite different for the defaulted loans, because the IRR is evaluated at times of repayment, but the ERR considers the risk-free reinvesting profit up until the end of the 36th MoB. All of the above show us that ERR is behaving similarly to IRR to capture the profit level and risk level of the loans, whereas the ERR can also distinguish the fully-paid and the prepaid.

Additionally, Fig. 5 and Table 5 indicate that the IRR is equal to the APR for a fully-paid loan. They also show that the IRR remains entirely the same as the APR, whenever a loan is prepaid. Thus,

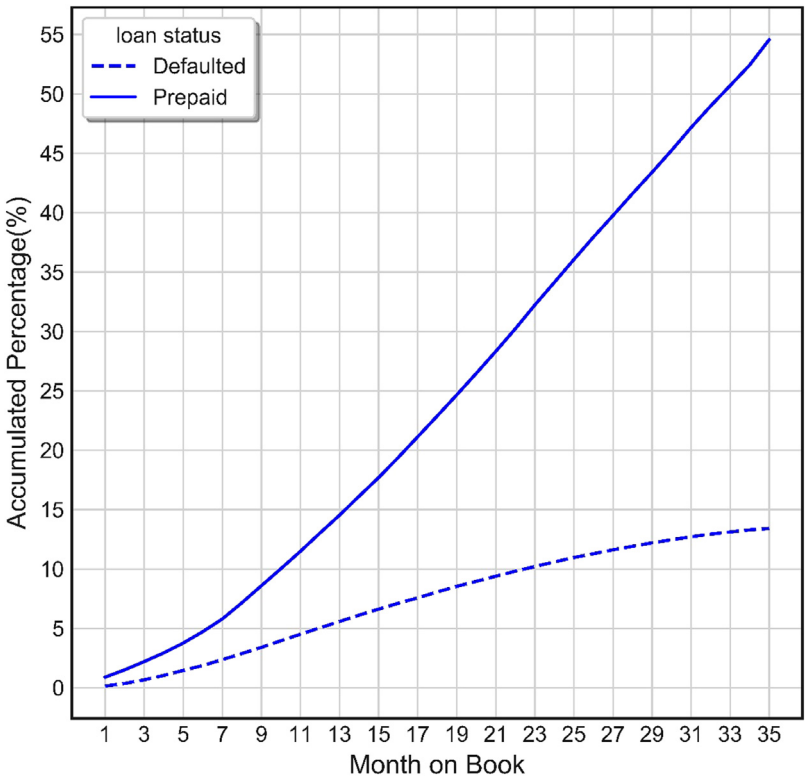

Fig. 4. The accumulated percentage of prepayments and defaults.

the IRR cannot distinguish the loans that were prepaid at different MoBs. Furthermore, all defaulted loans have a lower IRR than any prepaid one (see the upper three graphs of Fig. 5). It means that a defaulted loan is always judged to be 'worse' than the prepaid, even if it may provide higher profits, in rare but existing cases. In contrast, for a prepaid loan, a higher ERR is generated if the prepayment happens later. This illustrates that the ERR indeed distinguishes early-prepaid and late-prepaid loans. Moreover, a late-defaulted loan could still be profitable and have a higher ERR than an early-prepaid loan with the same APR. A few examples are highlighted in Table 5 and Fig. 5. With an APR of 5%, the ERR of a loan defaulted at the 36th MoB exceeds the ERR of any loan prepaid earlier than the 8th MoB. More obvious phenomena are seen for the loans defaulted at the 36th MoB with an APR of either 10% or 15%. All of the above show that the ERR is a profit measure with finer granularity than the traditional IRR.

Fig. 1 shows the ERR distribution across all loans, where the return is centred around 8% with a range spread from −26% to 16%. The negative part is contributed by the defaulted loans. However, the high interest rate plus possible recovery from collection may generate substantial profit for the defaulted loans. Therefore, 24% of the defaulted loans have a non-negative ERR, although they are still labelled as 'bad' loans. This further suggests that, to evaluate the profit, it would be preferable to look at the ERR, rather than resorting to traditional measurements.

Without loss of generality, we removed the loans with missing values and kept 114,000 loans for the experiments. The dataset is randomly divided into five training sets (from $DATA_{train\_1}$ to $DATA_{train\_5}$, each containing 20,000 loans) and five test sets (from $DATA_{test\_1}$ to $DATA_{test\_5}$, each containing 2800 loans). Then we divide each $DATA_{train\_i}$ into $DATA^{+}_{train\_i}$ with non-negative-ERR loans and $DATA^{-}_{train\_i}$ otherwise. In a real scenario, a portfolio selection may not be performed on a large set of 2800 candidate loans. Therefore, in the portfolio selection stage, we randomly divide each $DATA_{test\_i}$ into 10 subsets from $DATA^{(1)}_{test\_i}$ to $DATA^{(10)}_{test\_i}$, with 280 candidate loans in each subset.

4.2. Return prediction

To implement the SPEMs and the traditional models, we need the following steps:

(a) On each $DATA_{train\_i}$, we establish a model to predict $\pi = P(Z = 1|X)$. Among various models, logistic regression provides an explainable probabilistic result. Hence the fact that logistic regression with LASSO has been chosen here, to predict the probability of non-negative ERRs. The hyper-parameter of the LASSO is selected by cross-validation to maximise the area under the receiver operating characteristic curve (AUC). On each $DATA_{test\_i}$, apply the established model to predict $\hat{\pi}_j$ for each loan. The average predicted probabilities are shown in Table 6. It is found that the average predicted probability for non-negative part is higher than the one for negative part, which is consistent with the observed proportion.

(b) Separate regression models are established, based on $DATA^{+}_{train\_i}$ and $DATA^{-}_{train\_i}$, to predict the ERR. There are multiple choices of models for such a prediction task, from among which we considered three popular candidates: linear regression with LASSO (a classical statistical model), random forest (a classical bagging decision tree) and XGBoost (a classical boosting decision tree). The mainly involved hyper-parameters (shown in Table 7) were chosen by cross validation to minimise mean absolute error. On each $DATA_{test\_i}$, apply the established regression model to obtain $\hat{R}^{+}_j$ and $\hat{R}^{-}_j$ for every loan j.

(c) Following Eq. (9), we combine the results from both (a) and (b) to obtain the ERR prediction $\hat{R}_j$ for each loan j.

(d) Three traditional regression models (LASSO, XGBoost and random forest) are also established on each $DATA_{train\_i}$ to predict ERR without considering the sign of the observed ERR. These models are called the 'simple' models.

To compare SPEMs with the simple models, Tables 8 and 9 have been established. From the aspects of Mean AE, Median AE, Mean SE and Median SE, Table 8 shows the testing results on

Table 4
The descriptive statistics of IRR and ERR for different loan status.

LendingClub grade APR		A 6.0%–9.3%	B 6.0%–14.1%	C 6.0%–17.3%	D 6.0%–20.3%	E&E- 6.0%–25.9%	Total
Fully-paid	Proportion	36.3%	33.1%	29.4%	27.8%	26.6%	45,622 (31.9%)
	Median IRR	7.9%	12.1%	15.3%	18.5%	21.5%	
	Median ERR	**6.3%**	**8.4%**	**10.1%**	**11.8%**	**13.5%**	
Prepaid	Proportion	57.7%	55.6%	53.5%	50.4%	48.6%	78,022 (54.6%)
	Median IRR	7.9%	12.1%	15.3%	18.5%	21.5%	
	Median ERR	**5.9%**	**7.6%**	**9.0%**	**10.4%**	**11.7%**	
Defaulted	Proportion	6.0%	11.3%	17.1%	21.9%	24.8%	19,348 (13.5%)
	Median IRR	−41.9%	−41.0%	−42.3%	−48.9%	−57.5%	
	Median ERR	**−8.7%**	**−8.3%**	**−8.3%**	**−8.8%**	**−9.5%**	
Total	Proportion	100.0%	100.0%	100.0%	100.0%	100.0%	142,992 (100%)
	Median IRR	7.8%	12.0%	15.2%	18.4%	21.1%	
	Median ERR	**6.0%**	**7.7%**	**9.4%**	**10.8%**	**12.0%**	

Table 5
Cross table between the return rates and the loan status.

APR	Type of return rate	Loan status	MoB 1	MoB 2	MoB 3	MoB 4	…	MoB 33	MoB 34	MoB 35	MoB 36
5%	IRR	Prepay	5.0%	5.0%	5.0%	5.0%		5.0%	5.0%	5.0%	5.0%
		Default	−∞	−1164.0%	−973.5%	−775.3%		−3.0%	−0.8%	1.3%	3.2%
	ERR	Prepay	**2.1%**	**2.2%**	**2.3%**	**2.3%**		3.6%	3.6%	3.6%	3.6%
		Default	−26.0%	−25.3%	−24.6%	−24.0%		−0.3%	0.7%	1.7%	**2.6%**
10%	IRR	Prepay	10.0%	10.0%	10.0%	10.0%		10.0%	10.0%	10.0%	10.0%
		Default	−∞	−1161.3%	−964.2%	−763.0%		2.3%	4.5%	6.5%	8.3%
	ERR	Prepay	**2.2%**	**2.5%**	**2.7%**	**2.9%**		6.1%	6.1%	6.1%	6.2%
		Default	−26.0%	−25.3%	−24.5%	−23.8%		2.2%	3.2%	4.2%	**5.2%**
15%	IRR	Prepay	15.0%	15.0%	15.0%	15.0%		15.0%	15.0%	15.0%	15.0%
		Default	−∞	−1158.4%	−954.8%	−750.6%		7.7%	9.7%	11.6%	13.4%
	ERR	Prepay	**2.4%**	**2.7%**	**3.1%**	**3.4%**		8.7%	8.7%	8.7%	8.7%
		Default	−26.0%	−25.2%	−24.4%	−23.6		4.6%	5.7%	6.7%	**7.7%**

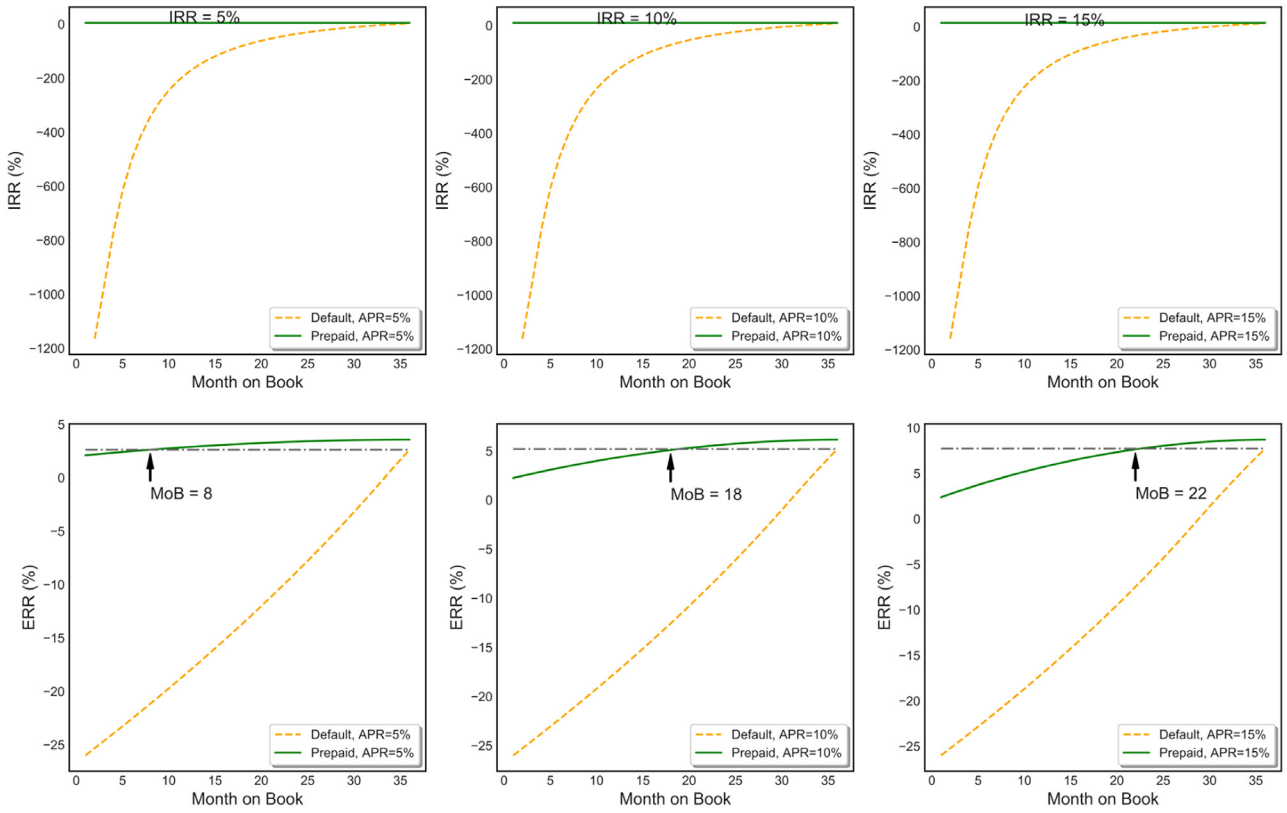

Fig. 5. IRR and ERR for prepay and default with different APRs.

the five independent test sets and the mean result across all five test sets. These results show that the SPEM has higher predictive accuracy than the simple model. For example, the mean MAE of a simple XGBoost across five test sets is 0.03768, and the mean

Table 6
The average predicted probabilities.

	Non-negative ERR	Negative ERR
$DATA_{train\_1}$	0.9028	0.0972
$DATA_{train\_2}$	0.9015	0.0985
$DATA_{train\_3}$	0.9019	0.0981
$DATA_{train\_4}$	0.9035	0.0965
$DATA_{train\_5}$	0.9016	0.0984

MAE of SPEM XGBoost is 0.03672, which **decreases 2.5%** from that of the simple model. The mean MAE of the simple random forest is 0.04277, and the mean MAE of the SPEM random forest is 0.03757, which **decreases 12%** from that of the simple model.

To examine whether the prediction accuracy is sensitive to the size of training set, we performed a sensitivity analysis with randomly selected 10,000 loans and 5000 loans, respectively. Table 9 shows that, on training sets with different sizes (i.e., 20,000, 10,000 or 5000), the prediction errors of SPEM are consistently smaller than that of the simple model.

Although Table 8 shows that the prediction performance varies for different applied regression models, XGBoost has better average performance, across all test sets, than other regression models. Additionally, XGBoost has been widely proven to be an efficient algorithm for regression task. Therefore, we will use the predicted ERRs given by XGBoost SPEM and the XGBoost simple model for further steps.

4.3. Portfolio selection

We firstly implement the kernel methods as follows: (a) Choose a Gaussian kernel for K in Eq. (13) in order to make the results comparable with those by the existing Guo's model [34] (named GM for short). (b) The bandwidth h can be chosen by leave-one-out-cross-validations on each $DATA_{train\_i}$. Table 10 records the h choices for the SPEM, the GM and the simple model. Fig. 6 shows an example of the grid search for h on $DATA_{train\_1}$. (c) For each loan in the 50 testing sets (from $DATA^{(1)}_{test\_1}$ to $DATA^{(10)}_{test\_5}$), the risk is predicted via Eqs. (12) and (13).

Then, for a given expected ERR and the total investment amount, the portfolio optimisation follows the Markowitz theory and is solved by quadratic programming. Multiple pairs of expected return and total amount are considered for a sensitivity analysis. Without loss of generality, we assume that the risk-free rate is 0.04. Equations Eqs. (12) and (13) do not need to know the true ERR of the candidate loans, but only need the true ERR of the historical loans and the estimated ERR of the candidate loans, so the results of portfolio selection are all out-of-sample.

Table 11 shows the out-of-sample Sharpe ratios of the optimal portfolios, which are calculated based on the true ERRs of candidate loans and the portfolios selected by the SPEM, the simple model and the GM. Most of the Sharpe ratios exceed 3.00, showing that all three methods are effective for portfolio selection. When the expected ERR is equal to 0.06, the SPEM's Sharpe ratio is slightly lower than that given by the simple model or by GM.

Table 7
The setting of main hyper-parameters in SPEM.

Dataset	Target	XGBoost				Random Forest			
		Number of estimators	Max depth	Min child weight	Learning rate	Number of estimators	Max depth	Min samples split	Min samples leaf
$DATA_{train\_1}$	$\hat{R}^+$	55	2	4	0.07	50	19	190	10
	$\hat{R}^-$	80	4	2	0.07	250	13	70	10
$DATA_{train\_2}$	$\hat{R}^+$	35	6	4	0.2	250	15	90	30
	$\hat{R}^-$	50	7	2	0.2	120	11	150	20
$DATA_{train\_3}$	$\hat{R}^+$	50	6	4	0.1	160	17	70	40
	$\hat{R}^-$	80	9	1	0.2	50	13	150	10
$DATA_{train\_4}$	$\hat{R}^+$	80	8	4	0.2	250	19	110	10
	$\hat{R}^-$	75	10	3	0.1	50	7	70	20
$DATA_{train\_5}$	$\hat{R}^+$	75	2	1	0.07	120	13	170	20
	$\hat{R}^-$	65	12	1	0.2	120	11	170	40

Table 8
MAE, Median AE, MSE and Median SE on five testing sets.

Mean absolute error	Simple			SPEM			Median absolute error	Simple			SPEM		
	LASSO	XGBoost	Random Forest	LASSO	XGBoost	Random Forest		LASSO	XGBoost	Random Forest	LASSO	XGBoost	Random Forest
Set 1	0.03850	0.03794	0.04253	0.03779	**0.03454**	0.03773	Set 1	0.02139	0.02097	0.02370	0.02040	**0.01514**	0.02026
Set 2	0.03885	**0.03809**	0.04324	0.03810	0.03837	0.03811	Set 2	0.02159	0.02103	0.02356	0.02058	0.02083	**0.02051**
Set 3	0.03771	0.03717	0.04145	0.03727	**0.03653**	0.03721	Set 3	0.02232	0.02156	0.02332	0.02166	**0.02017**	0.02158
Set 4	0.03822	0.03760	0.04335	0.03737	0.03758	**0.03734**	Set 4	0.02165	0.02133	0.02399	0.02028	0.02040	**0.02006**
Set 5	0.03832	0.03758	0.04328	0.03748	**0.03659**	0.03747	Set 5	0.02237	0.02113	0.02428	0.02090	**0.01970**	0.02089
Mean	0.03832	0.03768	0.04277	0.03760	**0.03672**	0.03757	Mean	0.02186	0.02120	0.02377	0.02077	**0.01925**	0.02066
Std.Dev.	0.00037	0.00032	0.00072	0.00030	0.00129	0.00032	Std.Dev.	0.00040	0.00022	0.00034	0.00050	0.00208	0.00054
Mean squared error	Simple			SPEM			Median squared error	Simple			SPEM		
	LASSO	XGBoost	Random Forest	LASSO	XGBoost	Random Forest		LASSO	XGBoost	Random Forest	LASSO	XGBoost	Random Forest
Set 1	0.00440	**0.00434**	0.00493	**0.00434**	0.00442	**0.00434**	Set 1	0.00046	0.00044	0.00056	0.00042	**0.00023**	0.00041
Set 2	0.00455	**0.00444**	0.00517	0.00449	0.00450	0.00449	Set 2	0.00047	0.00044	0.00056	**0.00042**	0.00043	**0.00042**
Set 3	0.00413	0.00408	0.00466	**0.00406**	0.00408	**0.00406**	Set 3	0.00050	0.00046	0.00054	0.00047	**0.00041**	0.00047
Set 4	0.00437	0.00430	0.00503	**0.00425**	0.00428	**0.00425**	Set 4	0.00047	0.00046	0.00058	0.00041	0.00042	**0.00040**
Set 5	0.00417	0.00408	0.00492	**0.00407**	0.00408	0.00409	Set 5	0.00050	0.00045	0.00059	0.00044	**0.00039**	0.00044
Mean	0.00432	0.00425	0.00494	**0.00424**	0.00427	0.00425	Mean	0.00048	0.00045	0.00057	0.00043	**0.00037**	0.00043
Std.Dev.	0.00016	0.00014	0.00017	0.00016	0.00017	0.00016	Std.Dev.	0.00002	0.00001	0.00002	0.00002	0.00007	0.00002

Table 9
Sensitivity analysis for the comparison between SPEM and simple model.

Size of training data	Model type	Regressor	Mean absolute error	Median absolute error	Mean squared error	Median squared error
20,000	Simple	LASSO	0.03832	0.02186	0.00432	0.00048
		XGBoost	0.03768	0.02120	0.00425	0.00045
		Random Forest	0.04277	0.02377	0.00494	0.00057
	SPEM	LASSO	0.03760	0.02077	**0.00424**	0.00043
		XGBoost	**0.03672**	**0.01925**	0.00427	**0.00037**
		Random Forest	0.03757	0.02066	0.00425	0.00043
10,000	Simple	LASSO	0.03850	0.02223	0.00432	0.00049
		XGBoost	0.03805	0.02170	0.00427	0.00047
		Random Forest	0.04277	0.02402	0.00492	0.00058
	SPEM	LASSO	0.03791	0.02133	**0.00426**	0.00046
		XGBoost	**0.03763**	**0.02093**	0.00427	**0.00044**
		Random Forest	0.03784	0.02120	**0.00426**	0.00045
5000	Simple	LASSO	0.03828	0.02183	0.00432	0.00048
		XGBoost	0.03811	0.02121	0.00432	0.00045
		Random Forest	0.04242	0.02318	0.00495	0.00054
	SPEM	LASSO	0.03768	0.02076	**0.00427**	0.00043
		XGBoost	**0.03763**	**0.02040**	0.00431	**0.00042**
		Random Forest	0.03769	0.02075	0.00428	0.00043

Table 10
Choices of bandwidth h for each model.

	SPEM	Simple	GM
$DATA_{train\_1}$	0.00330	0.00228	0.01635
$DATA_{train\_2}$	0.00308	0.00210	0.02194
$DATA_{train\_3}$	0.00321	0.00228	0.01528
$DATA_{train\_4}$	0.00323	0.00245	0.02612
$DATA_{train\_5}$	0.00390	0.00298	0.02112

However, **as the expected ERR increases, SPEM becomes significantly more efficient** compared to the other two methods. This implies that, if an investor is conservative on expected return, the simple model suffices. However, if an investor is aggressive to obtain a higher expected return, the SPEM is more helpful in selecting a portfolio in order to improve the Sharpe ratio.

Additionally, Fig. 7 illustrates the efficiency frontier of three models on the first testing set $DATA^{(1)}_{test\_1}$, while similar behaviours are observed on the rest of the forty-nine testing sets. It confirms the findings from Table 11. All of the above results show us that: (1) Compared to the simple model with an assumption of unimodal distribution, the portfolios optimised based on SPEM yield a higher return for the same amount of risk. (2) Comparing to the GM based on a similarity measure defined by PD, the SPEM based on the ERR-similarity is also more efficient.

5. Concluding remarks

In this study, we have taken peer-to-peer online lending as an example of online consumer lending in order to discuss profit evaluation and investment decision-making with prepayments.

Firstly, we propose a novel profit measure (i.e., the ERR) for the online loans. Unlike the traditional IRR, the ERR provides a profit evaluation subject to prepayment events. Based on a LendingClub dataset, we show that ERR can appropriately reflect the influence of prepayments and evaluate the profit of a loan, while the IRR totally ignores the prepayments. Secondly, with the discovery of the two-modal distribution of ERR, we figure out that the traditional regression models may fail to catch the characteristics of the ERR to make proper predictions. Therefore, we propose a SPEM model based on an assumption of mixture distribution. The empirical analysis shows that the SPEM truly have an improved precision for ERR prediction in contrast to the traditional regression models. Finally, we perform the portfolio selection based on the SPEM-predicted ERR. We compare the portfolio outcome with those given by either the traditional-model-predicted ERR or the predicted IRR. The results have illustrated that the portfolio optimised by the SPEM-predicted ERR exhibits better performance.

By charging transaction fees such as investor fees, borrower fees, and collection fees, the P2P platform has a different desire to make profits from the investors. The early-termination itself does not favour the platform since there is no prepayment fee. The platform claims that the investors gain the same return rate (IRR) from a prepaid loan. Then the investors are encouraged to re-invest other loans on the platform, so new fees can be charged by the platform. From the investors' perspective, the ERR provides a more objective profit measure while the prepayment is allowed. The investors may avoid to overestimate the loan return by the use of ERR and corresponding models with the consideration of prepayment uncertainty.

Although the ERR is shown to be an appropriate profit measure for loans with prepayments, it only represents the minimum return an investor would obtain if no idle fund from the investor's side is allowed during the predetermined term. Thus, the investment strategies based on ERR are conservative. On the other hand, since the SPEM requires to complete one classification task and

Table 11
Comparison of Sharpe ratios by three models.

Expected ERR		0.06			0.065			0.07			0.075			0.08		
Model choice		SPEM	Simple	GM	SPEM	Simple	GM	SPEM	Simple	GM	SPEM	Simple	GM	SPEM	Simple	GM
Invested amount	50,000	4.25	**4.50**	4.39	**4.77**	4.72	4.76	4.40	**4.51**	4.34	**4.01**	3.78	3.95	**3.53**	2.92	3.51
	70,000	4.25	**4.50**	4.39	**4.77**	4.72	4.76	4.40	**4.59**	4.33	**3.99**	3.78	3.93	**3.48**	2.74	3.42
	90,000	4.25	**4.50**	4.39	**4.77**	4.71	4.76	4.40	**4.66**	4.33	**3.97**	3.77	3.90	**3.48**	2.67	3.42

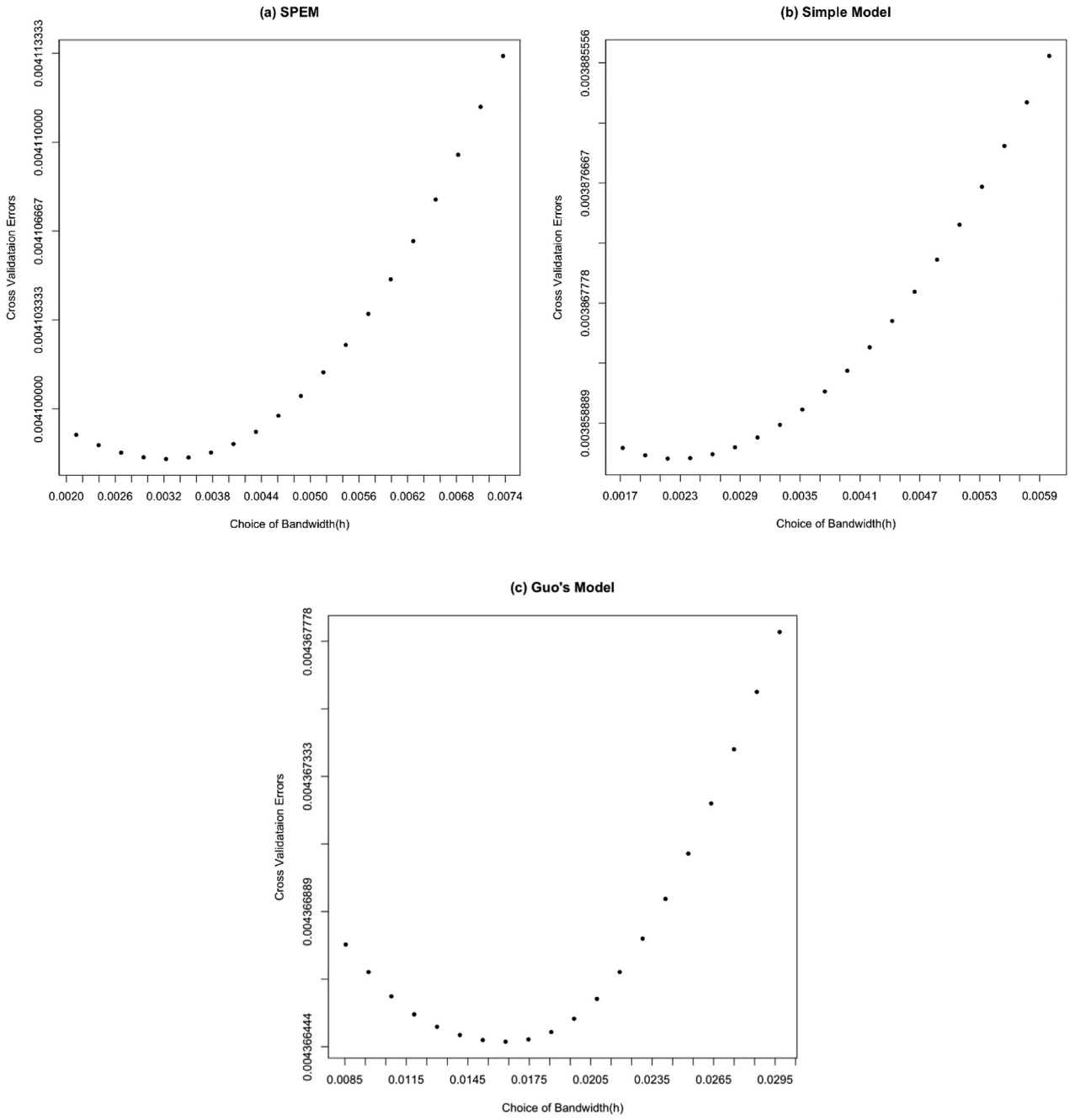

Fig. 6. Identifying the optimal bandwidth on $DATA_{train\_1}$ for three models.

two regression tasks, so it costs more computational resources than the traditional regression models. Therefore, whether or not use ERR should depends on the investor's risk preference and resource availability.

CRediT authorship contribution statement

Ke Li: Conceptualization, Methodology, Writing - original draft. **Fanyin Zhou:** Methodology, Formal analysis, Validation, Writing - review & editing. **Zhiyong Li:** Investigation, Resources, Data curation, Writing - original draft. **Wanqing Li:** Software. **Feng Shen:** Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities in China [JBK2003002], [JBK2103006], [JBK1806002], the National Natural Science Foundation of China [No. 72001178] and the Joint Lab of Data Science and Business Intelligence at Southwestern University of Finance and Economics in China.

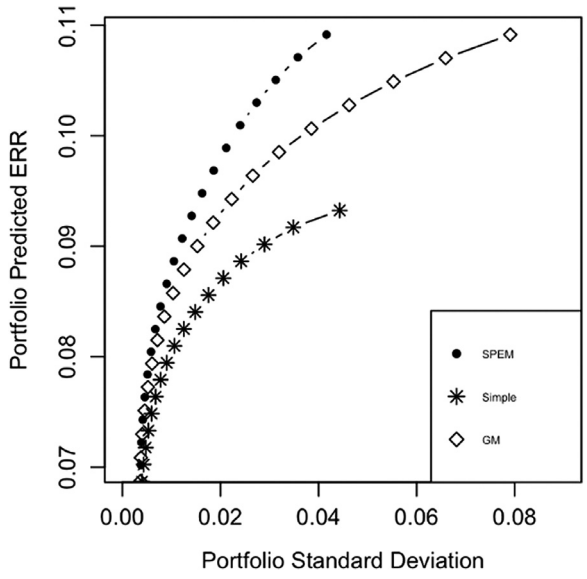

Fig. 7. The efficiency frontier of three models on $DATA^{(1)}_{test\_1}$.

References

[1] R. Anderson, The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation, Oxford University Press, New York, 2007.

[2] Y. Deng, J. Quigley, R. Order, Mortgage terminations, heterogeneity and the exercise of mortgage options, Econometrica 68 (2000) 275–307.

[3] S. Agarwal, B. Ambrose, S. Chomsisengphet, Determinants of automobile loan default and prepayment, Econom. Perspect. 32 (3) (2008) 2008.

[4] L. Thomas, Consumer Credit Models: Pricing, Profit, and Portfolios, Oxford University Press, New York, 2009.

[5] Z. Li, K. Li, X. Yao, Q. Wen, Predicting prepayment and default risks of unsecured consumer loans in online lending, Emerg. Mark. Finance Trade 55 (2019) (2018) 118–132.

[6] D.J. Hand, W.E. Henley, Statistical classification methods in consumer credit scoring: a review, J. R. Stat. Soc. Ser. A 160 (3) (1997) 523–541.

[7] M. Stepanova, L. Thomas, Survival analysis methods for personal loan data, Oper. Res. 50 (2) (2002) 277–289.

[8] B.C. Alves, J.G. Dias, Survival mixture models in behavioural scoring, Expert Syst. Appl. 42 (8) (2015) 3902–3910.

[9] R. Malhotra, D.K. Malhotra, Evaluating consumer loans using neural networks, Omega 31 (2) (2003) 83–96.

[10] M. Malekipirbazari, V. Aksakalli, Risk assessment in social lending via random forests, Expert Syst. Appl. 42 (10) (2015) 4621–4631.

[11] S. Maldonado, C. Bravo, J. López, J. Pérez, Integrated framework for profit-based feature selection and SVM classification in credits coring, Decis. Support Syst. 104 (2017) 113–121.

[12] N. Arora, P.D. Kaur, A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment, Appl. Soft Comput. 86 (2020) 105936.

[13] Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, Y. Wang, Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending, Inform. Sci. 525 (2020) 182–204.

[14] J.N. Crook, D.B. Edelman, L.C. Thomas, Recent developments in consumer credit risk assessment, European J. Oper. Res. 183 (3) (2007) 1447–1465.

[15] X. Dastile, T. Celik, M. Potsane, Statistical and machine learning models in credit scoring: A systematic literature survey, Appl. Soft Comput. 91 (2020) 106263.

[16] R. Iyer, A.I. Khwaja, K. Shue (, Screening peers softly: inferring the quality of small borrowers, Manage. Sci. 62 (6) (2016) 1554–1577.

[17] M.F. Lin, N.R. Prabhala, S. Viswanathan, Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending, Manage. Sci. 59 (1) (2013) 17–35.

[18] D. Liu, D.J. Brass, Y. Lu, D. Chen, Friendships in online peer-to-peer lending: pipes, prisms, and relational herding, MIS Q. 39 (3) (2015) 729–742.

[19] M. Papouskova, P. Hajek, Two-stage consumer credit risk modelling using heterogeneous ensemble learning, Decis. Support Syst. 118 (3) (2019) 33–45.

[20] Y. Tian, Z. Yong, J. Luo, A new approach for reject inference in credit scoring using kernel-free fuzzy quadratic surface support vector machines, Appl. Soft Comput. 73 (2018) 96–105.

[21] M.C. So, L. Thomas, H.V. Seow, C. Mues, Using a transactor/revolver scorecard to make credit and pricing decisions, Decis. Support Syst. 59 (2014) 143–151.

[22] J.M. Quigley, Interest rate variations, mortgage prepayments and household mobility, Rev. Econ. Stat. 69 (4) (1987) 636–643.

[23] J.C. Huang, G. Amromin, C. Sialm, The trade-off between mortgage prepayments and tax-deferred retirement savings, J. Publ. Econom. 91 (2007) 2014–2040.

[24] B. Ambrose, A. Sanders, Commercial mortgage-backed securities: prepayment and default, J. Real Estate Finance Econom. 26 (2003) 179–196.

[25] K.B. Dunn, C.S. Spatt, An analysis of mortgage contracting: prepayment penalties and the due-on-sale clause, J. Finance 40 (1) (1985) 293–308.

[26] Y. Varli, Y. Yildirim, Default and prepayment modelling in participating mortgages, J. Bank. Financ. 61 (2015) 81–88.

[27] J. Steinbuks, Effects of prepayment regulations on termination of subprime mortgages, J. Bank. Financ. 59 (2015) 445–456.

[28] W.J. Boyes, D.L. Homan, S.A. Low, An econometric analysis of the bank credit scoring problem, J. Econometrics 40 (1) (1989) 3–14.

[29] S. Finlay, Credit scoring for profitability objectives, European J. Oper. Res. 202 (2) (2010) 528–537.

[30] T. Verbraken, C. Bravo, R. Weber, B. Baesens, Development and application of consumer credit scoring models using profit-based classification measures, European J. Oper. Res. 238 (2) (2014) 505–513.

[31] L.J.S. Barrios, G. Andreeva, J. Ansell, Monetary and relative scorecards to assess profits in consumer revolving credit, J. Oper. Res. Soc. 65 (3) (2014) 443–453.

[32] L.J.S. Barrios, G. Andreeva, J. Ansell, Time-to-profit scorecards for revolving credit, European J. Oper. Res. 249 (2) (2016) 397–406.

[33] C. Serrano-Cinca, B. Gutirrez-Nieto, The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (p2p) lending, Decis. Support Syst. 89 (2016) 113–122.

[34] Y. Guo, W. Zhou, C. Luo, C. Liu, H. Xiong, Instance-based credit risk assessment for investment decisions in p2p lending, European J. Oper. Res. 249 (2) (2016) 417–426.

[35] J. López, S. Maldonado, Profit-based credit scoring based on robust optimization and feature selection, Inform. Sci. 500 (2019) 190–202.

[36] M. Markowitz, Foundations of portfolio theory, J. Finance 46 (2) (1991) 469–477.

[37] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall/CRC, 1986.