



Instance-based credit risk assessment for investment decisions in P2P lending



Yanhong Guo^a, Wenjun Zhou^{b,*}, Chunyu Luo^{a,c}, Chuanren Liu^{c,d}, Hui Xiong^c

^a Faculty of Management and Economics, Dalian University of Technology, Dalian, Liaoning 116024, China

^b Department of Business Analytics and Statistics, University of Tennessee, Knoxville, Tennessee 37996-0532, USA

^c Management Science and Information Systems Department, Rutgers University, Newark, New Jersey 07102, USA

^d Decision Sciences and MIS Department, Drexel University, Philadelphia, Pennsylvania 19104, USA

ARTICLE INFO

Article history:

Received 17 January 2014

Accepted 20 May 2015

Available online 29 May 2015

Keywords:

Data mining

P2P lending

Credit risk assessment

Instance-based method

Investment decisions

ABSTRACT

Recent years have witnessed increased attention on peer-to-peer (P2P) lending, which provides an alternative way of financing without the involvement of traditional financial institutions. A key challenge for personal investors in P2P lending marketplaces is the effective allocation of their money across different loans by accurately assessing the credit risk of each loan. Traditional rating-based assessment models cannot meet the needs of individual investors in P2P lending, since they do not provide an explicit mechanism for asset allocation. In this study, we propose a data-driven investment decision-making framework for this emerging market. We designed an instance-based credit risk assessment model, which has the ability of evaluating the return and risk of each individual loan. Moreover, we formulated the investment decision in P2P lending as a portfolio optimization problem with boundary constraints. To validate the proposed model, we performed extensive experiments on real-world datasets from two notable P2P lending marketplaces. Experimental results revealed that the proposed model can effectively improve investment performances compared with existing methods in P2P lending.

© 2015 Elsevier B.V. and Association of European Operational Research Societies (EURO) within the International Federation of Operational Research Societies (IFORS). All rights reserved.

1. Introduction

Peer-to-peer (P2P) lending, also known as person-to-person lending, allows individuals to directly lend to and borrow from each other on an Internet-based platform, without the involvement of traditional financial intermediaries. In this marketplace, borrowers submit applications for loans, called listings, by specifying loan details, such as loan amount and description. Then, prospective lenders are allowed to fund these listings partially by specifying the loan amounts they will provide. If the total dollar amount requested by a listing is fulfilled within a pre-specified period of time, the listing becomes a loan. Due to the elimination of a traditional financial intermediary, and a more dynamic environment that taps into the collective intelligence of the crowd, P2P lending has the potential to reduce financing costs and increase efficiency of the financial market.

To help personal investors manage risk, state-of-the-art P2P lending platforms, such as Prosper and Lending Club, provide risk ratings for each loan. Such ratings take into account many variables about the loan and the applicant, such as FICO scores, the amounts and terms

of the loans, borrowers' assets, debts, types of employment, and so on. This process of credit rating is similar to those traditional evaluation techniques commercial banks employ to evaluate the risk levels of borrowers. As a result, loans are categorized into a small number of risk groups. For example, Prosper uses a seven-level risk rating (i.e., AA, A, B, C, D, E, NR). This "rating-based" model provides basic evaluation of loans' credit risk, and the loans within each rating group are assumed to bear the same level of risk. In order to diversify their portfolios, investors are allowed to pick loans from different risk groups, according to their risk-return preferences.

Rating-based credit risk assessment models are practical and has been widely utilized by financial institutions for issuing loans. However, this model may be too coarse to meet the needs of personal investors in P2P lending. Since traditional financial institutions possess large amounts of cash, they are able to fund millions of loans, which facilitates myriad opportunities for diversification. For P2P lending, each personal investor has much less money for investment. However, because they are allowed to partially fund each loan, effective diversification has been made possible. Therefore, investors in the P2P lending marketplace not only need to decide which loans to fund, but also how much money to allocate to each of them, which minimizes risk for a given expected return. While this feature presents as a typical portfolio optimization problem, it is very challenging to

* Corresponding author. Tel.: +1 865 974 9198.

E-mail address: wzhou4@utk.edu, wzhou7@gmail.com (W. Zhou).

accurately assess the credit risk of each individual loan for P2P lending, which will serve as necessary input for portfolio optimization. Moreover, real-world P2P lending marketplaces typically impose minimum investment amount constraints. For example, the minimum required investment on each loan is \$25 at Prosper. Such constraints are substantial for individual investors in P2P lending, due to their limited amount of total funds. Thus, a new decision making tool is required for lucrative P2P lending.

In this study, we propose a more accurate credit risk model for P2P lending, which allows personal investors to optimize investment decisions. Without enough historical data of closed loans for each individual borrower, it is impossible to predict the performance of new loans based on a borrower's past ones. As a result, we identified past loans with similar attributes, and aggregated performance outcomes to predict the performance of a new loan. This is called an "instance-based" approach.

Specifically, in the proposed instance-based model, we first defined the similarity of loans as the difference between their probabilities to default, which was derived from a logistic regression of borrower's credit attributes. We then predicted the return of each loan as a weighted average of similar loans, and the risk as the weighted variance, where the optimal weights were learned within the mathematical framework of kernel regression. Furthermore, based on this instance-based credit risk model, we formulated the investment decisions in P2P lending as a portfolio optimization problem with boundary constraints. To validate the proposed model, we performed extensive experiments on real-world data from two state-of-the-art P2P lending marketplaces. Experimental results demonstrated that the proposed model effectively outperforms existing rating-based models.

The remainder of this paper is organized as follows. A literature review is performed in Section 2, and a description of the data is provided in Section 3. Section 4 is devoted to describing the instance-based model for credit risk assessment. Combined with a formal formulation of portfolio optimization in P2P lending, Section 5 describes the investment decision model. To better explain the model, an end-to-end example of the instance-based credit risk assessment and investment decision process is provided in the appendix. In Section 6, we report empirical results on effectiveness of the proposed model. Finally, Section 7 concludes this work.

2. Literature review

P2P lending has been introduced as a new e-commerce phenomenon in the financial field for its potential to provide more economical efficiencies, as supported by an Internet-based information system (Berger & Gleisner, 2009; Wang, Greiner, & Aronson, 2009). P2P lending is also known as online social lending (Hulme & Wright, 2006), microlending (Conlin, 1999) or crowdfunding (Belleflamme, Lambert, & Schwienbacher, 2014; Mollick, 2014), each of which emphasizes a different set of unique characteristics of P2P lending. Namely, the online social interactions among participants, the small amounts of transactions, and the large number of investors who may not be experts in loan investing.

Participants in the P2P lending marketplace can be roughly divided into two groups: borrowers and lenders. Like traditional credit marketplaces, risk assessment and decision making can be viewed from their different perspectives (Wu & Hsu, 2012). From the borrower's perspective, a common goal is securing loan funding. Among all the information provided by the borrower, researchers aim to find determinants of the success of a loan. For instance, Larrimore, Jiang, Larrimore, Markowitz, and Gorski (2011) analyzed the impact of language used in the listing, which can help borrowers strengthen persuasiveness for getting funds. Puro, Teich, Wallenius, and Wallenius (2010) proposed a borrower decision aid system to help borrowers quantify their strategic options, such as starting interest rate and the

amount of loan to request, in order to have their listings successfully funded. In spite of a focus on borrower's decisions, their models may also be adapted for lenders, aiming for predicting loans' pay-back statuses. Specifically, they formulate a query-based approach, which looks for similar listings in the past to make an overall prediction for a new listing. They report that this technique is equivalent to the logistic regression model in most cases, but may become less stable when only a small number of similar loans can be found. Instead of identifying similar loans in the past, our framework of instance-based model uses all loans in the past, and optimize their weights through kernel regression. Other works in borrower decision support include Wu and Xu (2011), which proposes an intelligent agent system that make recommendations to the borrowers. With a focus on system design rather than analytical modeling, their work is not being compared with ours.

When considering the lenders' perspective, researchers investigate lender's investment decision making and bidding behavior. For example, Sonenshein, Herzenstein, and Dholakia (2010) examined the roles of social accounts in influencing lenders' decisions about lending money to borrowers. Klafft (2008) examined whether lenders can profit from this new market as claimed. Based on data from Prosper.com, they found that it is possible for lenders to achieve satisfactory returns if they employ a sound strategy. Moreover, Krumme and Herrero-Lopez (2009) found that many transactions are based upon sub-optimal decisions. Puro, Teich, Wallenius, and Wallenius (2011) provided bidding strategies in the context of small loan auctions. Their study showed that bidding behavior is not homogeneous among bidders, but bidders use many different bidding strategies. It has been found that investors are more likely to bid on loans with more existing bids (Herzenstein, Dholakia, & Andrews, 2011). This phenomenon is known as herding, and is commonly observed in online social loan auctions. Existing studies provide valuable insights into how lenders screen and select loans, and succeed in bidding. However, effective decision support is lacking to guide personal investors' selection of investments, and determination of optimal amounts to put forward in each of them.

Since making decisions on whether or not to fund certain loans is a central task for investors in the P2P lending marketplace, traditional loan evaluation techniques also relate to P2P lending. In order to distinguish good loan applications (i.e., those expected to be successfully paid back) from bad ones (i.e., those predicted to default), many different data mining techniques have been used in the literature. A number of examples include logistic regression (Thomas, 2009; Wiginton, 1980), linear discriminate analysis (Rosenberg & Gleit, 1994), k-nearest neighbors (Chatterjee & Barcun, 1970), classification trees (Feldman & Gross, 2005), Markov chains (Frydman, Kallberg, & Kao, 1985; So & Thomas, 2011), survival analysis (Andreeva, Ansell, & Crook, 2007; Stepanova & Thomas, 2002), linear and nonlinear programming (Bugera, Konno, & Uryasev, 2002; Mangasarian, 1965), neural networks (Malhotra & Malhotra, 2002; Yang, Li, Ji, & Xu, 2001), support vector machines (Huang, Chen, Hsu, Chen, & Wu, 2004; Kim & Sohn, 2010), genetic methods (Desai, Conway, Crook, & Overstreet, 1997; Huang, Tzeng, & Ong, 2006), and so on. These studies examine the classification of each loan, but do not scrutinize loan investment portfolio as a whole.

In recent years, kernel methods have been broadly applied in many credit scoring and optimization problems (Yang, 2007; Zhang, Gao, & Shi, 2014). In this study, we exploit the mathematical framework of the kernel regression (Nadaraya, 1965) and extract the regression coefficients as the optimal weights for credit risk assessment. As a flexible statistical technique to study nonlinear relationships, kernel regression is increasingly utilized in many financial and economical studies, such as nonparametric VaR measurements (Ait-Sahalia & Lo, 2000), nonlinear relationships between international real interest rates (Mancuso, Goodwin, & Grennes, 2003), and time-varying diffusion processes for forecasting financial

crashes (Fernandes, 2006). Our study illuminates suitability of kernel regression for assisting individual investors in P2P lending to make decisions.

3. Data description

Advances in the development of P2P lending markets have provided large amounts of real-world P2P lending transaction data. Our empirical study is based on public datasets from two state-of-the-art P2P lending marketplaces.

The first dataset is from Lending Club, which consists of 2016 loans. The second dataset is from Prosper, which consists of 4128 loans. In both datasets, a predictive model is used to transform loan attributes into the probability of default. These attributes include the borrower's credit score from FICO, the amount of the loan, the borrower's number of inquires in the past six months, the borrower's debt-to-income ratio, the borrower's number of current delinquencies, and the borrower's home ownership status.

Many predictive models are available for quantifying the likelihood of loans to default. Choice and implementation of such a predictive model is beyond the scope of our study, since we assume the probability of default of each loan is known. Among the many models that exist to predict the probabilities of loan default, logistic regression has been the most widely used in the literature. For example, Puro et al. (2010) uses logistic regression to determine the probability of successful loan funding in the P2P lending marketplace. As a data preprocessing step, we use logistic regression to predict each loan's probability of default in both datasets.

4. Instance-based credit risk assessment

In this section, we first describe an instance-based credit risk assessment model to quantitatively assess the return and risk of loans in P2P lending. Then, we transform loans' default likelihood distances into optimized weights, exploiting the mathematical framework of kernel regression.

4.1. Instance-based models

It is common practice to predict the future using data from the past. Unlike working with stocks, there are very few historical observations for each individual borrower to analyze to make predictions on his or her new loan request. Alternatively, a feasible approach is to use similar loans, such as borrowers with similar profiles, to predict the credibility of a new loan request. Directly using other instances to make predictions about a particular instance is called an "instance-based" method.

In particular, for loan portfolio optimization, we would like to predict the return and risk of each loan. For a given loan i (i.e., the "focal" instance), based on n past loans (i.e., the "voting" members), each with an observed return rate R_j ($j = 1, 2, \dots, n$), we can directly predict loan i 's return, μ_i , using a weighted average of the performance of past loans

$$\mu_i = \sum_{j=1}^n w_{ij} R_j. \quad (1)$$

Here, w_{ij} represents the weight of loan j for predicting the return of loan i . This is similar to a "voting" scheme, where each vote is counted with a different weight. A critical step for instance-based approaches is to determine the weights of the "voting" instances. Our solution for determining optimized weights will be presented in Section 4.2.

Using a weighted average of similar loans' performance to predict the return of a new loan is straightforward. However, for the prediction of risk, we cannot take such a weighted average directly, since

the risk values of past loans are not quantified. As a result, we quantify the risk of a new loan i , σ_i^2 , as the weighted variance among the voting members

$$\sigma_i^2 = \sum_{j=1}^n w_{ij} (R_j - \mu_i)^2, \quad (2)$$

where μ_i is the weighted average given in Eq. (1), and w_{ij} carries the same meaning as in Eq. (1). Since a low-risk loan tends to bear a stable return, it is intuitive that the variation of the return rates among voting members can serve as an indicator of risk.

The weights for instance-based models are based on the proximity of the focal instance to its voting members. Intuitively, more similar items should have more weights, whereas less similar ones should have less weights. For loan assessment, the proximity between instances are primarily tied to their probability of default. In particular, we define the *default likelihood distance* between loans i and j , as

$$d_{ij} = |p_i - p_j|, \quad (3)$$

where p_i and p_j are the probabilities of default for loans i and j , respectively. Intuitively, the more default likelihood distance, the less weight should be presented in w_{ij} . Naïve weighting schemes, such as equal weights of nearest neighbors and reciprocal of distances, are easy to implement, but are not optimized for the most accurate predictions. To determine the function that maps the raw distances into optimal weights, we will employ kernel regression, to be presented in the next subsection.

4.2. Kernel weights for instance-based modeling

Kernel regression (Nadaraya, 1965) is a statistical technique to find non-linear relation between a pair of random variables. In particular, suppose that each observation is evaluated on two dimensions, X and Y , where X is the predictive variable and Y is the response variable. With observations of n instances, $\{(x_j, y_j) | j = 1, 2, \dots, n\}$, the estimation of an outcome y given its predictive observation x , will be

$$y = \hat{f}(x) = \frac{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)y_j}{\sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)}. \quad (4)$$

Here, $K(\cdot)$ is a kernel function that is usually specified as Gaussian. More specifically,

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}. \quad (5)$$

In this way, more weights are assigned to observations closer to x and less weights to those further apart. The parameter h ($h > 0$) is called the "bandwidth," which determines the proportion of local versus remote information used in the summation.

In the context of our instance-based risk modeling, the probability of default p_j serves as the predictive variable, and the return rate R_j is the response variable. Thus, we may rewrite Eq. (4) as

$$\mu_i = \frac{\sum_{j=1}^n K\left(\frac{x_i-x_j}{h}\right)y_j}{\sum_{j=1}^n K\left(\frac{x_i-x_j}{h}\right)} = \frac{\sum_{j=1}^n K\left(\frac{p_i-p_j}{h}\right)R_j}{\sum_{j=1}^n K\left(\frac{p_i-p_j}{h}\right)}, \quad (6)$$

where μ_i , R_j , p_i , and p_j follow the same definitions as in Section 4.1. By comparing the predicted return rate in Eq. (1) with the kernel regression in Eq. (6), it is straightforward to extract the optimal weight as

$$w_{ij} = \frac{K\left(\frac{p_i-p_j}{h}\right)}{\sum_{j=1}^n K\left(\frac{p_i-p_j}{h}\right)} = \frac{K\left(\frac{d_{ij}}{h}\right)}{\sum_{j=1}^n K\left(\frac{d_{ij}}{h}\right)}, \quad (7)$$

where d_{ij} is given in Eq. (3). In other words, the loan weight w_{ij} is computed with the default likelihood distance d_{ij} , and loans with smaller default likelihood distances will carry more weights.

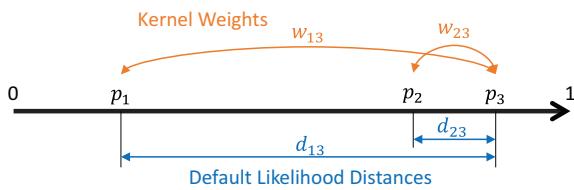


Fig. 1. We estimate the investment return of L_3 by computing the weights between L_3 and L_1, L_2 , respectively.

To fit the kernel regression model, the bandwidth h needs to be optimized based on training data. To this end, Clark (1975) proposed the leave-one-out least-squares cross-validation method. Specifically, the bandwidth is chosen to minimize the following cross-validation error

$$CV(h) = \frac{1}{n} \sum_{i=1}^n \left(\hat{f}_h(x_{-i}) - y_i \right)^2, \quad (8)$$

where $\hat{f}(x_{-i})$ is the leave-one-out estimation of y_i , using kernel regression. In other words,

$$\hat{f}_h(x_{-i}) = \frac{\sum_{j=1, j \neq i}^n K\left(\frac{x_i - x_j}{h}\right) y_j}{\sum_{j=1, j \neq i}^n K\left(\frac{x_i - x_j}{h}\right)}. \quad (9)$$

In our context of instance-based risk modeling, we may rewrite Eq. (8) as

$$CV(h) = \frac{1}{n} \sum_{j=1}^n \left[\frac{\sum_{k=1, k \neq j}^n K\left(\frac{p_j - p_k}{h}\right) R_j}{\sum_{k=1, k \neq j}^n K\left(\frac{p_j - p_k}{h}\right)} - R_j \right]^2. \quad (10)$$

To find the optimal bandwidth efficiency, we adopt a mixed bandwidth selection strategy recommended by Silverman (1986). This strategy searches for the optimal bandwidth between $0.25h_0$ and $1.5h_0$, where h_0 is given by

$$h_0 = \left(\frac{4}{3n} \right)^{1/5} \sigma. \quad (11)$$

Since we use Gaussian kernel, σ is the standard deviation of the predictive variable (i.e., the probability of default in our case).

4.3. Distances and kernel weights: an illustrative example

In Fig. 1, we provide an illustrative example of default likelihood distances and kernel weights. In particular, the horizontal axis represents the space for probabilities of default for all loans, which ranges from 0 to 1. For demonstrative purposes, we consider three loans: L_1 , L_2 , and L_3 , each of which is placed on the axis properly, according to their respective probabilities of default, p_1 , p_2 , and p_3 . We further assume L_3 is our focal loan instance, whereas L_1 and L_2 are the voting members.

A straightforward way of assessing the similarity between different loans is to calculate the default likelihood distances, as defined in Eq. (3). As a result, we may calculate $d_{13} = |p_1 - p_3|$, and $d_{23} = |p_2 - p_3|$. From Fig. 1, we can tell clearly that $d_{13} > d_{23}$, and thus the voting weight from L_1 should be lower than that from L_2 . Transformation from such distances to voting weights can be done using an arbitrary non-increasing function, but an optimal transformation can be learned as a kernel function whose parameters can be optimized by kernel regression. In our case, we follow Eq. (7), so that $w_{13} < w_{23}$. This is because the Gaussian kernel, as described in Eq. (5), provides a larger output value for inputs with smaller absolute values. Compared with parametric regression of loan returns, the non-parametric kernel regression is able to approximate arbitrary relationships without assuming oversimplified data distributions.

5. Investment decision model

In this section, we first formulate investment decisions in P2P lending as a portfolio selection problem with boundary constraints. Then, we combine the instance-based credit risk assessment model described in Section 4 with portfolio selection, and summarize the unified investment decision process. Finally, we describe our model and baseline models for experimental comparison.

5.1. Portfolio selection in P2P lending

Different from traditional banks, personal investors in P2P lending not only need to distinguish good loans from bad ones, but also need to know how to spread optimal proportions of available money across different loans to minimize risk for a given level of expected return. As a result, investing in P2P lending can be formulated as a typical portfolio selection problem, whose basis is modern portfolio theory (Markowitz, 1991). Specifically, by ignoring the negligible correlation between different loans, the objective of the portfolio selection problem is to minimize the investment risk

$$\sum_i \lambda_i^2 \sigma_i^2,$$

subject to $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$, and a given level of expected return

$$R^* = \sum_i \lambda_i \mu_i.$$

The optimization variable λ_i represents the optimal proportion of money invested into the i -th loan.

Moreover, constraints on investment amounts need to be considered. On the one hand, P2P lending marketplaces usually require a minimum investment amount to each loan. For example, at Prosper, the minimum amount on each loan is $m = \$25$. Investors in P2P lending are especially sensitive to this constraint (Kraft & Steffensen, 2013; Woodside-Oriakhi, Lucas, & Beasley, 2011), because personal investors tend to have much less funds than commercial banks. On the other hand, for the i -th loan, the investor can only lend an amount less than the loan's requested amount e_i . As a result, suppose M is the total investment amount an investor has available, for a loan that he or she invests in (i.e., $\lambda_i > 0$), we require

$$m \leq \lambda_i M \leq e_i.$$

Otherwise, we have

$$\lambda_i = 0.$$

5.2. Procedure of investment decisions

As summarized in Algorithm 1, the investment decision process may be described as follows.

There are two input datasets and three parameters in the model. $DataH$ is the historical performance dataset that includes past loans whose pay-back statuses are known. We call this set the “training data.” $DataI$ is the current set of listings that are open for investment, and thus their pay-back statuses are unknown. We call this set “testing data.” The minimum investment amount (m) is determined universally at the P2P lending marketplace. The total investment amount M and the expected overall return rate R^* allow for personalizing to the needs or preferences of any individual investor.

In the initialization and training processes (Lines 1–3), we first find out the total number of historical loans in $DataH$ (n), the performance of each historical loan (R_j), the number of listings that are open for investment (I), and the amount requested by each listing (e_i). Then, we build a logistic regression model with $DataH$ and extract the default probability (p_j) of each loan. Finally, we use cross-validation to optimize the bandwidth (h).

Algorithm 1: Investment decision model.

Input : $DataH$: set of n historical loans;
 $DataI$: set of l listings that are open for investment;
 m : the minimum investment amount required for funding a loan;
 M : the total amount available for investment;
 R^* : the expected overall return rate.
Output: OM_i : recommended investment amount for listing i ($i = 1, 2, \dots, l$).

```

// Initialization and training
1  $(n, \{R_j\}_{j=1}^n, l, \{e_i\}_{i=1}^l) \leftarrow \text{Preprocess}(DataH, DataI);$ 
2  $(\{\hat{\beta}_k\}_{k=0}^p, \{\hat{p}_j\}_{j=1}^n) \leftarrow \text{LogisticRegression}(DataH);$ 
3  $h \leftarrow \text{KernelCrossValidation}(DataH);$ 
// Instance-based credit risk assessment
4 for  $i \leftarrow 1$  to  $l$  do
5    $\hat{p}_i \leftarrow \frac{1}{1+e^{-(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_p X_{pi})}};$ 
6   for  $j \leftarrow 1$  to  $n$  do
7      $d_{ij} \leftarrow |\hat{p}_i - \hat{p}_j|;$ 
8      $w_{ij} \leftarrow \text{KernelWeight}(d_{ij}, h);$ 
9   end
10   $\mu_i \leftarrow \sum_{j=1}^n w_{ij} R_j;$ 
11   $\sigma_i^2 \leftarrow \sum_{j=1}^n w_{ij} (R_j - \mu_i)^2;$ 
12 end
// Portfolio selection and investment recommendation
13 Minimize:  $\sum_{i=1}^l \lambda_i^2 \sigma_i^2;$ 
14 Subject to:
15    $\sum_{i=1}^l \lambda_i \mu_i = R^*$ 
16    $\sum_{i=1}^l \lambda_i = 1$ 
17    $\begin{cases} m \leq \lambda_i M \leq e_i, & \text{if loan } i \text{ is invested} \\ \lambda_i = 0, & \text{otherwise} \end{cases}$ 
18  $\{\lambda_i\} \leftarrow \text{Prog}(m, M, R^*, \{\mu_i, \sigma_i^2, e_i\}_{i=1}^l)$ 
19 Output  $\{OM_i = \lambda_i M\}$ 

```

In the credit risk assessment process (Lines 4–12), we first predict the default probability (p_i) of each loan in $DataI$ using the logistic regression model trained in Line 2. Then, for each loan i in $DataI$, we compute d_{ij} , its default likelihood distance to each loan j in $DataH$ ($i = 1, 2, \dots, l; j = 1, 2, \dots, n$). Furthermore, we compute the kernel weights w_{ij} based on d_{ij} and optimize bandwidth h by minimizing Eq. (10). Finally, we assess the return (μ_i) and the risk (σ_i^2) based on the kernel weights (w_{ij}) and the performance (R_j) of each historical loan in $DataH$ using Eqs. (1) and (2).

In the portfolio selection and investment recommendation process (Lines 13–19), we implement the model described in Section 5.1.

To further demonstrate how the instance-based credit risk assessment model works, we provide an end-to-end numeric example in Appendix A.

5.3. Model description

In this paper, we propose an instance-based decision model for risk assessment in P2P lending. In order to show its effectiveness, we compare it with two rating-based baselines. The models are described in detail below.

RBM is a basic rating-based model, where all the loans are segmented into seven grades according to the loans' credit risk. Loans in the same grade are assumed to have the same return and risk.

RBM+ is a refined rating-based model, where each grade in RBM is further segmented into five sub-grades, and the number of sub-grades is selected based on cross-validation. In other words, RBM+ is a rating-based model with optimized grade segmentation.

IOM is the instance-based model proposed in this study. Each loan is assessed using kernel weights and the historical performance of similar loans.

Table 1
Comparison of models' prediction accuracy.

	Prosper		Lending club	
	Distance	R^2	Distance	R^2
RBM	24.3816	0.2497	14.8607	0.1625
RBM+	24.3646	0.2523	14.7801	0.18943
IOM	24.3123	0.2603	14.7241	0.2070

In summary, RBM is the basic version of rating-based credit risk model that is used practically. RBM+ is an enhanced version of rating-based credit risk model, which is used to compare the results when finer levels of risk grades are implemented.

Moreover, we divide each dataset into k subsets and use k -fold cross-validation to verify the effectiveness of our model. In other words, each time we use one of the subsets as the testing set ($DataI$) and all others as the training set ($DataH$). Specifically, we divide the Lending Club dataset randomly into 12 subsets, and each subset contains about 168 loans. Similarly, the Prosper dataset is randomly divided into 24 subsets, and each subset contains about 172 loans.

We compare the three models by the following procedure:

1. Train each model with the training set, and use the trained model to predict the return rate (μ_i) and risk (σ_i) of each loan in the testing set.
 - For rating-based models, including RBM and RBM+, we first find the mean and standard deviation of the return rate of loans in each rating grade in the training set. Then, we predict the credit rating of each testing loan. The mean and standard deviation from that credit grade will be used as the predicted return and the predicted risk, respectively.
 - For the IOM model, we first train the kernel weights of all training loans (Eq. 7), and use such weights for predicting the return of a testing loan with the weighted average (Eq. 1) and risk with the weighted variance (Eq. 2).
2. Check the prediction accuracy of the models by comparing the predicted return rate (μ_i) and the actual return rate (R_i).
3. For each model, feed the predicted return μ_i and risk σ_i of each testing loan into the portfolio selection algorithm, and compute the return of investment on the optimal portfolio.
4. Compare the return rate and Sharpe ratio of the three models.

6. Results and discussions

Extensive experiments have been performed on real-world datasets described in Section 3. Specifically, we compare our model with two rating-based models as described in Section 5.3 with respect to prediction accuracy and investment performances.

We predict the default probability p_i using the logistic regression model and attribute variables of each loan. Then we run the credit risk model on the Lending Club and the Prosper datasets to predict the return rate μ_i and the risk σ_i^2 . Fig. 2 shows the relationship between the default probability and the return rate (or the risk) on the Lending Club and the Prosper datasets, respectively. For either dataset, the overall trend is that as the default probability increases, the predicted rate of return decreases, and the risk increases. It is obvious that the curves of IOM is smoother than RBM and RBM+, which demonstrates that IOM has the capacity of predicting the return and risk with better granularity, and provides more accurate overall prediction of loan returns.

Furthermore, we compare the predicted return rate μ_i by each model against the actual return rate R_i . Specially, we computed the Euclidean distance and R^2 as the accuracy metrics. As shown in Table 1, IOM model has better prediction accuracy than either the RBM or RBM+ model on both datasets.

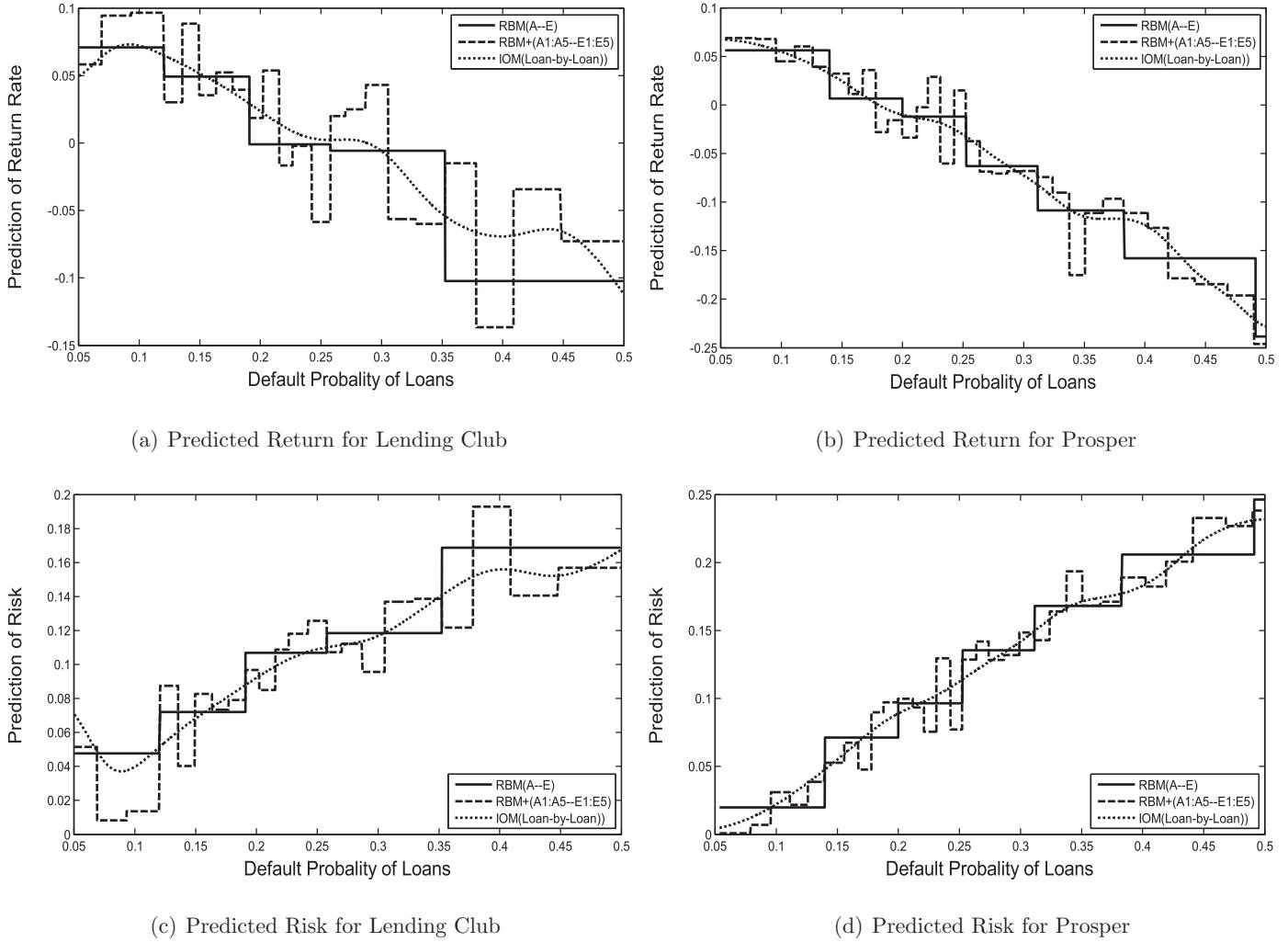


Fig. 2. Predicted rate of return (top row) and predicted risk (bottom row) versus probability of default for Lending Club (left column) and Prosper (right column).

Based on the predicted return μ_i and risk σ_i^2 from the output of each credit risk model, we execute the portfolio selection to find the optimal portfolio, and compute the overall rate of return. By default, the parameters are investment amount $M = \$15,000$ and target return rate $R^* = 0.06$. We also assume risk-free return rate to be 0.025, which is equivalent to the average return from T-Bills at the same period of time.

Table 2 shows the investment return rate on each testing subset and the average performance on Lending Club dataset. Both the average investment return rate and the Sharpe ratio demonstrate that the IOM model performs better than both the RBM and the RBM+ models. Similar results are found for the Prosper dataset, as shown in **Table 3**.

Finally, we find that the IOM credit risk assessment model can effectively improve the investment decisions for different investors as compared to rating-based credit models, when they have different investment amounts and expected returns. We run portfolio selection on both the Lending Club and the Prosper datasets with different combinations of input parameters.

As listed in **Table 4**, we consider nine different combinations of investment amounts and investors' expected return rates. For each combination of parameters, we use the same portfolio selection algorithm to find the optimal asset allocation (i.e., amount of money to be allocated to each loan), as described in **Section 5.1**. The only difference is that the input for portfolio selection (i.e., the predicted return and risk for each loan) is produced by each of the RBM, RBM+,

Table 2
Rate of return from the optimal portfolio on the Lending Club dataset.

Subset	RBM	RBM+	IOM
1	0.0487	0.0747	0.0302
2	0.0369	0.0244	0.0672
3	0.0588	0.0504	0.0554
4	0.0236	0.0398	0.0481
5	0.0276	0.0178	0.0624
6	0.0512	0.0617	0.0673
7	0.0421	0.0431	0.0335
8	0.0556	0.0763	0.0991
9	0.0373	0.0390	0.0701
10	0.0248	0.0670	0.0558
11	0.0110	0.0739	0.0775
12	0.0836	0.0712	0.0743
Average	0.0418	0.0533	0.0617
Std. dev.	0.0195	0.0204	0.0190
Sharpe ratio	0.8593	1.3846	1.9298

and IOM models, respectively, as described in **Section 5.3**. We compute the actual return of investment and Sharpe ratio on all loans for each set of parameters in **Table 4** as well as the average, as shown in **Fig. 3**.

Fig. 3 demonstrates a performance comparison of the three models on the Lending Club and Prosper datasets in terms of overall return and the Sharpe ratio. We can see that for each combination of

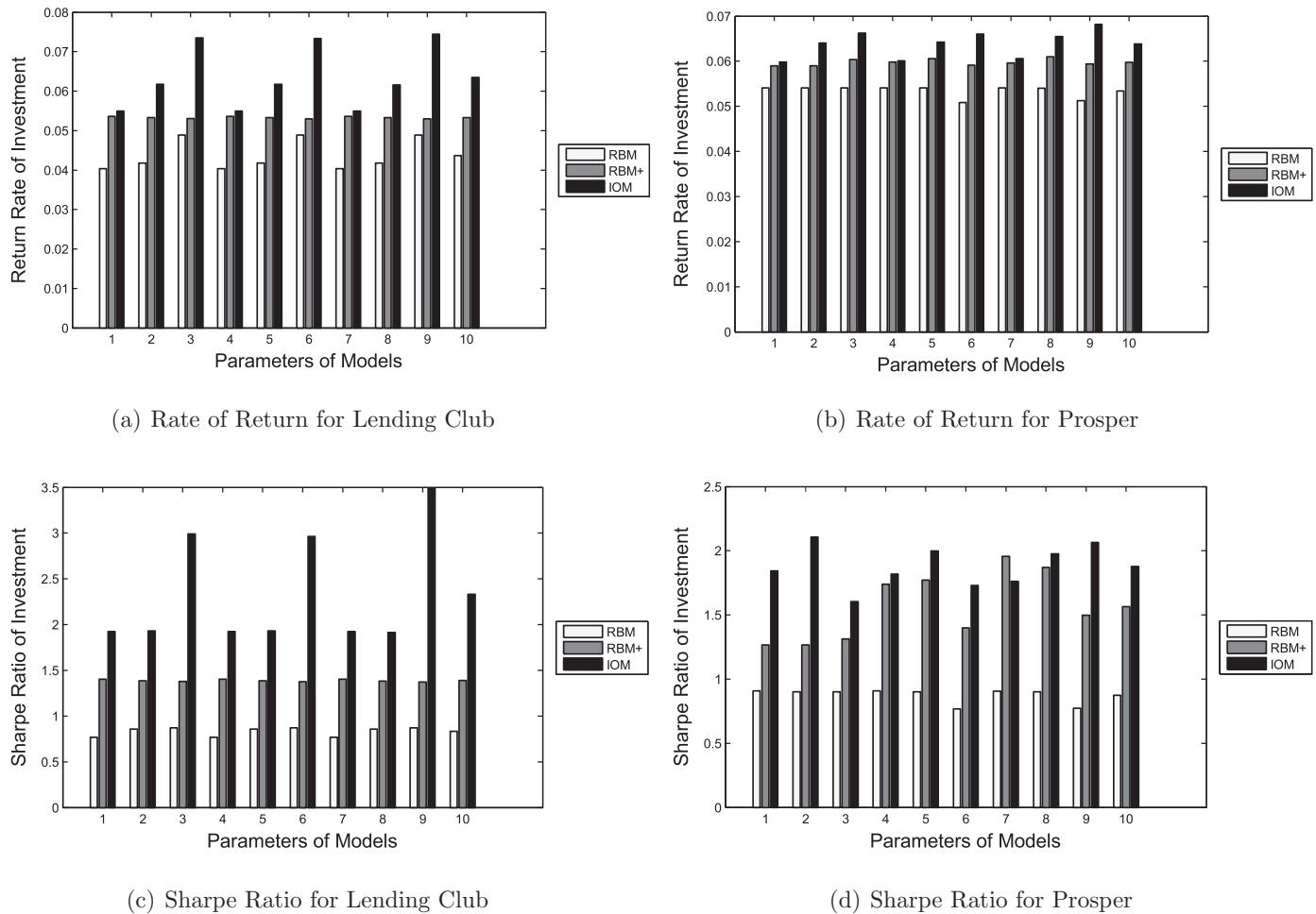
**Fig. 3.** Sensitivity analysis.

Table 3
Rate of return from the optimal portfolio on the Prosper dataset.

Subset	RBM	RBM+	IOM
1	0.0783	0.0679	0.0787
2	0.0840	0.0759	0.0761
3	0.0779	0.0634	0.0820
4	0.0690	0.0693	0.0765
5	0.0478	0.0609	0.0513
6	0.0498	0.0455	0.0727
7	0.0676	0.0722	0.0761
8	0.0849	0.0707	0.0752
9	0.0775	0.0752	0.0756
10	-0.0443	0.0289	0.0147
11	0.0878	0.0765	0.0768
12	0.0878	0.0671	0.0765
13	0.0373	0.0711	0.0780
14	-0.0034	0.0571	0.0701
15	0.0622	0.0674	0.0710
16	0.0482	0.0659	0.0549
17	0.0672	0.0725	0.0798
18	0.0590	0.0665	0.0717
19	0.0119	0.0633	0.0447
20	0.0654	0.0684	0.0738
21	0.0163	0.0674	0.0424
22	0.0496	0.0560	0.0708
23	0.0398	0.0420	0.0310
24	0.0760	-0.0173	0.0207
Average	0.0541	0.0606	0.0642
Std. dev.	0.0322	0.0201	0.0196
Sharpe ratio	0.9019	1.7725	1.9976

Table 4
Investors' choices of input parameters for portfolio selection.

Set	Investment amount (M)	Expected return (R^*)
1	\$10,000	5.5 percent
2	\$10,000	6.0 percent
3	\$10,000	6.5 percent
4	\$15,000	5.5 percent
5	\$15,000	6.0 percent
6	\$15,000	6.5 percent
7	\$20,000	5.5 percent
8	\$20,000	6.0 percent
9	\$20,000	6.5 percent

parameters (sets 1 through 9 from Table 4, as listed on the x-axis in each subplot) as well as the average (number 10 on the x-axis in each subplot), the IOM model performs better than both RBM and RBM+ models on both datasets, by having higher average investment return and higher Sharpe ratio.

In summary, our experiments show that the IOM credit risk assessment makes more accurate predictions of risk and returns. Moreover, when its output is used for portfolio selection, it yields better portfolios of investment than the rating-based models.

7. Conclusions

In this study, we proposed an instance-based credit risk assessment model to quantify the risk of each loan in the P2P lending marketplace, and formulated the investment decision in P2P

lending as a portfolio optimization problem with boundary constraints. The proposed credit risk assessment model has at least three major advantages. Firstly, it has the ability of quantifying the credit risk of each individual loan, instead of categorizing the loans into a small number of risk groups. Such granularity provides an easy way to compare risks of different loans, and the potential of more accurate predictions. Secondly, this model can assess the credit risk of a new loan based on its default likelihood distances to past loans, without requiring detailed historical observations of the same borrower. Finally, the weights of past loans are determined within the mathematical framework of kernel regression, which guarantees the optimality and does not require strong statistical assumptions. As demonstrated by experimental results on two real-world datasets, the proposed model has better performances than existing models available at the P2P lending marketplaces.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments. The research presented in this paper was partially supported by Natural Science Foundation of China (No. 71028002, 71372083, 71402014), Humanities and Social Science Program of the Ministry of Education in China (No. 14YJCZH044) and Fundamental Research Funds for the Central Universities (No. DUT15RW116).

Appendix A. A numeric example

In this section, we present an end-to-end example to demonstrate the instance-based credit risk assessment model for making investment decisions in P2P lending.

Step 1: Data preparation (Line 1 in [Algorithm 1](#)) Suppose that a historical dataset (*DataH*) is provided in the first seven columns in [Table A.5](#), which includes n training instances ($n = 30$). The input

Table A.5

The training set (*DataH*).

j	X_1	X_2	X_3	X_4	X_5	Y	R	\hat{p}
1	695	1	1050	1	0.0722	0	0.1578	0.2165
2	695	5	3500	0	0.2870	0	0.2614	0.0206
3	650	1	13000	1	0.1598	1	-0.5352	0.7269
4	670	7	7500	1	0.0546	1	-0.9015	0.9300
5	650	7	3600	1	0.1981	1	-0.0441	0.7432
6	670	2	10000	1	0.1316	0	0.1348	0.6392
7	695	0	3000	1	0.0690	1	-0.7739	0.2194
8	695	1	15000	1	0.0650	0	0.1706	0.7404
9	650	1	5000	1	0.0872	1	-0.7304	0.5992
10	730	7	7000	1	0.0760	1	-0.8400	0.6680
11	695	1	9000	1	0.1876	0	0.1531	0.2420
12	670	3	8500	0	0.0638	0	0.2040	0.3030
13	650	8	5100	1	0.0699	1	-0.8651	0.9483
14	695	2	6000	1	0.0857	1	-0.2786	0.4290
15	695	0	7000	1	0.1429	0	0.1736	0.2052
16	650	2	4000	1	0.0898	0	0.2225	0.6331
17	695	2	4000	1	0.0707	0	0.1563	0.3867
18	650	3	12000	0	0.0974	1	-0.3377	0.4833
19	695	1	3000	1	0.0360	0	0.1575	0.3519
20	670	7	8000	1	0.2540	1	-0.9301	0.6659
21	695	0	4000	1	0.0808	1	-0.3478	0.2272
22	695	0	9000	1	0.1468	1	-0.5105	0.2557
23	650	1	5000	1	0.0000	1	-0.6612	0.7803
24	650	3	11000	1	0.1852	1	-0.2322	0.7437
25	800	1	12000	1	0.0888	0	0.1218	0.0783
26	670	6	4800	1	0.2272	0	0.1734	0.5263
27	650	3	4000	1	0.1506	0	0.2216	0.5679
28	695	2	6000	1	0.1791	0	0.1641	0.2292
29	695	1	3000	1	0.0871	0	0.1639	0.2465
30	670	1	5000	0	0.0000	0	0.1911	0.1931

Table A.6

Description of variables in [Tables A.5](#) and [A.7](#).

Variable	Description
X_1	The FICO credit score of the borrower
X_2	The number of inquiries of the borrower in the last 6 months
X_3	The monetary amount of the loan
X_4	The homeownership status of the borrower (0 = rent, 1 = own)
X_5	The debt-to-income ratio of the borrower
Y	Dependent variable (0 = paid, 1 = default)

Table A.7

The testing set (*DataI*).

i	X_1	X_2	X_3	X_4	X_5	\hat{p}	μ	σ^2	λ
1	730	6	7500	0	0.0703	0.1597	-0.0170	0.1089	0.00
2	670	3	3300	1	0.1740	0.3534	-0.0214	0.0881	0.00
3	730	1	2100	1	0.0236	0.1731	-0.0227	0.1103	0.00
4	695	1	12375	1	0.0821	0.6112	-0.2481	0.2047	0.00
5	730	0	10000	1	0.0744	0.2464	-0.0325	0.1093	0.00
6	800	0	5000	0	0.0456	0.0034	0.1427	0.0440	0.65
7	730	0	18000	1	0.1101	0.4564	-0.0782	0.1020	0.00
8	730	1	16000	1	0.1923	0.2722	-0.0299	0.1059	0.00
9	760	3	2300	1	0.1160	0.0702	0.0619	0.0812	0.35
10	695	4	4000	1	0.2250	0.2092	-0.0315	0.1115	0.00

Table A.8

Estimated coefficients for the logistic regression model in [Eq. \(A.1\)](#).

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
1.524×10	-2.658×10^{-2}	3.313×10^{-1}	1.622×10^{-4}	2.166	-9.923

variables (Columns 2–6 in [Table A.5](#)) and the target variable (Column 7 in [Table A.5](#)) are described in [Table A.6](#).

Also suppose that a testing dataset (*DataI*) of l instances ($l = 10$) is provided in the first six columns in [Table A.7](#), with the same input variables, but values of the target variable are unknown.

Step 2. Logistic regression (Line 2 in [Algorithm 1](#)) Based on the training set *DataH*, fit a logistic regression model. More specifically, estimate the parameters $\hat{\beta}_k$ ($k = 0, 1, \dots, 5$) in the following equation

$$\text{logit}(\hat{p}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5. \quad (\text{A.1})$$

Based on the training set *DataH*, we find the estimated coefficients as listed in [Table A.8](#). Using [Eq. \(A.1\)](#), we find the predicted probability of default for all training instances, as listed in Column 9 of [Table A.5](#).

Note that since logistic regression modeling is beyond the scope of our study, and that we are using a small dataset for illustrative purposes, we ignore many issues in logistic regression, such as assumption checking, variable transformation, model selection, and diagnosis. Here we assume that these tasks are robustly performed without distraction from our main procedure.

Step 3: Scoring the testing data (Line 5 in [Algorithm 1](#)) Using [Eq. \(A.1\)](#), find the predicted probability of default for all testing instances, as listed in the column of \hat{p} in [Table A.7](#).

We can then calculate the similarity between each pair of testing instance and training instance, using [Eq. \(3\)](#). More specifically, we produce the matrix of d_{ij} in [Table A.9](#), where rows correspond to training instances and columns correspond to testing instances. These raw distances only measure proximity among loans in terms of probability to default, but do not provide weights of voting from training instances whose return rates are known. In the next few steps, we find such weights by utilizing kernel regression.

Step 4: Bandwidth optimization (Line 3 in [Algorithm 1](#)) The bandwidth is to be optimized by using the training data only. Since the sample size is $n = 30$ and the standard deviation of \hat{p} is $\sigma = 0.2581$, according to [Eq. \(11\)](#), we find that

$$h_0 = \left(\frac{4}{3 \times 30} \right)^{1/5} \times 0.2581 = 0.1385.$$

Table A.9
Distances in probability of default between testing and training samples

	1	2	3	4	5	6	7	8	9	10
1	0.0568	0.1369	0.0434	0.3947	0.0299	0.2131	0.2399	0.0557	0.1463	0.0073
2	0.1391	0.3328	0.1525	0.5906	0.2258	0.0172	0.4358	0.2516	0.0496	0.1886
3	0.5672	0.3735	0.5538	0.1157	0.4805	0.7235	0.2705	0.4547	0.6567	0.5177
4	0.7703	0.5766	0.7569	0.3188	0.6837	0.9266	0.4736	0.6578	0.8598	0.7209
5	0.5835	0.3897	0.5701	0.1319	0.4968	0.7397	0.2867	0.4709	0.6729	0.5340
6	0.4795	0.2858	0.4661	0.0280	0.3928	0.6358	0.1828	0.3670	0.5690	0.4300
7	0.0597	0.1340	0.0463	0.3919	0.0270	0.2159	0.2371	0.0528	0.1491	0.0102
8	0.5807	0.3870	0.5673	0.1291	0.4940	0.7369	0.2840	0.4682	0.6702	0.5312
9	0.4395	0.2457	0.4261	0.0121	0.3528	0.5957	0.1427	0.3270	0.5289	0.3900
10	0.5083	0.3145	0.4949	0.0567	0.4216	0.6645	0.2115	0.3958	0.5977	0.4588
11	0.0823	0.1114	0.0689	0.3692	0.0044	0.2386	0.2144	0.0302	0.1718	0.0328
12	0.1433	0.0504	0.1299	0.3082	0.0566	0.2996	0.1534	0.0308	0.2328	0.0939
13	0.7886	0.5949	0.7752	0.3371	0.7019	0.9449	0.4919	0.6761	0.8781	0.7391
14	0.2693	0.0756	0.2559	0.1822	0.1826	0.4256	0.0274	0.1568	0.3588	0.2198
15	0.0455	0.1482	0.0321	0.4060	0.0412	0.2018	0.2512	0.0670	0.1350	0.0040
16	0.4734	0.2796	0.4599	0.0218	0.3867	0.6296	0.1766	0.3608	0.5628	0.4239
17	0.2270	0.0333	0.2136	0.2246	0.1403	0.3832	0.0698	0.1145	0.3164	0.1775
18	0.3236	0.1298	0.3102	0.1280	0.2369	0.4798	0.0268	0.2110	0.4130	0.2741
19	0.1922	0.0015	0.1788	0.2593	0.1056	0.3485	0.1045	0.0797	0.2817	0.1428
20	0.5062	0.3125	0.4928	0.0547	0.4195	0.6625	0.2095	0.3937	0.5957	0.4567
21	0.0675	0.1262	0.0541	0.3841	0.0192	0.2238	0.2292	0.0450	0.1570	0.0180
22	0.0960	0.0977	0.0826	0.3555	0.0093	0.2523	0.2007	0.0165	0.1855	0.0466
23	0.6206	0.4269	0.6072	0.1690	0.5339	0.7768	0.3238	0.5081	0.7100	0.5711
24	0.5840	0.3903	0.5706	0.1325	0.4973	0.7403	0.2873	0.4715	0.6735	0.5345
25	0.0814	0.2751	0.0948	0.5330	0.1681	0.0748	0.3781	0.1939	0.0080	0.1309
26	0.3666	0.1729	0.3532	0.0849	0.2799	0.5229	0.0699	0.2541	0.4561	0.3171
27	0.4082	0.2145	0.3948	0.0433	0.3215	0.5645	0.1115	0.2957	0.4977	0.3587
28	0.0695	0.1242	0.0561	0.3820	0.0172	0.2258	0.2272	0.0430	0.1590	0.0200
29	0.0868	0.1070	0.0734	0.3648	0.0001	0.2430	0.2100	0.0258	0.1762	0.0373
30	0.0334	0.1603	0.0200	0.4182	0.0533	0.1896	0.2634	0.0791	0.1228	0.0161

Table A.10
Different bandwidths and corresponding cross-validation errors.

h/h_0	$CV(h)$	h/h_0	$CV(h)$	h/h_0	$CV(h)$	h/h_0	$CV(h)$
0.25	0.1663	0.60	0.1517	0.95	0.1528	1.30	0.1535
0.30	0.1635	0.65	0.1515	1.00	0.1529	1.35	0.1536
0.35	0.1605	0.70	0.1516	1.05	0.1531	1.40	0.1537
0.40	0.1578	0.75	0.1518	1.10	0.1532	1.45	0.1538
0.45	0.1554	0.80	0.1521	1.15	0.1532	1.50	0.1540
0.50	0.1536	0.85	0.1524	1.20	0.1533		
0.55	0.1524	0.90	0.1526	1.25	0.1534		

Table A.11
Kernel weights between testing samples and training samples.

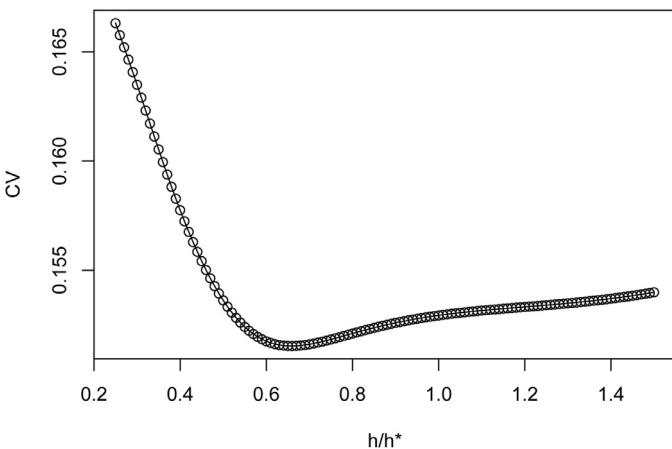


Fig. A.4. Identifying the optimal bandwidth h that minimizes $CV(h)$.

Using Eq. (8), i.e., leave-one-out least-squares cross validation, calculate the $CV(h)$ for all possible values of h , varying from $0.25h_0$ to $1.5h_0$. For any given bandwidth h , the calculation of $CV(h)$ is based on Eq. (10).

For each possible value of h/h_0 , ranging from 0.25 to 1.5, in increments of 0.05, Table A.10 lists the calculated $CV(h)$ based on the training set. Fig. A.4 visualizes the trend of $CV(h)$ with respect to h/h_0 , which clearly shows a minimal point roughly at $h/h_0 = 0.65$. This indicates that the optimal bandwidth will be

$$h = 0.65h_0 = 0.65 \times 0.1385 = 0.0900.$$

Step 5: Calculating kernel weights (Line 8 in Algorithm 1) Using Eq. (7) and the optimized bandwidth h from the previous step, find the kernel weight w_{ij} between each testing loan i to each training loan j . The calculated weights are listed in Table A.11.

Step 6: Predicting return and risk (Lines 10 and 11 in Algorithm 1) Based on w_{ij} and R_j in historical data, find the predicted return (μ) and risk (σ^2), as listed in the 9th and 10th columns of Table A.7.

Step 7: Portfolio selection (Lines 13–19 in Algorithm 1) Given expected return $R^* = 0.08$, and the minimum investment $m = \$25$, based on the average rate of return μ and risk σ^2 , as listed in the 9th and 10th columns of Table A.7, find the weights of investments, as listed in the 11th column of Table A.7.

Given total amount for investment $M = \$5000$, the investment amounts include Loan 6 of $OM_6 = \$3250$, and Loan 9 of $OM_9 = \$1750$.

References

- Ait-Sahalia, Y., & Lo, A. W. (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, 94(1–2), 9–51.
- Andreeva, G., Ansell, J., & Crook, J. (2007). Modelling profitability using survival combination scores. *European Journal of Operational Research*, 183(3), 1537–1549.
- Belleflamme, P., Lambert, T., & Schwienbacher, A. (2014). Crowdfunding: Tapping the right crowd. *Journal of Business Venturing*, 29(5), 585–609.
- Berger, S., & Gleisner, F. (2009). Emergence of financial intermediaries in electronic markets: The case of online P2P lending. *Business Research*, 2(1), 39–65.
- Bugera, V., Konno, H., & Uryasev, S. (2002). Credit cards scoring with quadratic utility functions. *Journal of Multi-Criteria Decision Analysis*, 11(4–5), 197–211.
- Chatterjee, S., & Barcun, S. (1970). A nonparametric approach to credit screening. *Journal of the American Statistical Association*, 65(329), 150–154.
- Clark, R. M. (1975). A calibration curve for radiocarbon dates. *Antiquity*, 49(196), 251–266.
- Conlin, M. (1999). Peer group micro-lending programs in Canada and the United States. *Journal of Development Economics*, 60(1), 249–269.
- Desai, V. S., Conway, D. G., Crook, J. N., & Overstreet, G. A. (1997). Credit-scoring models in the credit-union environment using neural networks and genetic algorithms. *IMA Journal of Management Mathematics*, 8(4), 323.
- Feldman, D., & Gross, S. (2005). Mortgage default: Classification trees analysis. *The Journal of Real Estate Finance and Economics*, 30(4), 369–396.
- Fernandes, M. (2006). Financial crashes as endogenous jumps: Estimation, testing and forecasting. *Journal of Economic Dynamics and Control*, 30(1), 111–141.
- Frydman, H., Kallberg, J. G., & Kao, D. L. (1985). Testing the adequacy of Markov chain and Mover–Stay models as representations of credit behavior. *Operations Research*, 33(6), 1203–1214.
- Herzenstein, M., Dholakia, U. M., & Andrews, R. L. (2011). Strategic herding behavior in peer-to-peer loan auctions. *Journal of Interactive Marketing*, 25(1), 27–36.
- Huang, J. J., Tzeng, G. H., & Ong, C. S. (2006). Two-stage genetic programming (2SGP) for the credit scoring model. *Applied Mathematics and Computation*, 174(2), 1039–1053.
- Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, 37(4), 543–558.
- Hulme, M. K., & Wright, C. (2006). Internet based social lending: Past, present and future. *Social Futures Observatory*, 174(2), 1–115.
- Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research*, 201(3), 838–846.
- Klafft, M. (2008). Online peer-to-peer lending: A lenders' perspective. In *Proceedings of the international conference on E-learning, E-business, enterprise information systems, and E-government* (pp. 371–375).
- Kraft, H., & Steffensen, M. (2013). A dynamic programming approach to constrained portfolios. *European Journal of Operational Research*, 229(2), 453–461.
- Krumme, K., & Herrero-Lopez, S. (2009). Do lenders make optimal decisions in a peer-to-peer network? In *Proceedings of the IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technology* (pp. 124–127).
- Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, 39(1), 19–37.
- Malhotra, R., & Malhotra, D. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research*, 136(1), 190–211.
- Mancuso, A. J., Goodwin, B. K., & Grennes, T. J. (2003). Nonlinear aspects of capital market integration and real interest rate equalization. *International Review of Economics & Finance*, 12(3), 283–303.
- Mangasarian, O. L. (1965). Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13(3), 444–452.
- Markowitz, H. M. (1991). *Portfolio selection: Efficient diversification of investments* (2nd ed.). Wiley.
- Mollick, E. R. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing*, 29(1), 1–16.
- Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10(1), 186–190.
- Puro, L., Teich, J. E., Wallenius, H., & Wallenius, J. (2010). Borrower decision aid for people-to-people lending. *Decision Support Systems*, 49(1), 52–60.
- Puro, L., Teich, J. E., Wallenius, H., & Wallenius, J. (2011). Bidding strategies for real-life small loan auctions. *Decision Support Systems*, 51(1), 31–41.
- Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: A survey. *Operations Research*, 42(4), 589–613.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall/CRC.
- So, M. M., & Thomas, L. C. (2011). Modelling the profitability of credit cards by Markov decision processes. *European Journal of Operational Research*, 212(1), 123–130.
- Sonenstein, S., Herzenstein, M., & Dholakia, U. M. (2010). How accounts shape lending decisions through fostering perceived trustworthiness. *Organizational Behavior and Human Decision Processes*, 115(1), 69–84.
- Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, 50(2), 277–289.
- Thomas, L. C. (2009). *Consumer credit models: pricing, profit and portfolios*. USA: Oxford University Press.
- Wang, H., Greiner, M., & Aronson, J. E. (2009). People-to-people lending: the emerging e-commerce transformation of a financial market. *Value Creation in E-Business Management*, 36(13), 182–195.
- Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, 15(03), 757–770.
- Woodside-Oriakhi, M., Lucas, C., & Beasley, J. E. (2011). Heuristic algorithms for the cardinality constrained efficient frontier. *European Journal of Operational Research*, 213(3), 538–550.
- Wu, J., & Xu, Y. (2011). A decision support system for borrowers loan in P2P lending. *Journal of Computers*, 6(6), 1183–1190.
- Wu, T.-C., & Hsu, M.-F. (2012). Credit risk assessment and decision making by a fusion approach. *Knowledge-Based Systems*, 35, 102–110.
- Yang, B., Li, L. X., Ji, H., & Xu, J. (2001). An early warning system for loan risk assessment using artificial neural networks. *Knowledge-Based Systems*, 14(5–6), 303–306.
- Yang, Y. (2007). Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research*, 183, 1521–1536.
- Zhang, Z., Gao, G., & Shi, Y. (2014). Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. *European Journal of Operational Research*, 237, 335–348.