Portfolio Selections in P2P Lending: A Multi-Objective Perspective

Hongke Zhao[1] Qi Liu[1]* Guifeng Wang[1] Yong Ge[2] Enhong Chen[1]
[1]School of Computer Science and Technology, University of Science and Technology of China
{zhhk, wgf1109}@mail.ustc.edu.cn {qiliuql, cheneh}@ustc.edu.cn
[2]Eller College of Management, University of Arizona, ygestrive@gmail.com

ABSTRACT

P2P lending is an emerging wealth-management service for individuals, which allows lenders to directly bid and invest on the loans created by borrowers. In these platforms, lenders often pursue multiple objectives (e.g., *non-default probability, fully-funded probability and winning-bid probability*) when they select loans to invest. How to automatically assess loans from these objectives and help lenders select loan portfolios is a very important but challenging problem. To that end, in this paper, we present a holistic study on portfolio selections in P2P lending. Specifically, we first propose to adapt *gradient boosting decision tree*, which combines both *static features* and *dynamic features*, to assess loans from multiple objectives. Then, we propose two strategies, i.e., *weighted objective optimization strategy* and *multi-objective optimization strategy*, to select portfolios for lenders. For each lender, the first strategy attempts to provide one optimal portfolio while the second strategy attempts to provide a Pareto-optimal portfolio set. Further, we design two algorithms, namely $\mathcal{DPA}$ and $\mathcal{EVA}$, which can efficiently resolve the optimizations in these two strategies, respectively. Finally, extensive experiments on a large-scale real-world data set demonstrate the effectiveness of our solutions.

Keywords

P2P Lending; Portfolio Selection; Multi-objective Optimization; Dynamic Feature

1. INTRODUCTION

Recent years have witnessed the rapid development of online P2P lending platforms, e.g., Prosper[1], Lendingclub[2]. As a new emerging wealth-management service for individuals, P2P lending allows individuals to borrow and lend money directly from one to another without going through any traditional financial intermediaries. Indeed, P2P lending has become a fast growing investment market which attracts many users (i.e., borrowers and lenders) and generates massive lending transactions. For instance, the total loan issuance amount of Lendingclub had reached more than \$13.4 billion at the end of 2015.

The prevalence of P2P lending and the availability of transaction data have attracted many researchers' attentions, which mainly focused on risk evaluation [23, 11], social relation analysis [20, 13] and fully-funded analysis [28, 25]. Recently, authors in [34] proposed to study loan recommendations for lenders. However, due to the specific working mechanism of P2P lending, the problem of loan/investment recommendations in these platforms is still largely underexplored.

In P2P lending, there are mainly two kinds of roles: the *borrowers* who want to borrow money from others and the *lenders* who lend money to borrowers. Trading in these markets follows the *Dutch Auction Rule*[3] [17, 31]. Specifically, for borrowing money, a borrower will first create a listing to solicit bids from lenders by describing herself, the reason of lending (e.g., for wedding), the required amount (e.g., \$1,000) and the maximal interest rate (e.g., 10%). Then, if a lender wants to lend to this loan within its soliciting duration (e.g., one week), a bid is created by describing both how much money she wants to lend (e.g., \$50) and the minimum interest rate (e.g., 9.5%). If this listing receives more than its required amount in its soliciting duration, those bids with lower rates will succeed/win, and other bids with higher rates will be outbid/fail. In contrast, if this listing can't receive enough bids in time, it would be expired and all the previous bids would also fail [34, 6]. Based on this trading rule, a rational lender *Alice* may have the following two considerations while selecting loans to bid. **Multi-objective.** While selecting loans, *Alice* may evaluate a loan from the probability of this loan being fully funded, the probability of winning the bid, as well as the loan risk (i.e., default probability) [4]. **Portfolio.** To be a successful lender, *Alice* also has the *portfolio* [24] perspective in her mind, i.e., she usually wants to select more than one loan (i.e., a portfolio) to bid in each investment. Indeed, some platforms (e.g., Prosper) already instruct lenders to diversify their money on multiple loans to reduce risk. However, it is difficult and boring for lenders to select dozens of loans in each time. Thus, developing an automatic approach to recommend portfolios for lenders is very needed.

In this paper, we present a holistic approach to help P2P lenders select investment portfolios, which can satisfy lender-

*Corresponding author.

[1] https://www.prosper.com/

[2] https://www.lendingclub.com/

[3] There exists another kind of trading rule in P2P lending, in which the platform determines posted rates for loans [31]. This trading can be treated as a special case of our studied scenario.

KDD '16, August 13-17, 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-4232-2/16/08…\$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939861

Table 1: Mathematical notations.

Notation	Description
$VA = \{v_1, ..., v_{\|VA\|}\}$	the set of being-auctioned loans currently
$UA = \{u_1, ..., u_{\|UA\|}\}$	the set of current active lenders
$\boldsymbol{x} = (x_1, ..., x_{\|VA\|})$	a selected loan portfolio
Re_i^u	lender u_i's preference on rate expectation
Re_j^v	loan v_j's declared interest rate in auction
$\boldsymbol{P_j} = [R_j, T_j, C_j]$	loan v_j's assessed profile
$\boldsymbol{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, \alpha_{i3})$	lender u_i's personalized weighted vector

s' interest rate expectations, minimize investment risk (i.e., *default probabilities*) and maximize trading efficiency (i.e., *fully-funded probabilities and winning-bid probabilities*) simultaneously. Specifically, we first identify active lenders in current market, i.e., the lenders who are most likely to invest in the following period, as our target users. Second, we assess each being-auctioned loan from multi-objective views, i.e., the non-default probability, fully-funded probability and winning-bid probability. Here, different from previous works which only used *static features* in assessments, we also extract *dynamic features* of loans, and adapt an ensemble method (i.e., Gradient Boosting Decision Tree) to combine both static features and dynamic features to improve the prediction performances. Finally, given the identified active lenders and assessed loans, we attempt to select portfolios for each active lender. As we described above, the selection should take into account multiple economic factors/objectives, and the recommendation for each lender should also be portfolios rather than single loans. Specifically, we propose two strategies, i.e., *weighted objective optimization strategy* and *multi-objective optimization strategy* to solve portfolio selections. Weighted objective strategy combines three objectives into a single objective based on a weighted objective vector, and provides each lender with an optimal portfolio. Multi-objective optimization strategy optimizes three objectives simultaneously and gets a Pareto-optimal solution set (portfolios) for each lender. For these two strategies, two efficient algorithms, i.e., $\mathcal{DPA}$ (dynamic programming) and $\mathcal{EVA}$ (evolutionary algorithm), are designed to solve the optimization problems respectively. The contributions of this paper can be summarized as follows.

- To the best of our knowledge, this is the first work on assessing loans from a multi-objective perspective in P2P lending. Furthermore, we also propose to extract dynamic features from both bidding lenders and auctions, which are very helpful for loan assessments.
- With our two portfolio selection strategies, we attempt to recommendation portfolios rather than single loans for lenders. Especially in the multi-objective optimization strategy, for each lender, we get a Pareto-optimal (skyline) portfolio set. These distinguish our study from other recommendation works very much.
- We develop two algorithms, $\mathcal{DPA}$ and $\mathcal{EVA}$, which can select portfolios in two strategies effectively. Particularly, $\mathcal{EVA}$ optimizes for all lenders one by one with a special inherited initialization, which is more effective and efficient than conventional algorithms.
- We construct extensive experiments on a real-world data set. The experimental results clearly demonstrate the effectiveness of our solutions.

2. PRELIMINARIES AND ASSESSMENTS

In this section, we first introduce the preliminaries of portfolio selections. Then, we identify the active lenders and learn their preferences on rate expectation. Finally and most importantly, we introduce how to assess being-auctioned loans from multiple objectives. For better illustration, Table 1 lists some mathematical notations used in this paper.

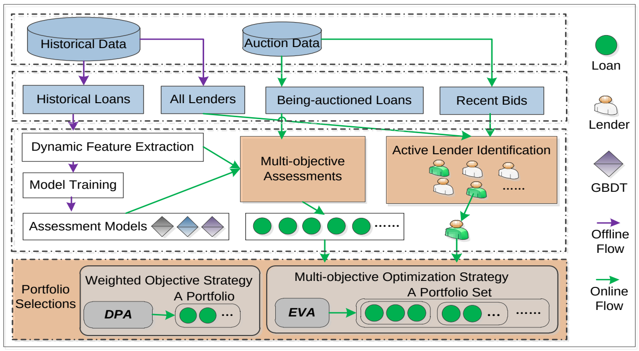


Figure 1: Flowchart of portfolio selections.

2.1 Preliminaries

Problem Statement. Formally, given the lenders' historical bidding records, and current being-auctioned loans $VA = \{v_1, ..., v_{|VA|}\}$ in market, our goal is to select loan portfolios from VA for each active lender. Active lenders $UA = \{u_1, ..., u_{|UA|}\}$ are those who are most likely to lend in the following period. A portfolio $\boldsymbol{x} = (x_1, ..., x_{|VA|})$ is an optimal combination of multiple loans, if $x_j = 1$, the *j-th* loan in VA is selected and put into the portfolio. For each lender $u_i \in UA$, the selected portfolios should satisfy her preference on rate expectation Re_i^u with minimum risk (maximal non-default probability) and maximum transaction efficiency (fully-funded probability, winning-bid probability).

Framework Overview. For tackling the above problem, we propose a solution framework which is show in Figure 1. There are three major steps (brown backgrounds): (1) identifying the active lenders UA and learning their preferences on rate expectation $Re_i^u, i \in \{1, ..., |UA|\}$; (2) assessing each being-auctioned loan $v_j \in VA$ on multiple objectives, (i.e., Non-default probability R_j, Fully-funded probability T_j, Winning-bid probability C_j); and (3) selecting portfolios for all active lenders.

We identify active lenders online (green arrows) and learn lenders' preferences on rate expectation from their historical investment records. For assessing being-auctioned loans, we train multiple assessment models using the historical loans offline (purple arrows). Given the identified active lenders and the assessed loans, we propose two strategies to select portfolios. Weighted objective optimization strategy provides an optimal portfolio and multi-objective optimization strategy provides a Pareto-optimal (skyline) portfolio set for each lender. In these two strategies, two algorithms, namely $\mathcal{DPA}$ and $\mathcal{EVA}$, are designed. This portfolio selection step is also achieved online. The first two steps will be introduced in the rest of this section, and the portfolio selections will be introduced in the next section.

2.2 Active Lenders and Rate Preferences

According to the data analysis and observation, we find that, in a certain period of time, only a small part of lenders (usually less than 10%) rather than all lenders will bid on current loans. We call these lenders who are most likely to bid in the following period *active lenders*. Identifying active lenders in advance can achieve accurate service of portfolio selection, also improve the service efficiency and user experience. Indeed, most lenders often bid periodically, and invest

Algorithm 1: $\mathcal{ALI}$: Active lenders identification.

```
Input: Lenders $U = \{u_1, ..., u_{|U|}\}$,
       Time threshold $TR$,
       Current time $CT$;
Output: Active Lenders $UA = \{u_1, ..., u_{|UA|}\}$,
        Rate expectation preference $Re_i^u$ of $u_i \in UA$;
Initialize: $UA = \phi$;
for each $u_i \in U$ do
    if $u_i$'s latest bid time in the range of $TR$ before $CT$ then
        $UA = UA \cup \{u_i\}$,
        Compute $Re_i^u$ according to Equation (1);
return $UA$ and $Re^u = \{Re_1^u, ..., Re_{|UA|}^u\}$.
```

in a series of loans continuously. Thus, the lenders who bid in recent days have higher probabilities to invest in the following period than other lenders. In our study, we select the lenders who have bidding behaviors in a time range TR (e.g., 10 days) before current time CT as our current target users, i.e., active lenders.

In investment, lender's preference is mainly reflected in her return expectation. In P2P lending, the interest rate of a loan declared before auction is often treated as the return [23, 34]. Thus, we can get a lender's preference on return rate through her historical investments. Specifically, lender u_i's preference Re_i^u on rate expectation can be calculated as:

$$Re_i^u = \frac{1}{|VU_i|} \sum_{v_j \in VU_i} Re_j^v, \tag{1}$$

where VU_i is the loan set that lender u_i invested in the past, Re_j^v is the declared interest rate of loan v_j. Algorithm 1 shows the process of identifying active lenders and the calculation of their preferences on rate expectation.

2.3 Multi-objective Assessments for Loans

In this subsection, we show the way of assessing the profiles of being-auctioned loans on multiple objectives. According to the previous study [4], there are three major economic factors that lenders take into account when deciding to bid on a loan: lenders' belief about the probability of a loan being fully funded, the probability of winning the bid, as well as the interest rate. Similarly, we also formalize the loan assessment from a multi-objective view. In our study, we assess a loan v_j on the following three objectives.

Non-default Probability(R_j). Since the loan interest rate Re_j^v is given explicitly in market and often taken as a searching criterion by many lenders, we take the rate expectation as one important constraint, and formalize another widely-used assessment metric: Risk, i.e., default probability [23, 19, 34]. For consistency, we maximize loans' non-default probabilities rather than minimizing default probabilities on risk assessment. Formally, non-default probability is the estimated probability that one loan may repay the principal and interest to lenders in time.

Fully-funded Probability(T_j). Fully-funded probability is the estimated probability that one loan may receive enough bids in its auction. According to the trading rule, the Transactions are valid only if the corresponding loan can receive enough bids. Thus, fully-funded probability is another important aspect to assess loans [14, 4].

Winning-bid Probability[4](C_j). Winning-bid probability is the estimated probability that a lender's bid on this loan under current loan status will finally success or participate after the loan's entire auction. Since some popular loans may receive more bids than their required amount, Competition among bids on a loan will take place and some bids will fail. Thus, winning-bid probability reflects the biding competition on a loan [4, 18].

Table 2: Feature examples.

Name	Description	Class
Bor_Rat	the maximum interest rate the borrower is willing to pay	Loan (*static*)
Category	borrowing purpose	
Cre_Dat	the created date of the listing of this loan	
...	...	
Deb_Inc	debt to income ratio of the borrower	Borrower (*static*)
Credit	credit grade of the borrower	
...	...	
Am_Rem	the amount which remains to be funded	Auction (*dynamic*)
Auc_Pro	time interval from Cre_Date to current CT	
...	...	
Def_Per	lenders' past default loan percent	Lender (*dynamic*)
Fun_Per	lenders' past fully-funded loan percent	
...	...	

Non-default probability R_j mainly affects lenders' profits, while fully-funded probability T_j and winning-bid probability C_j mainly affect the investment efficiency, i.e., helping avoid invalid bids. In summary, for each being-auctioned loan v_j, we can adopt a three-element vector $\boldsymbol{P_j} = [R_j, T_j, C_j]$ to denote its profile. Next, we will introduce how to estimate these profile terms on specific objectives for given loans.

2.3.1 Features for Assessments

Here, we introduce extracting features for loan assessments. In previous works, researchers explored some classification or regression models on some single objective assessment tasks, e.g., default [29, 9, 34], fully-funded [14, 28]. However, these studies only explored the *static features* from both loan (e.g., *rate, amount, purpose,*) and borrower (e.g., *credit, debt*). In [23], authors evaluated a loan by the characteristics extracted from the bidding lenders, i.e., average and variance of the past real returns of bidding lenders. Indeed, a loan receives bids from lenders one by one during its auction. Thus, the features extracted from the lenders who have bid on this loan are dynamic. Besides, we can also extract some dynamic features from loan auctions, e.g., *auction phase/time, percent fund.* These dynamic features in auction reflect the popularity of a loan directly. In summary, in this paper, we explore both static features and dynamic features for better multi-objective loan assessments.

Static Features. Static features are given explicitly before auction, which are extracted from the loan's properties and the associated borrower's properties. These static features directly reflect the basic properties of loans and borrowers. For example, borrowers with high credits are more likely to repay in time than borrowers with poor credits.

Dynamic Features. Dynamic features are extracted from the temporal auction and incremental bidding lenders of a loan, which are changing from time to time during an auction. Dynamic features can reflect the popularity of a loan. For example, popular loans may receive massive bids rapidly, thus, are more likely to be fully funded. Further, the dynamic features extracted from lenders reflect the lenders' views, such that, loans received many bids from experienced lenders (lenders whose investments success frequently and default rarely) may be better than other loans. Especially, dynamic features, *Def_Per, Fun_Per, Win_Per,* extracted from the incremental bidding lenders' historical investments in line with our three assessment objectives, will have great help to the prediction performances on corresponding objectives.

We represent all the features as numerics. For temporal

[4] This objective doesn't exist on the loans with posted rates [31].

features, such as *Cre_Dat*, we convert raw features to a serial date number, which represents the whole and fractional number of days from a fixed preset date [6]. For categorical features, such as *Category* and *Credit*, we convert a variable with n categories into a n-dimensional binary vector, in which only the value in the corresponding category is set to one. All features are normalized for comparability. For example, *Def_Per* represents the fraction of all default loans of bidding lenders to their total bidding number in the past. Please note that, for non-default objective and fully-funded objective, we extract dynamic features every day, while for winning-bid objective, we extract dynamic features every bid (i.e., the features for winning-bid objective are constructed according to the latest bid on this loan at current). Thus, the training instance numbers are different on different objectives but the feature dimensions are same on all objectives. In summary, we use a total of 26 features (i.e., 9 loan features, 4 borrower features, 7 auction features and 6 lender features) in our study and some examples of these features are shown in Table 2.

2.3.2 Loan Assessments

After the feature extraction and preparation, in this part, we introduce the assessment models. Specifically, we adopt the gradient boosting decision tree ($\mathcal{GBDT}$) [10, 12] to assess loans, which has been used to predict P2P lending transaction in [6]. The reason for choosing $\mathcal{GBDT}$ is as follows. First, we should evaluate the being-auctioned loans with numerical or ranking output which is the foundation in the following portfolio selection formalization. $\mathcal{GBDT}$ can estimate the probabilities (non-default probability, fully-funded probability and winning-bid probability) for each loan. More importantly, $\mathcal{GBDT}$ is an ensemble method where an individual learner is a decision tree which only uses one variable at each node when it is trained/constructed as well as when it is applied to test data. This characteristic prevents us from worrying about heterogeneity in the features we generated [6]. What's more, compared with conventional machine learning models, the performances of ensemble methods are significantly better and well demonstrated [10, 32, 15, 1].

Suppose we have n training instances $\{(z_1, y_1), ..., (z_n, y_n)\}$ for a specific task (e.g., non-default objective), where z_i is the feature vector of loan instance v_i, and y_i is the label of v_i on this objective (e.g., if v_i repay in time, y_i=1; otherwise, y_i=0) . We train $\mathcal{GBDT}$ ($G(z)$) with M weak learners, each is a decision tree $h(z)$ with a weight coefficient γ_m,

$$G(z) = \sum_{m=1}^{M} \gamma_m h_m(z). \tag{2}$$

Similar to other boosting algorithms, $\mathcal{GBDT}$ builds the additive model in a forward stage-wise fashion. At each stage the decision tree $h_m(z)$ is chosen to minimize the loss function $\mathcal{L}$ given the current model G_{m-1} and its fit $G_{m-1}(z_i)$.

$$\begin{aligned} G_m(z) =& G_{m-1}(z) + \gamma_m h_m(z), \\ =& G_{m-1}(z) + \arg\min_{h} \sum_{i=1}^{n} L(y_i, G_{m-1}(z_i) - h(z_i)). \end{aligned} \tag{3}$$

$\mathcal{GBDT}$ attempts to solve this above minimization problem numerically via steepest descent, whose direction is the negative gradient of the loss function evaluated at the current model G_{m-1},

$$G_m(z) = G_{m-1}(z) + \gamma_m \sum_{i=1}^{n} \bigtriangledown_G L(y_i, G_{m-1}(z_i)), \tag{4}$$

the weight coefficients γ_m is calculated by:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, G_{m-1}(z_i) - \gamma \frac{\partial L(y_i, G_{m-1}(z_i))}{\partial G_{m-1}(z_i)}). \tag{5}$$

Repeat this building trees process until $h_M(z)$, and then $\widehat{G(z)} = G_M(z)$. After we get the models $\mathcal{GBDT}^i$, $i \in \{1, 2, 3\}$, on three objectives. For each being-auctioned loan v_j, we can get its estimations on three objectives, i.e., the terms in profile $\boldsymbol{P_j}$ by the model output probabilities on positive classes of these objectives.

$$\begin{aligned} VA, \{v_1, ..., v_{|VA|}\} \xrightarrow[i \in \{1,2,3\}]{\mathcal{GBDT}^i} & \{(R_1, T_1, C_1), ..., \\ & (R_{|VA|}, T_{|VA|}, C_{|VA|})\}. \end{aligned} \tag{6}$$

For the scale consistency of different objectives, we reprocess the values in each dimension of $\boldsymbol{P_j}$ by their relative values, e.g., inverse ranking values in all the being-auctioned loans. For example, $\boldsymbol{P_j} = [101, 111, 51]$ means loan v_j outperforms other 100, 110, 50 loans on three objectives respectively. Now, we assessed the being-auctioned loans on three objectives, in the following, we will make our main effort to help active lenders select portfolios.

3. PORTFOLIO SELECTIONS

In this section, we introduce the detailed information of portfolio selections. Generally, portfolio is a famous theory in finance that attempts to maximize portfolio expected return for a given amount of portfolio risk, or equivalently minimize risk for a given level of expected return, by carefully choosing the proportions of various assets [24]. Similarly, in our study, we maximize the formalized multiple objectives of portfolio for a given level of expected return (Re_i^u). In P2P lending, for a given lender, her bidding amount on specific loans are not significantly different, which are often the smallest allowed amount (i.e., $25) or integer multiple of smallest amount (e.g., $50). Thus, when selecting portfolios, we don't care about the investment amount of a lender and we mainly focus on solving a discrete selection optimization problem. Meanwhile, we also assume "lenders are experienced and rational", which is one of the most fundamental assumptions in economics [24, 26].

Specifically, given the active lenders UA with their rate expectation preferences $Re_i^u, i \in \{1, ..., |UA|\}$ and the profiles of being-auctioned loans i.e., $\boldsymbol{P_j}, j \in \{1, ..., |VA|\}$, we propose two strategies, i.e., weighted objective optimization strategy and multi-objective optimization strategy to help lenders select loan portfolios. The first strategy combines the three objectives into one single objective via a predefined weighted vector and provides one optimal portfolio for each lender through a dynamic programming algorithm that we designed, i.e., $\mathcal{DPA}$. Further, the multi-objective optimization strategy optimizes all objectives simultaneously, and provides all the Pareto-optimal portfolios for each lender through a novel evolutionary algorithm, i.e., $\mathcal{EVA}$.

3.1 Weighted Objective Strategy

Weighted objective strategy supposes there is a weight vector $\boldsymbol{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, \alpha_{i3})$ from lender u_i to balance the importance of three objectives. This setting is practical since some famous platforms, e.g., Prosper or Lendingclub allows users to preset some criteria, e.g., their expected interest rates, before searching. The weight vector can also be integrated into the user profiles.

Through the multi-objective assessments in section 2.3, we get the profile terms of each loan, i.e., $\boldsymbol{P_j}$. Thus, for lender u_i, we can obtain a single weighted objective on all being-auctioned loans VA: $f(\boldsymbol{x}) = \sum_{j=1}^{|VA|} \boldsymbol{\alpha_i^T} \boldsymbol{P_j} x_j$, where x_j is the boolean selecting label of loan v_j. In this case, the portfolio selection problem for lender u_i can be formalized as:

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = \sum_{j=1}^{|VA|} \boldsymbol{\alpha_i^T} \boldsymbol{P_j} x_j, \\ s.t. \quad & \frac{1}{||\boldsymbol{x}||_0} \sum_{j=1}^{|VA|} Re_j^v x_j \geq Re_i^u, \\ & K \geq ||\boldsymbol{x}||_0, \end{aligned} \tag{7}$$

where $\boldsymbol{x} = (x_1, ..., x_{|VA|})$ is the selection vector and x_j=1 means the *j-th* loan in VA is selected and put into u_i's portfolio. K is the portfolio size which is given by lender u_i[5]. The objective of this function is to maximize the weighted single objective, i.e., $f(\boldsymbol{x})$. The first constraint means that the selected portfolio should satisfy the lender u_i's preference on rate expectation Re_i^u. This formalization is a discrete constrained optimization problem, which is NP-hard and will need $\sum_{i=1}^{K} \binom{|VA|}{i}$ computations in the worst case.

This problem is difficult to solve directly by conventional algorithms since there are two constraints. In this study, we design a dynamic programming algorithm, namely $\mathcal{DPA}$, with two main loops to obtain the optimal solution. First, we convert the declared rate value Re_j^v of each loan into a positive integer representation, e.g., 10.50% → 1050 (Re_j^v is a percentage with two decimal places in our data), and denote the largest rate value as $Rmax$, (i.e., $Rmax = \max_{j \in \{1,...,|VA|\}} Re_j^v$). The weighted objective value of each loan v_j is donated as Pr_j (i.e., $Pr_j = \boldsymbol{\alpha_i^T} \boldsymbol{P_j}$). $\mathcal{DPA}$ is shown in Algorithm 2, in which $D[p][q]$ denotes weighted objective of the portfolio with p loans and q summation of rate.

In $\mathcal{DPA}$, the first loop only considers the second constraint in problem (7) and can get all the optimal selections under any p and q, and the computational complexity is $O(\frac{1}{2}|VA|K^2 Rmax)$. Further, the second loop gets the final optimal solution by satisfying the first constraint, and the computational complexity is $O(K^2 Rmax)$. Please note that, only the second loop takes the lender's rate expectation into consideration, thus the optimal selections provided by the first loop are the same for all lenders. If we want to get the final solutions for all lenders, we need to compute for all of them in the second loop, and thus, the overall computational complexity is $O(\max\{\frac{1}{2}|VA|K^2 Rmax, |UA|K^2 Rmax\})$.

Through $\mathcal{DPA}$, we can get an exact optimal portfolio for each lender. However, in many cases, for most lenders, they are ambiguous or not sensitive about the weights on the multiple objectives. Thus, we propose another strategy, multi-objective optimization, to solve portfolio selections.

Algorithm 2: $\mathcal{DPA}$: Dynamic programming algorithm.

```
Input: A given lender $u_i$, with her rate expectation $Re_i^u$,
       All being-auctioned loans $VA$, and each loan $v_j \in VA$,
       with its declared rate $Re_j^v$ and objective profit $Pr_j$;
Output: Portfolio selection vector $\boldsymbol{x}$,
        Maximum objective profit $f(\boldsymbol{x})$;
Initialize: $\boldsymbol{x}, x_j = 0, j = 1, ..., |VA|$;
            $D[0][0] = 0$, other $D[p][q] = -1$;
            All $G[p][q] = \phi$;
for $j$ = 1 to $|VA|$ do
    for $p$ = K-1 down to 0 do
        for $q$ = p*Rmax down to 0 do
            if $(D[p][q] \geq 0) \wedge (D[p][q] + Pr_j > D[p+1][q + Re_j^v])$
            then
                $D[p+1][q + Re_j^v] = D[p][q] + Pr_j$,
                $G[p+1][q + Re_j^v] = G[p][q] \cup \{j\}$;
maxD=0,maxp=maxq=0;
for $p$ = 1 to K do
    for $q$ = 0 to K*Rmax do
        if $q/p \geq Re_i^u \wedge D[p][q] > maxD$ then
            maxD = D[p][q],
            maxp=p,  maxq=q;
for $j \in G[maxp][maxq]$ do
    $x_j = 1$;
return $\boldsymbol{x}$, and $f(\boldsymbol{x}) = maxD$.
```

3.2 Multi-objective Optimization Strategy

Different from the weighted objective optimization strategy, multi-objective optimization strategy does not need a weight vector and it optimizes the multiple objectives simultaneously and obtains a Pareto-optimal portfolio set instead of a single portfolio for each lender. In this strategy, the three objective functions on being-auctioned loans VA can be respectively defined by their corresponding profile terms, i.e., $\boldsymbol{P_j} = [R_j, T_j, C_j], j \in \{1, ..., |VA|\}$. Thus,

$$f_1(\boldsymbol{x}) = \sum_{j=1}^{|VA|} R_j x_j, \quad f_2(\boldsymbol{x}) = \sum_{j=1}^{|VA|} T_j x_j, \quad f_3(\boldsymbol{x}) = \sum_{j=1}^{|VA|} C_j x_j. \tag{8}$$

Pareto-optimal portfolio set (A^*). Suppose $\boldsymbol{x}$ and $\boldsymbol{x'}$ are two feasible solutions/portfolios in the solution space Ω. $\boldsymbol{x}$ is said to dominate $\boldsymbol{x'}$ (denoted as $\boldsymbol{x} \prec \boldsymbol{x'}$) if and only if $\forall e \in \{1,2,3\}$, $f_e(\boldsymbol{x}) \geq f_e(\boldsymbol{x'})$ and $\exists e \in \{1,2,3\}$, $f_e(\boldsymbol{x}) > f_e(\boldsymbol{x'})$. Thus, a solution $\boldsymbol{x*} \in \Omega$ is Pareto-optimal if there is no other $\boldsymbol{x} \in \Omega$ dominates $\boldsymbol{x*}$, i.e., $\neg \exists \boldsymbol{x} \in \Omega, \boldsymbol{x} \prec \boldsymbol{x*}$. The set A^* of all Pareto-optimal solutions/portfolios in the decision space is called Pareto-optimal or skyline portfolio set, and each of those solutions can be treated as a specific trade-off among these contradictory objectives.

In fact, Pareto optimality is a state of allocation or selection in which it is impossible to make any objective better off without making at least one objective worse off. Each selection portfolio in Pareto-optimal set is assessed under multiple objectives and no other option can categorically outperform any of the members in this Pareto-optimal set. In the multi-objective optimization strategy, we try to get the Pareto-optimal portfolio set as the recommendation candidates to each lender. The multi-objective optimization problem for lender u_i can be formalized as follows:

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & \boldsymbol{\mathcal{F}(x)} = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), f_3(\boldsymbol{x})), \\ s.t. \quad & \frac{1}{||\boldsymbol{x}||_0} \sum_{j=1}^{|VA|} Re_j^v x_j \geq Re_i^u, \\ & K \geq ||\boldsymbol{x}||_0. \end{aligned} \tag{9}$$

This formalized problem is much more complicated than the weighted optimization problem (7). Getting the exact Pareto-optimal solution set for this problem is unpractical. Inspired by previous studies on multi-objective optimization [7, 8, 33], evolutionary algorithms are well-suited to solve multi-objective optimization problems because their inherent parallelism allows them to find a set of Pareto-optimal solutions in a single run. Being population-based stochastic search approaches, evolutionary algorithms use concepts

[5] Lenders may have different portfolio size K. In practice, we group lenders based on their K values. At each time, $\mathcal{DPA}$ precess all lenders in a group together with a same K, which is the same in $\mathcal{EVA}$.

of population and of recombination inspired by Darwinian evolution. An iterative process is executed, initialized by a randomly chosen population (portfolios) which is also called chromosome (i.e., $\boldsymbol{x}$ in our study, which also conforms binary encoding). In each generation, new solutions or offsprings are generated through recombination and mutation. Besides, selection operation in each generation leads the evolutionary process according to the defined fitness function. For multi-objective problems, fitness functions are often designed based on Pareto domination, e.g., [8, 16]. Evolutionary algorithms are often used to solve many complicated problems with providing approximate solutions efficiently.

However, the conventional multi-objective evolutionary algorithms, e.g., NSGA-II [8], MOEA/D [33], were designed for one multi-objective optimization problem rather than a series of similar problems, e.g., portfolio selections for all lenders in our study. Thus, the conventional algorithms have to run $|UA|$ times independently for all active lenders in our problem which will take much computation cost. In fact, for many lenders, their selection problems are similar, and the only difference is the rate expectation preference Re_i^u in the second constraint. Especially for some lenders with similar preferences, their final optimal selections may also be very similar. Thus, there is no need to reprocess from beginning for all lenders. In this study, we propose a new evolutionary algorithm, namely $\mathcal{EVA}$, to solve these series of optimizations for all active lenders more effectively.

$\mathcal{EVA}$ first ranks all active lenders in descending order according to their rate expectation preferences Re_i^u. Then, we have the following theorem:

Theorem 1. *If A_i^* is the Pareto-optimal solution set for lender u_i, each solution $\boldsymbol{x} \in A_i^*$ must be a feasible solution (satisfy constraints) for each lender $u_{i'}, i' \in \{i+1, ..., |UA|\}$, and A_i^* is a good approximation of the Pareto-optimal solution set of u_i's next neighbor lender, i.e., u_{i+1}.*

Considering Theorem 1, we can simplify and speed up the convergence processes for lenders $u_{i+1}, i \in \{1, ..., |UA|\}$ by taking u_i's Pareto-optimal set as initialization. Besides, $\mathcal{EVA}$ adopts a random greedy strategy to repair the infeasible individual solutions during evolution. Suppose $\boldsymbol{x}$ is an infeasible individual, (i.e., $\frac{1}{||\boldsymbol{x}||_0} \sum_{j=1}^{|VA|} Re_j^v x_j < Re_i^u$, or $K < ||\boldsymbol{x}||_0$), in each repair, determine an objective $e \in \{1, 2, 3\}$ randomly, and select $j' \in \{1, ..., |VA|\}$ such that:

$$j' = \arg \min_{j \in \{1,...,|VA|\}} (f_e(\boldsymbol{x}) - f_e(\boldsymbol{x}^{j-}))Re_j^v, \qquad (10)$$

where $\boldsymbol{x}^{j-}$ is different from $\boldsymbol{x}$ only in the position j, and $x_j^{j-} = 0$. Set $x_{j'} = 0$, and repeat repairing until $\boldsymbol{x}$ is feasible. What's more, $\mathcal{EVA}$ selects population in each generation according to their dominations, which is same with [8, 16]. The whole process of $\mathcal{EVA}$ is shown in Algorithm 3.

In each iteration, the computational complication is $O(N^2)$ (N is the population size) since $\mathcal{EVA}$ adopts the same selection operation with that in [8]. Thus, the whole computational complication is $O(|UA|GN^2)$ (G is the iteration generations). In fact, the iteration times for lenders $u_{i+1}, i \in \{1, ..., |UA|\}$ are much less than that for u_1 because of the special initialization approach in $\mathcal{EVA}$.

Each solution in the Pareto-optimal set is the best under a certain tradeoff of multiple objectives. After getting the Pareto-optimal portfolio set, a lender can select one portfolio from this set autonomously or randomly.

Algorithm 3: $\mathcal{EVA}$: Evolutionary algorithm.

```
Input: Active lenders $UA = \{u_1, ..., u_{|UA|}\}$,
       their preferences on rate expectation $\{Re_1^u, ..., Re_{|UA|}^u\}$;
       All being-auctioned loans $VA$; and each $v_j \in VA$,
       with its declared rate $Re_j^v$ and assessed profile $\boldsymbol{P}_j$;
Output: Pareto-optimal portfolio selections
        $A_i^* = \{\boldsymbol{x}^1, ..., \boldsymbol{x}^i, ...\}$ for lender $u_i$, $i \in \{1, ..., |UA|\}$;
Initialize: Rank $UA$ according to $Re_i^u$ in descending order;
for $i$ from 1 to $|UA|$ do
    if $i$==1 then
        Initialize $N$ solutions $A_1 = \{\boldsymbol{x}^1, ..., \boldsymbol{x}^N\}$ randomly;
        Go to Evolution;
    else
        Initialize $N - |A_{i-1}|$ solutions $\{\boldsymbol{x}^1, ..., \boldsymbol{x}^{N-|A_{i-1}|}\}$
        randomly, $A_i = A_{i-1}^* \cup \{\boldsymbol{x}^1, ..., \boldsymbol{x}^{N-|A_{i-1}|}\}$;
    Evolution:
    for $\boldsymbol{x} \in A_i$ do
        Repair $\boldsymbol{x}$ according to Equation (10);
    Reproduction:
    for $j$ from 1 to $N$ do
        Select two individuals randomly from $A_i$;
        Generate one offspring $\boldsymbol{x}$, repair $\boldsymbol{x}$, and $A_i = A_i \cup \{\boldsymbol{x}\}$;
    Select $N$ $\boldsymbol{x}$ from $A_i$ according to their dominations;
    Stopping Criterion:
    If stopping criterion is not satisfied, go to Reproduction;
    Post-processing:
    Remove the dominated solutions from $A_i$, and get $A_i^*$;
    return $A_i^*$.
```

In summary, for portfolio selections, we first identify the active lenders in market and get their preferences on rate expectation, and meanwhile, assess the being-auctioned loans on multiple objectives. Further, we propose two strategies to help active lenders select portfolios. Weighted objective strategy works efficiently and provides an optimal portfolio for each lender, which depends on a weighted vector from lenders. Multi-objective optimization strategy can automatically provide an approximate Pareto-optimal portfolio set for each lender. From the formalizations of these two strategies, we can see that both of them can easily expand to other or more objectives. We will evaluate the advantages and disadvantages of these two strategies in experiments.

4. EXPERIMENTS

In this section, we will construct extensive experiments on a large-scale real-world data set. First, we make data analysis and explore the lenders' bidding behaviors from multi-objective and portfolio perspectives. Second, we evaluate the performances of multi-objective assessments. Finally, we evaluate our portfolio selections holistically.

4.1 Experimental Data and Analysis

The experimental data set is collected from one famous P2P lending platform in America, i.e., Prosper[6]. This data contains the transaction records in this platform of almost 6 years. We mainly use three tables of this data set. *Listing table* contains the loan temporal status on auction and some basic credit features of borrowers. This table is mainly used to extract static features. *Bid table* contains the specific time and some information of bid, e.g., bid amount of money, and the bidding result of each lender on a certain loan. These investment records are the basis to construct dynamic features. *Loan table* is used to evaluate the performances of a loan, e.g., default and fully-funded.

We partition the data into five groups, and take four groups as training data and the remaining one group as test data.

[6] https://www.prosper.com/tools/DataExport.aspx

Table 3: Experimental data statistics.

#Loan	#Lender	$\#Tr_1$	$\#Tr_2$	$\#Tr_3$	#VA	#UA
387,848	62,782	19,311	310,159	5,355,873	2,233	4,800

In experiments, we adopt five-fold cross-validation and all results are the average of five test rounds. The number of training instances for $\mathcal{GBDT}_i$ is denoted as Tr_i, and the number of test instances (being-auctioned loans) is denoted as Te (i.e., VA). In the test process or portfolio selections, we select portfolios for lenders every a certain period, e.g., two weeks, instead of every day. This is because loans being auctioned on neighbor days are almost completely overlapped. Table 3 shows the basic statistics of the experimental data.

4.1.1 *Analysis from Multi-objective Perspective*

In this part, we analyze the lenders' bidding behaviors from the multi-objective perspective. Figure 2(a) shows the lender distribution on risk, where the X-axis represents the non-default loan percent in one lender's past investments and Y-axis represents the lender percent. The blue bars are raw results and red line is fitting curve by Gauss distribution. Figure 2(b) shows the lender distribution on fully-funded objective, in which the lender distribution is not uniform. Figure 2(c) shows the lender distribution on winning-bid objective. We can see that, only small part of lenders' bids perform well. In other words, many lenders often suffer from risk and investment failure. Our goal is to help lenders select loan portfolios with higher probabilities on all objectives. Figure 2(d) shows the lender distribution on lenders' rate expectations (Re_i^u) which is treated as a personalized restrict in portfolio selections. We can see that most lenders prefer medium and similar rates, e.g., 0.1-0.15, rather than highest rates. This characteristic of similar rate expectation is taken used when designing $\mathcal{DPA}$ and $\mathcal{EVA}$.

4.1.2 *Analysis from Portfolio Perspective*

In this part, we analyze lenders' bidding behaviors from portfolio perspective. We randomly select 10 lenders with more than 100 bidding records. Firstly, we obtain individual loans for each lender, besides, we partition each lender's bidding loans into different portfolios based on the time intervals between two neighbor bids, i.e., if the interval of two neighbor bids is more than 30 days, we partition them into different portfolios. Thus, we get the loan portfolios for each lender, and we compute the average rate of loans in one portfolio as this portfolio's rate. Figure 4 shows the rates of lenders using box plot, in which the blue boxes represent the rates of single loans and green boxes represent the rates of portfolios of lenders. We can see that, different lenders have different rate preferences; furthermore, for a specific lender, portfolio rates are much more stable and focused than single loan rates. In other word, rate expectation constraint or preference based on portfolio is more reasonable than that based on single loans. In fact, in an investment, a lender may bid several loans whose rates may be quite different, but the portfolio rate or average rate of these loans is stable and personalized. This finding demonstrates the rationality of portfolio recommendation in our study rather than conventional single loan recommendation.

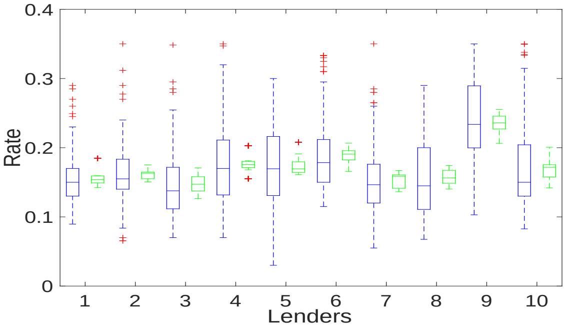


Figure 4: Rate v.s. single loan and portfolio.

4.1.3 *Result of Active Lender Identification*

In this part, we report the result of identifying active lenders. Figure 5 shows the statistical result of active lender identification. We select the lenders who have lending behaviors in the previous cumulative days (i.e., TR) before current time CT in each test round, and get their bidding results in the following auctions. We can see that, identified small part of lenders (blue bars) contribute the most bids (red line), e.g., 6% lenders contribute 85% bids. Thus, in portfolio selections, we mainly take the lenders who have bids in previous 10 days before current time CT as the current active lenders.

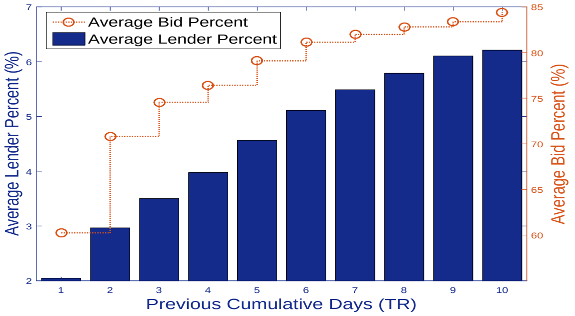


Figure 5: Lender/bid percent v.s. TR.

4.2 Evaluations of Loan Assessments

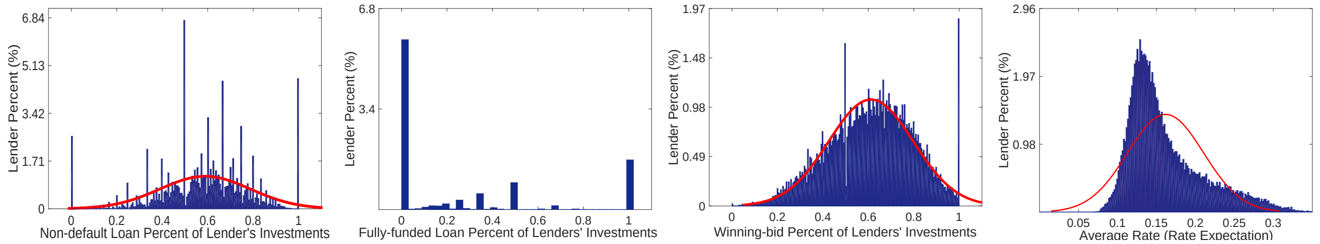
In this subsection, we evaluate the performances of multi-objective assessments for loans. Specifically, we evaluate both the utilities of dynamic features and the effectiveness of $\mathcal{GBDT}$ models on different objectives.

4.2.1 *Setups and Baselines*

We denote the $\mathcal{GBDT}$ models using all features as GBDT, and $\mathcal{GBDT}$ models only using static features as GBDT_S. We also adopt two comparison models, e.g., Logistic Regression and Decision Tree, which have been used in previous studies on P2P lending [9, 34, 22]. Similarly, we denote the comparison models as LR, LR_S, DT, DT_S respectively. All these models get their best performances and parameters through training and validation processes. We adopt two widely-used metrics, i.e., ROC curve and AUC [2] (area under ROC curve) to evaluate the assessment results.

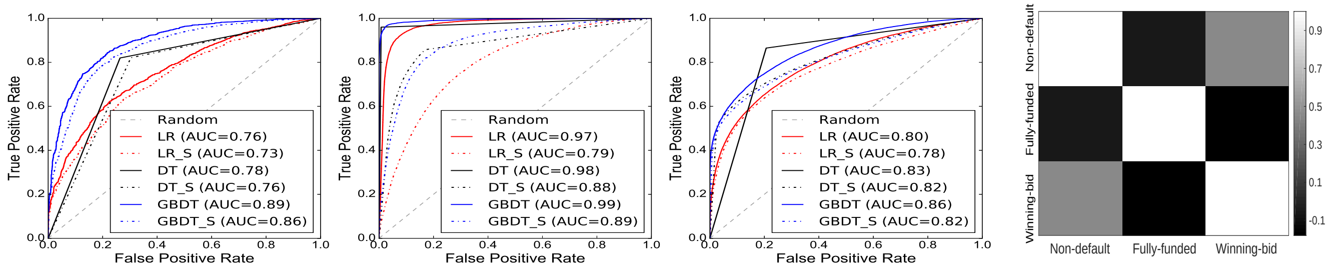
4.2.2 *Assessment Results*

The assessment results are shown in Figure 3. We can see that, on three objectives, extracted dynamic features can improve the performances of all models significantly, especially on the fully-funded objective prediction, i.e., about 10% improvements on AUC of all models. On the other hand, $\mathcal{GBDT}$ models, i.e., GBDT and GBDT_S, outperform other models on all three assessment tasks, especially on the non-default objective with more than 10% improvements. Thus, in portfolio selections, we adopt $\mathcal{GBDT}$ with all features to assess the being-auctioned loans.



(a) (b) (c) (d)

Figure 2: Lender distributions/percents v.s. different objectives.



(a) Non-default assessments (b) Fully-funded assessments (c) Winning-bid assessments (d) Correlations

Figure 3: Prediction performances of assessments.

Besides, we also show the Pearson correlations of assessed loans' profile values on different objectives in Figure 3(d). This figure tells us that the correlations between loans' estimated profile values on different objectives are not significant (i.e., values are less than 0.5). Thus, we can't get the optimal loans on all objectives by recommending or selecting loans through conventional one-objective techniques.

4.3 Evaluations of Portfolio Selections

In this subsection, we evaluate our portfolio selection approaches holistically.

4.3.1 *Experimental Setups*

In $\mathcal{EVA}$, the population size is set as 100; the generations are set as 150 for first lender and 30 for other lenders; the crossover rate is set as 0.5; and the mutation rate is set as $50K/|VA|$. The portfolio size K is given by each lender. In selections, actually, we can remove some individual loans in advance that are dominated by many (i.e., 50) other loans before selecting, since they are almost impossible to be chosen into any portfolios.

Baseline Methods. To the best of our knowledge, there are not relevant works on portfolio selections from multiple objectives in P2P lending or other domains. Thus, in this experiment, we compare $\mathcal{DPA}$, $\mathcal{EVA}$ and their variants. Specifically, we set the comparison methods as follows.

$\mathcal{DPA}$ **series**. For $\mathcal{DPA}$, we adopt multiple different parameters $\boldsymbol{\alpha}$, (i.e., (0.5, 0.3, 0.2),(0.3, 0.4, 0.3),(0.3, 0.5, 0.2),(0.2, 0.6, 0.2)), and the corresponding algorithms are denoted as DPA_1, DPA_2, DPA_3, DPA_4.

$\mathcal{EVA}$ **series**. Since $\mathcal{EVA}$ provides a portfolio set rather than a single portfolio for each lender, for comparison, we respectively select the portfolios with maximum non-default objective, fully-funded objective and winning-bid objective from the Pareto-optimal candidate set as the final selections. The corresponding methods are denoted as EVA_1, EVA_2 and EVA_3. Besides, we can also select a portfolio randomly from the Pareto-optimal portfolio set as the final result, and this method is denoted as EVA_R.

$\mathcal{EVAR}$ **series**. As a variant of $\mathcal{EVA}$, $\mathcal{EVAR}$ is a conventional evolutionary algorithm similar to $\mathcal{EVA}$, which always adopts random initializations and needs 150 generations for all lenders. The other setups in $\mathcal{EVAR}$ are the same with these in $\mathcal{EVA}$. Similar to $\mathcal{EVA}$ series, we also get $EVAR_1$, $EVAR_2$, $EVAR_3$, and $EVAR_R$.

$SELF$. We also get the results selected by lenders themselves or recommending by criteria match, e.g., rate match, in Prosper, which are denoted as $SELF$.

Metrics. We adopt the real percents of loans with positive labels on different objectives (Non-default, Fully-funded, Winning-bid) in the selected portfolio as our evaluation metrics. Besides, we simulate the bidding and repayment processes of selected portfolios, i.e., a good loan means this loan can be fully funded, a lender bid at current will success after its auction and this loan will repay in time. Thus, through simulation, we can compute the average return rates of portfolios for lenders theoretically, which can be treated as an overall metric considering three objectives. In our data, for repayment, we only know the default boolean labels without knowing the specific installments. Thus, when computing the simulative average return rates, we suppose the defaulting borrowers will default a certain *percent* principal amount, which is set as 30% and 60% respectively, (i.e., Return-0.3 and Return-0.6 are our two overall metrics).

4.3.2 *Experimental Results*

The selecting results are shown in Table 4. We can see that, $\mathcal{DPA}$ series can get good results with some certain tradeoffs on different objectives and $\mathcal{EVA}$ series perform best, i.e., EVA_1, EVA_2, EVA_3 perform best on Non-default, Fully-funded, Winning-bid metrics respectively. Comparatively, $\mathcal{EVAR}$ series also perform well and a little worse than $\mathcal{EVA}$ series in most cases. However, $\mathcal{EVAR}$ needs much more time cost than $\mathcal{EVA}$ which will be reported later. Even the random selections, i.e., EVA_R, from Pareto-optimal set are much better than the lenders' own selections on all three metrics. On the two overall metrics, i.e., Return-0.3 and Return-0.6, most methods will improve lenders' returns to varying degrees. Further, EVA_3 and EVA_1 can provide the highest average return rates, i.e., 11.5% and 10.3% returns, respectively. In other words, EVA could provide lenders

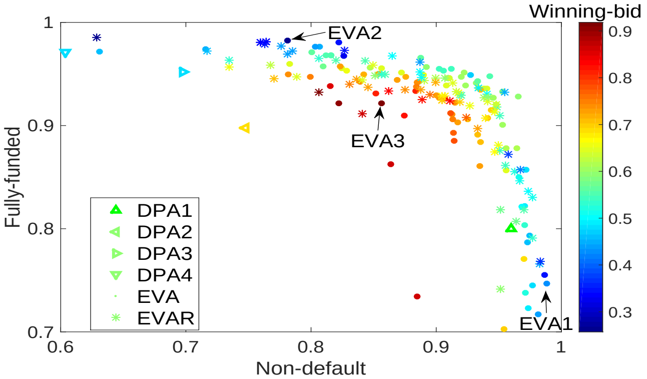


Figure 6: Solutions found by different algorithms.

with more economic profits. *SELF* and several methods with low performances on non-default objective even get negative return rates on Return-0.6.

Figure 6 visualizes the solutions getting from different methods for a random lender, in which one point represents a portfolio. X-axis represents the non-default assessment values (relative values mapping into 0-1), Y-axis represents the fully-funded assessment values and the color represents the winning-bid assessment values. We can see that, $\mathcal{DPA}$ series get one portfolio selection while $\mathcal{EVA}$ (and $\mathcal{EVAR}$) gets a Pareto-optimal portfolio set firstly. In fact, $\mathcal{DPA}$ with a certain parameter $\boldsymbol{\alpha}$ is a specific tradeoff on three objectives. Further, the Pareto-optimal portfolio set getting via $\mathcal{EVA}$ is a little better than $\mathcal{EVAR}$' since the point skyline or envelope surface of $\mathcal{EVA}$ dominates $\mathcal{EVAR}$'. That may be because $\mathcal{EVA}$ adopts the special initialization strategy which can lead to faster convergence and is more conducive to inherit the good individuals/solutions.

We also compare the efficiency results of different algorithms. Since $\mathcal{DPA}$ only provides one solution for each lender while $\mathcal{EVA}$ gets a solution set, for comparison, we also report the results of $\mathcal{DPA}$ running 50 (Pareto set size) times, which is denoted as DPAP. The running time results of seconds (in log scale) are shown in Figure 7. We can see that $\mathcal{DPA}$ and DPAP are most efficient. $\mathcal{EVA}$ needs more time than DPAP does. $\mathcal{EVAR}$ takes much more time costs compared with $\mathcal{EVA}$ since it needs more evolutionary generations. These comparisons of $\mathcal{EVA}$ and $\mathcal{EVAR}$ demonstrate the effectiveness and efficiency of $\mathcal{EVA}$.

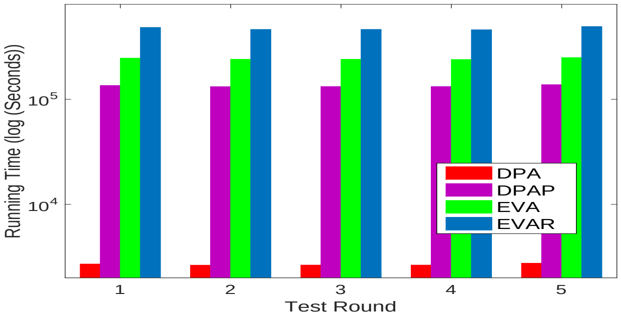


Figure 7: Running time results.

5. RELATED WORK

To the best of our knowledge, there are few existing works on portfolio selections in P2P lending. However, there are some relevant studies in the P2P lending domain, e.g., loan risk assessment [9, 23], bidding behavior analysis [4, 18].

Risk assessment and fully-funded prediction are two hot research topics in P2P lending. For assessing the loan risk or credit, some conventional classification models were adopted, such as Logistic Regression [9, 34] and Neural Network [3]. Fully-funded probability is another important aspect to assess loans in P2P lending. Herzenstein et al. [14] studied both the borrower-related determinants and loan loan-related determinants of funding success in Prosper. Their results indicated that borrower-related financial determinants affected funding the most, while loan-related variables mediate affected the likelihood of funding success. Ryan et al. [28] proposed two regression models combining personal and social determinants and financial determinants for funded percent and number of bids respectively. These works focused on a certain objective assessments, e.g., risk or fully-funded, and none of them adopted or combined the dynamic features extracted from lenders and auctions.

Bidding analysis especially *herding* [4, 18], and social community [13, 6] were also well studied in previous studies. In [6], authors found lender team was an important community to help lenders make decision and promoted lending activities in P2P lending. In [5], authors predicted how likely a given lender would fund a new loan through a gradient boosting tree method. In [34], authors proposed to recommend single personalized loans to lenders by considering both lender preference and loan risk. However, there are few studies in P2P lending from the multi-objective assessments or select/recommend portfolios to lenders in P2P lending.

In addition to the studies in P2P lending, works on recommendation may be also relevant to our portfolio selections to some extent. Recommender system provides suggestions of items that may interest users and aims to predict users' preferences with high accuracy [21, 27]. However, in P2P lending or other financial domains, accuracy may be not as important as in the traditional recommendations, e.g., electronic commerce. In [35, 30], portfolio theory was used on recommendation and information retrieval, but these works still aim to get single items rather than combination items like loan portfolios in our study.

6. CONCLUSIONS

In this paper, we proposed a holistic study on portfolio selections in P2P lending. First, we assessed loans on multiple objectives by a gradient boosting decision tree. We extracted dynamic features from lenders and auctions for better assessment performances. Then, to help lenders select portfolios, we proposed two strategies, i.e., weighted objective optimization strategy and multi-objective optimization strategy. The first strategy provided an optimal portfolio and the multi-objective optimization strategy provided a Pareto-optimal portfolio set for each lender. In two strategies, two selection algorithms, i.e., $\mathcal{DPA}$ and $\mathcal{EVA}$ were designed respectively.

For evaluating our approach, we constructed extensive experiments on Prosper data. The analysis and experimental results demonstrated the significance of our study and the effectiveness of our solutions. Specifically, the extracted dynamic features and $\mathcal{GBDT}$ models significantly improved the assessment performances. Further, the portfolio selections, especially the multi-optimization strategy and $\mathcal{EVA}$ algorithm provided lenders with more economic profits by selecting Pareto-optimal portfolios both effectively and efficiently.

7. ACKNOWLEDGEMENTS

This research was partially supported by National Science Foundation for Distinguished Young Scholars of China

Table 4: Performances of portfolio selections.

Result\Alg	DPA_1	DPA_2	DPA_3	DPA_4	EVA_1	EVA_2	EVA_3	EVA_R	$EVAR_1$	$EVAR_2$	$EVAR_3$	$EVAR_R$	$SELF$
Non-default	0.831	0.789	0.734	0.731	**0.952**	0.714	0.856	*0.806*	0.923	0.715	0.857	0.801	*0.710*
Fully-funded	0.178	0.220	0.323	0.342	0.256	**0.543**	0.214	*0.289*	0.256	0.529	0.202	0.267	*0.123*
Winning-bid	0.778	0.856	0.833	0.802	0.824	0.689	**0.921**	*0.843*	0.801	0.678	0.902	0.810	*0.741*
Return-0.3	0.067	0.073	0.069	0.070	0.112	0.080	**0.115**	*0.105*	0.105	0.085	0.097	0.086	*0.057*
Return-0.6	0.035	0.021	-0.026	-0.027	**0.103**	-0.063	0.069	*0.044*	0.099	-0.049	0.064	0.032	*-0.046*

(61325010), Natural Science Foundation of China (61403358, 61572032, 71571093), National High Technology Research and Development Program of China (2014AA015203) and National Center for International Joint Research on E-Business Information Processing (2013B01035). Qi Liu gratefully acknowledges the support of the Youth Innovation Promotion Association of CAS and acknowledges the support of the CCF-Tencent Open Research Fund.

8. REFERENCES

[1] R. J. Agrawal and J. G. Shanahan. Location disambiguation in local searches using gradient boosted decision trees. In *18th SIGSPATIAL*, pages 129–136. ACM, 2010.

[2] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[3] A. Byanjankar, M. Heikkila, and J. Mezei. Predicting credit risk in peer-to-peer lending: A neural network approach. In *Computational Intelligence, 2015 IEEE Symposium Series on*, pages 719–725. IEEE, 2015.

[4] S. Ceyhan, X. Shi, and J. Leskovec. Dynamics of bidding in a p2p lending service: effects of herding and predicting loan success. In *20th WWW*, pages 547–556. ACM, 2011.

[5] J. Choo, C. Lee, D. Lee, H. Zha, and H. Park. Understanding and promoting micro-finance activities in kiva. org. In *7th WSDM*, pages 583–592. ACM, 2014.

[6] J. Choo, D. Lee, B. Dilkina, H. Zha, and H. Park. To gather together for a better world: Understanding and leveraging communities in micro-lending recommendation. In *23rd WWW*, pages 249–260. ACM, 2014.

[7] C. A. C. Coello, D. A. Van Veldhuizen, and G. B. Lamont. *Evolutionary algorithms for solving multi-objective problems*, volume 242. Springer, 2002.

[8] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Trans.Evo.Com.*, 6(2):182–197, 2002.

[9] G. Dong, K. K. Lai, and J. Yen. Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science*, 1(1):2463–2468, 2010.

[10] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[11] Y. Guo, W. Zhou, C. Luo, C. Liu, and H. Xiong. Instance-based credit risk assessment for investment decisions in p2p lending. *European Journal of Operational Research*, 249(2):417–426, 2016.

[12] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[13] S. Herrero-Lopez. Social interactions in p2p lending. In *3rd Workshop on Social Network Mining and Analysis*, page 3. ACM, 2009.

[14] M. Herzenstein, R. L. Andrews, U. M. Dholakia, and E. Lyandres. The democratization of personal consumer loans? determinants of success in online peer-to-peer lending communities. *Boston University School of Management Research Paper*, (2009-14), 2008.

[15] M. Jahrer, A. Töscher, and R. Legenstein. Combining predictions for accurate recommender systems. In *16th SIGKDD*, pages 693–702. ACM, 2010.

[16] H. Jain and K. Deb. An improved adaptive approach for elitist nondominated sorting genetic algorithm for many-objective optimization. In *Evolutionary Multi-Criterion Optimization*, pages 307–321. Springer, 2013.

[17] M. Kumar and S. I. Feldman. Internet auctions. In *Proceedings of the 3rd USENIX Workshop on Electronic Commerce*, volume 31, 1998.

[18] E. Lee and B. Lee. Herding behavior in online p2p lending: An empirical investigation. *Electronic Commerce Research and Applications*, 11(5):495–503, 2012.

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

[20] D. Liu, D. Brass, Y. Lu, and D. Chen. Friendships in online peer-to-peer lending: Pipes, prisms, and relational herding. *Mis Quarterly*, 39(3):729–742, 2015.

[21] Q. Liu, E. Chen, H. Xiong, C. H. Ding, and J. Chen. Enhancing collaborative filtering by user interest expansion via personalized ranking. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans. on*, 42(1):218–233, 2012.

[22] B. Luo and Z. Lin. A decision tree model for herd behavior and empirical evidence from the online p2p lending market. *Information Systems and e-Business Management*, 11(1):141–160, 2013.

[23] C. Luo, H. Xiong, W. Zhou, Y. Guo, and G. Deng. Enhancing investment decisions in p2p lending: an investor composition perspective. In *17th SIGKDD*, pages 292–300. ACM, 2011.

[24] H. Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952.

[25] E. Mollick. The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing*, 29(1):1–16, 2014.

[26] T. Odean. Volume, volatility, price, and profit when all traders are above average. *The Journal of Finance*, 53(6):1887–1934, 1998.

[27] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[28] J. Ryan, K. Reuk, and C. Wang. To fund or not to fund: Determinants of loan fundability in the prosper. com marketplace. *WP, The Standord Graduate School of Business*, 2007.

[29] L. C. Thomas. *Consumer Credit Models: Pricing, Profit and Portfolios: Pricing, Profit and Portfolios*. OUP Oxford, 2009.

[30] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *32nd SIGIR*, pages 115–122. ACM, 2009.

[31] Z. Wei and M. Lin. Market mechanisms in online peer-to-peer lending. *Management Science*, April 1, 2016.

[32] J. Xie, V. Rojkova, S. Pal, and S. Coggeshall. A combination of boosting and bagging for kdd cup 2009-fast scoring on a large database. In *KDD Cup*, pages 35–43, 2009.

[33] Q. Zhang and H. Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Trans.Evo.Com.*, 11(6):712–731, 2007.

[34] H. Zhao, L. Wu, Q. Liu, Y. Ge, and E. Chen. Investment recommendation in p2p lending: A portfolio perspective with risk management. In *ICDM*, pages 1109–1114. IEEE, 2014.

[35] H. Zhu, H. Xiong, Y. Ge, and E. Chen. Mobile app recommendations with security and privacy awareness. In *20th SIGKDD*, pages 951–960. ACM, 2014.