Data-driven optimization of peer-to-peer lending portfolios based on the expected value framework

Ajay Byanjankar | József Mezei | Markku Heikkilä

1Faculty of Social Sciences, Business and Economics, Åbo Akademi University, Turku, Finland
2Avaintec Oy, Helsinki, Finland

Correspondence
József Mezei, Fänriksgatan 3A, 20500, Turku, Finland.
Email: jmezei@abo.fi

Funding information
Liikesivistysrahasto

Abstract
In recent years, peer-to-peer (P2P) lending has been gaining popularity amongst borrowers and individual investors. This can mainly be attributed to the easy and quick access to loans and the higher possible returns. However, the risk involved in these investments is considerable, and for most investors, being nonprofessionals, this increases the complexity and the importance of investment decisions. In this study, we focus on generating optimal investment decisions to lenders for selecting loans. We treat the loan selection process in P2P lending as a portfolio optimization problem, with the aim being to select a set of loans that provide a required return while minimizing risk. In the process, we use internal rate of return as the measure of return. As the starting point of the model, we use machine-learning algorithms to predict the default probabilities and calculate expected values for the loans based on historical data. Afterwards, we calculate the distance between loans using (i) default probabilities and, as a novel step, (ii) expected value. In the calculations, we utilize kernel functions to obtain similarity weights of loans as the input of the optimization models. Two optimization models are tested and compared on data from the popular P2P platform Lending Club. The results show that using the expected-value framework yields higher return.

KEYWORDS
expected-value framework, peer-to-peer lending, portfolio optimization

1 INTRODUCTION

Peer-to-peer (P2P) lending is an online financial service providing micro-financing solutions. Through an online platform, P2P connects borrowers and lenders to perform loan transactions (Bachmann et al., 2011). From both the borrowers’ and lenders’ perspectives, there are significant differences compared with traditional financial service offerings. The absence of collateral and traditional financial intermediaries in the service, combined with the rapid automated processing, provides easy and quick access to loans at a lower cost (Klafft, 2008). Hence, these attributes attract borrowers who have difficulties in securing loans from traditional banks (Iyer, Khwaja, Luttmer, & Shue, 2009; Serrano-Cinca, Gutiérrez-Nieto, & López-Palacios, 2015). The loan amount requested through P2P lending generally ranges from small to medium and borrowed for a short period of time. Investing in loans through P2P platforms, lenders get a higher return than with similar traditional investments, such as bank deposits. Considering the easy and quick access to loans and the higher return, P2P lending has seen an increasing popularity and growth in recent years (Kumar, Natarajan, Keerthana, Chinmayi, & Lakshmi, 2016).

Additional to these attractive benefits, there is also a high risk of loss in P2P lending investments. The primary source of risk in P2P
lending is the absence of collateral, which implies that lenders can lose their entire investment in the worst-case scenario. With the P2P platform acting as a simple intermediary with no credit services, the increased risk is mostly imposed on the lenders (Li, Hsu, Chen, & Chen, 2016). The risk of failure in P2P investments is also related to the lack of analytical skills of individual investors when making investment decisions and the information asymmetry resulting from the nature of online services (Klafft, 2008; Yum, Lee, & Chae, 2012). Most of the lenders, being nonprofessional investors, have difficulties in screening borrowers; manual assessment of credit risk becomes difficult and requires a high level of expertise (Verstein, 2011; Wang & Greiner, 2011).

The process of P2P lending, as described by Suryono, Purwandari, and Budi (2019), starts with a borrower submitting an application through an online service, including the necessary information, such as demographic and financial data and credit history. The application goes through an underwriting process and the platform evaluates the credit risk; as a result of this evaluation, each loan application is assigned to a risk group/grade. The applications that fulfill the requirements of the evaluation process are listed on the online platform to bid for by potential lenders. After receiving the required funding amount, the application becomes a loan. Lenders can partially fund a loan and spread their investments to multiple loans in order to diversify their portfolio. This provides the opportunity for lenders to select loans from different risk groups based on their requirements of return and risk. Besides loan selection, the amount to invest in a loan also becomes an important issue. These considerations warrant the use of traditional portfolio optimization models to minimize the risk and maximize return (Guo, Zhou, Luo, Liu, & Xiong, 2016).

In creating a low-risk and high-return portfolio, the selection of loans becomes very critical and relies on accurately identifying the risk. There are numerous studies in the literature studying the credit risk in P2P lending. The studies have applied different approaches for studying borrowers’ behavior and predicting credit risk in assisting lenders in loan selection. In recent years, there have been a large number of studies that use credit scoring as the main approach for evaluating borrowers’ risk in P2P lending. By applying machine-learning techniques, the studies focus on accurately predicting the risk, which is typically measured as the probability of default (Byanjankar, Heikkilä, & Mezei, 2015; Emekter, Tu, Jirasakuldech, & Lu, 2015; Kumar et al., 2016; Malekipirbazari & Aksakalli, 2015).

In addition to risk prediction, few studies have implemented profit-scoring approaches that predict the expected profit of a P2P loan (Byanjankar & Viljanen, 2019; Serrano-Cinca & Gutiérrez-Nieto, 2016). However, these approaches either isolate the risk and return or evaluate the risk and return only for a single loan at a time. Since lenders in P2P lending have the ability to spread their investment on multiple loans to create a portfolio, the ideal scenario for them would be to evaluate the overall risk and return of the portfolio. Hence, in this study, we treat the problem of loan selection in P2P lending as a portfolio optimization problem to create a selection of loans with a minimum risk for a required return. We extend the existing literature by proposing to use a previously untapped approach, the expected-value framework by Provost and Fawcett (2013), as an alternative of default probability to be used as the input of the optimization model. As there are no sufficient historical data for each individual borrower, we can identify past loans with similar properties to be used in an instance-based model as presented by Guo et al. (2016). The similarity of loans is defined using the expected value of a loan, and it is determined by the best performing of several advanced machine-learning models. To validate the proposed model, we perform experiments on data from the popular Lending Club P2P platform.

The rest of the article is structured as follows. Section 2 presents a brief literature review on the use of machine learning in P2P default prediction and discusses existing portfolio optimization approaches. The methodology is presented in Section 3, and the data used in the experiments are discussed in Section 4. Section 5 describes the modeling process and presents the results, sensitivity analysis, and discussion. Finally, some conclusions and future research directions are provided in Section 6.

2 | BACKGROUND AND LITERATURE REVIEW

From an investor’s perspective, in the P2P market there are two main tasks (Guo et al., 2016): (i) assessing the quality of individual loans, and based on this, (ii) identifying the optimal allocation of the available investment budget into a combination of loans. Regarding the academic literature, one can find significantly more contributions focusing on the first step of the investment process, i.e., credit risk analysis. As in traditional financial applications, we can approach this problem from the borrowers’ or the lenders’ perspective. While there is a long list of contributions focusing on the borrowers’ decision-making choices, in the following we mainly summarize results from the lenders’ perspective as it is the main focus of the models developed in the article. It is important to note that recent findings support the observation that some simple characteristics impact the decision of lenders to a great extent. For example, Dorfleitner et al. (2019) identified the existence of a third-party trustee and information in the description texts to be the main predictors of successful funding. In Han et al. (2018), the persuasiveness of the information voluntarily provided by lenders is assessed, and the authors found that positive sentiment of the provided information is positively associated with loan success.

In order to support the decision-making of lenders, the most important task is to assess the risk associated with a given loan. This is typically done by constructing statistical and machine-learning models that can estimate the probability of default for individual loans (Kumar et al., 2016). In recent years, we have seen an increase of contributions utilizing various machine-learning techniques to assess credit risk in P2P lending. The algorithms utilized range from basic techniques, such as logistic regression (Emekter et al., 2015) and decision trees (Xia, Liu, & Liu, 2017), to more advanced techniques, such as random forests (Jiang, Wang, Wang, & Ding, 2018) and neural
networks (Byanjankar et al., 2015). The common feature of the majority of the models introduced in the literature is that they focus purely on misclassification error; that is, minimizing the number of loans (or an alternative classification performance metric) that the algorithm classifies incorrectly. While considering more sophisticated evaluation metrics, one can focus on minimizing precision or recall according to the stated preferences, but this perspective does not place any emphasis on understanding and quantifying the cost of making a mistake. Still, this can be an important perspective to consider, especially when an investor aims to create a portfolio of loans and perform investment selection in an optimal way. Consequently, investing in P2P lending can be formulated as a typical portfolio selection problem, the cornerstone of modern portfolio theory as discussed by Markowitz (1959).

2.1 Portfolio optimization in the P2P lending literature

In the following, we briefly summarize the main ideas and components of portfolio optimization models and discuss the important applications in the domain of P2P lending. The main goal of a portfolio optimization model is to derive the optimal weights associated with individual investment choices. Though there are various approaches, the main underlying ideas are typically captured by the Markowitz mean-variance optimization model (Markowitz, 1959); that is, minimizing the risk and maximizing the return. Suppose that we have \( n \) possible assets to allocate our investment into, with associated loan amount \( L_i, i = 1, \ldots, n \), with expected return \( \mu_i \), and covariance between assets \( i \) and \( j \) as \( \sigma_{ij} \). Denoting the weight values (i.e., the fraction of the budget we intend to invest in investment \( i \)), as \( \lambda_i \) and total amount available for investment as \( A \), a bi-objective portfolio optimization model can be formulated as follows (Mansini, Ogryczak, & Speranza, 2015):

\[
\begin{aligned}
\min & \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \sigma_{ij} \\
\max & \sum_{i=1}^n \lambda_i \mu_i \\
\text{s.t.} & \lambda_i A \leq L_i \quad \forall i \\
& \sum_{i=1}^n \lambda_i = 1 \\
& \lambda_i \geq 0 \quad \forall i 
\end{aligned}
\]  

(1)

This is a general form of the optimization model, and various specific versions with some extra constraints are considered in the literature. The first general assumption (and simplification) concerns the parameters \( \sigma_{ij} \). All the papers in the literature applying portfolio optimization in P2P lending consider the values \( \sigma_{ij} \) to be zero when \( i \neq j \) (Babaei & Bamdad, 2020a). This can be attributed to the assumption of negligible correlation between individual loan applications in the P2P market. Consequently, the model is simplified as the objective becomes to minimize the weighted average of \( \sigma_{ij} \) values, from now on denoted simply as \( \sigma_i \) (i.e., a measure of variance associated with the loan application). There is an additional constraint also present in all P2P-related portfolio optimization models limiting the amount that we can invest in a specific loan. On the one hand, an investor cannot assign more money to a loan than the full amount \( L_i \) for loan \( i \). Additionally, most of the P2P lending platforms specify a minimum amount, denoted \( m \), that needs to be invested in order to buy part of a loan. This amount can vary from platform to platform; in our empirical study, we consider data from the Lending Club platform, where the minimum amount is $25.

Beyond these basic notions, the core part of the optimization model, the objective function(s), can be defined according to different preferences. The aforementioned multi-objective formulation can rarely be found in the literature; the main reason for this is that the algorithmic and computational requirements of these types of problems are significantly higher than similar single-objective formulations. The only recent example attempting to formulate and solve a multi-objective formulation was presented by Babaei and Bamdad (2020a), in which they utilized one of the most widely used evolutionary algorithm procedures, NSGA2. The authors found that the solution to the multi-objective formulation offered a solution with lower return and risk than with a profit-scoring model. The comparison with the single-objective model, however, was limited, as portfolio return was constrained to have a lower value than the optimal solution of the multi-objective model. In conclusion, that study alone does not provide sufficient evidence that one can obtain significantly better results by increasing model complexity. In this paper, we opted to focus on a single-objective formulation and attempt to improve on previous results from different perspectives.

Regarding single-objective portfolio optimization models, there are two main approaches: (i) to minimize risk with specifying a minimum acceptable value for return; or (ii) to maximize return with specifying a maximum acceptable value of risk. In the P2P literature, the first approach is utilized in the majority of contributions, and for this reason we also employ this structure.

After the specific form of the optimization model is chosen, a key component of the process is to specify how the expected return and the measure of risk associated with the loan is determined. As there are no historical data available in the same vein as in traditional financial markets, one needs different approaches to estimate the relevant measures. There are two main approaches considered in the literature: instance-based estimations and aggregated/rating-level-based estimations (Chi, Ding, & Peng, 2019).

As there are a limited number of observations available about each individual borrower to make appropriate predictions purely based on that information, the main general idea is to identify similar loans and create predictions using the information of similar borrowers. This can be done, in the simplest case, by relying on some available characterizing variables of a borrower. One of these variables, and the most informative when assessing the credibility of a loan, is the credit rating (Cho et al., 2019). For example, in the Lending Club platform, there are seven grades according to the loan’s credit risk. With the assumption that loans that have the same grade tend to behave similarly (i.e., can be assumed to have the same return and
risk), we can use loans with the same grade to create predictions for new loans.

While using credit rating as the basis of establishing similarity, one can obtain better results by calculating empirical, or so-called instance-based, similarity: We identify loans with similar profiles and use those as the basis of estimations. There are several different approaches that can be used for this method. The most important category of these relies on kernel-based models, as introduced by Guo et al. (2016) and then used in most of the related articles. Kernel regression can be used to assign a weight for each pair of loans measuring the similarity between them. The return and risk of a new loan can be estimated as the similarity weighted average of returns and risks of the other loans. The same approach has been used by Chi et al. (2019) and Babaei and Bamdad (2020a, 2020b), for example, and was found to overperform grade-based similarity in all the cases. As an alternative, Ding, Cheng, and Jiang (2020) utilized graph theory and constructed bipartite graphs based on the data collected from a P2P platform. The network constructed is then used to establish the optimal portfolio.

3 | METHODOLOGY

The research process followed in this article is shown in Figure 1. The process has three critical components:

1. Creating a credit-risk model and estimating risk measures.
2. Calculating the similarities of loan applications based on the estimated risk values (default probability and expected value).
3. Solving the optimization models (one for each of the two similarity values based on the two risk estimates).

In the following, we briefly discuss the three components as applied later in the empirical work. The last stage of the methodology, performance evaluation, takes the output (i.e., the optimal portfolio) of the two optimization models (based on the two risk estimators) and evaluates these two portfolios based on a selected financial indicator (as we discuss later, we chose the internal rate of return [IRR]). We can perform this realistic evaluation as we consider only finished loans in our experiments, and this allows for the direct comparison of the optimal portfolios.

3.1 | Credit-risk model and the expected-value framework

The first step of the investment decision-making process concerns collecting and processing the data about past loans, and then building a model that estimates the level of risk and/or expected return of a loan application. This step of distinguishing good and bad-quality applications has been approached increasingly with various machine-learning approaches. The main goal of this step is typically to estimate the probability of default (Kumar et al. 2016). There are numerous machine-learning models that can be utilized for this purpose; we chose three different algorithms to be used for this purpose:

- **Logistic regression.** This is one of the most widely used approaches. It does not require extensive data processing and is
not sensitive to missing data and provides easily interpretable solutions. Logistic regression has been applied in the P2P literature for example by Kumar et al. (2016) and Xia (2019). Logistic regression models are typically used as a baseline in order to obtain a first model with reasonable performance.

- **Random forest.** The most widely used ensemble methodology that provides very accurate solutions in many applications and can be used efficiently on large data sets. Random forests were used by Jiang et al. (2018) and Malekipirbazari and Aksakalli (2015), and they were found to perform better than other traditional classification models, such as logistic regression and support vector machines.

- **XGBoost.** Extreme gradient boosting is an enhancement over traditional boosting machines that solves many problems in an efficient, flexible, and portable manner. It is becoming a widely used model also in the P2P lending credit-risk model literature, with important contributions by Xia et al. (2017), Ma et al. (2018), and Li, Ding, Wang, Chen, and Yang (2020).

Credit-risk models based on machine-learning algorithms typically calculate the probability of default and use that for making investment decisions. However, in order to assess loans in a more appropriate manner, one also needs to consider the potential loss or gain, not only a simple quantification of risk. The expected-value framework by Provost and Fawcett (2013) can offer a solution by taking into consideration all the possible outcomes that can take place in a situation. As per the traditional definition, the expected value is calculated as the weighted average of the values of all the possible outcomes, where the weights are the probability assigned to the outcomes (Provost & Fawcett, 2013). When implementing a classification model for predicting the final outcome of a loan, all the possible outcomes and the probabilities can be summarized by a confusion matrix, as shown in Figure 2. To create the confusion matrix, one needs to apply a threshold value to determine a binary outcome. In order to optimally select this threshold, one can choose, for example, the value at which the $F_2$ score is maximum for classifying the loans. As a modification of the traditional $F_1$ score, the mean of precision and recall, the $F_2$ score assigns higher weight to the recall than to precision. Consequently, we assign more importance to (correctly identifying) the positive class: the default loans. $F_2$ (and $F_1$) is a special cases of the $F_\beta$ measure, which can be defined as follows (Sasaki, 2007):

$$F_\beta = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2 FN + FP}$$

Substituting 2 in the place of $\beta$, we obtain $F_2$, which, according to the formula, emphasizes more avoiding false negative (FN) errors. In other words, by establishing the thresholds based on the $F_2$ score, we aim at selecting models that identify more true default loans correctly with the price of potentially misclassifying more nondefault loans.

With the confusion matrix, one can further calculate more metrics required for expected value calculation as follows:

- True positive (TP) rate, $tpr = TP / (TP + FN)$
- FN rate, $fnr = FN / (FN + TP)$
- False positive (FP) rate, $fpr = FP / (FP + TN)$
- True negative (TN) rate, $tnr = TN / (TN + FP)$.

Applying the confusion matrix metrics and the prediction from the model, the expected value of a loan can be calculated as follows:

$$EV = p(P)[tpr \times b(TP) + fnr \times c(FN)] + p(N)[fpr \times c(FP) + tpr \times b(TN)]$$

where $p(P)$ is the probability of default of a loan and $p(N)$ is the probability of the loan being good.

The values for the possible outcomes of the loan can be represented by a cost–benefit matrix, as shown in Figure 3, with cost for misclassification and benefit for correct classification. Assigning the value of costs and benefits is an important and, at the same time, challenging task in real-world applications (Oreski, Oreski, & Oreski, 2012). In this study, the benefit of correctly classifying defaults $b(TP)$ is defined to be zero as no investment is made. The cost of falsely identifying good loans as defaults $c(FP)$ is also assigned the value zero, as again no investment would be made because the loan is predicted to be default. There could be (lost) opportunity cost associated with an FP
loan, but in this study we do not consider this perspective (Wang, Kou, & Peng, 2020).

When classifying defaults as good loans, there is a cost $c(FN)$ that represents the loss due to default. To estimate this loss, we calculate the loss when a loan defaults with the charged off loans. We then use the average loss given default based on a loan grade to estimate the loss. Finally, when correctly classifying good loans to be good, there is a benefit $b(TN)$ received from the interest on loans. We estimate the benefit from the monthly installments paid on the loans. For simplicity, we calculate the benefit as the difference between the sum of all the monthly installments to be paid and the funded loan amount. However, in a real scenario, the benefit could be affected by the service charges and taxation, which are not considered in this study.

3.2 Loan similarity and the instance-based approach

In order to estimate return and risk for new loan applications, one can use the information from similar previous loans. In the instance-based approach, the expected return and risk of each loan are estimated as a weighted average of historical observations of closed loans (Chi et al., 2019). The weights are derived based on the similarity of instances. For the similarity calculations, we utilize the framework proposed by Guo et al. (2016): kernel weights computed using a distance measure. The originally proposed distance is based on the default likelihood distance, as used, for example, by Guo et al. (2016) and Chi et al. (2019). Our proposal is to use the output of the expected-value calculation as the basis of weight calculation, and the main goal of this study is to assess whether we can improve prediction performance.

After the weights are determined, for a new loan $i$, we can calculate the expected return as the weighted average of the historical returns of the past loans. For the historical return value, we choose to calculate and use the IRR measure as proposed by Serrano-Cinca and Gutiérrez-Nieto (2016). IRR can be computed from the cash outflow (the loan amount) followed by several cash inflows (the payments), accounting for irregular repayment schedules. The advantages of IRR include (Serrano-Cinca & Gutiérrez-Nieto, 2016) that, as a continuous variable, it allows for more precise information when compared with a dichotomous variable and that it takes into account not only loan payments but also loan interest rates.

After the kernel weights and the IRR values are calculated, we can compute the weighted average as

$$\mu_i = \sum_{j=1}^{n} w_i \text{IRR}_j$$  \hspace{1cm} (4)

where $\text{IRR}_j$ is the observed IRR of loan $j$ and $w_i$ is the weight of loan $j$ for predicting the return of the new loan $i$. For calculating the expected risk, we can use a similar formula as follows (Guo et al., 2016):

$$\sigma_i = \sum_{j=1}^{n} w_i (\text{IRR}_j - \mu_i)^2$$  \hspace{1cm} (5)

3.3 Optimization model

The expected return and risk calculated in Section 3.2 are used as the input for a portfolio optimization model. Based on the observations in discussing existing portfolio optimization models in the P2P lending literature, one can identify the following portfolio optimization problem as the most widely used in the literature (Guo et al., 2016), using the notation introduced earlier herein:

$$\min \sum_{i=1}^{n} \lambda_i^2 \sigma_i^2$$

s.t. $\sum_{i=1}^{n} \lambda_i \mu_i \geq R$

$$\sum_{i=1}^{n} \lambda_i = 1$$

$m \leq \lambda_i A \leq L_i \ \forall i \ \text{when} \ \lambda_i \neq 0$

$$\lambda_i \geq 0 \ \forall i$$

In order to solve this model, we have to deal with a quadratic objective function and to deal with the constraint of minimum investment. A typical approach is to introduce new binary variables that specify whether we invest in the loan or not. The resulting problem becomes a mixed nonlinear optimization formulation that, in general, can be difficult to solve. A good introduction to this topic, in particular focusing on portfolio optimization, can be found in Mansini et al. (2015).

| Grade | Count  | Default rate (%) | Interest rate (%) | FICO score | Debt to income (%) |
|-------|--------|------------------|------------------|------------|-------------------|
| A     | 72,133 | 6                | 6.93             | 721        | 16.28             |
| B     | 108,765| 13               | 10.04            | 696        | 17.98             |
| C     | 107,410| 22               | 13.29            | 688        | 19.62             |
| D     | 55,178 | 32               | 16.72            | 685        | 21.30             |
| E     | 29,762 | 41               | 19.29            | 684        | 21.56             |
| F     | 8,530  | 50               | 23.63            | 683        | 21.47             |
| G     | 1,957  | 54               | 26.84            | 681        | 20.14             |
| Average | 20.21  | 12.43            | 695              | 18.96      |
| Total  | 383,735|                  |                  |            |                   |

**TABLE 1** Data summary statistics
The growth in P2P lending has generated large amounts of data related to P2P transactions. In the empirical study, we use publicly available data from Lending Club, a leading P2P platform in the USA, including information of all the loans issued in year 2015.

The data include two sources of information on the loans. The first source includes the information on loan and borrower characteristics and borrower credit history. It provides information on attributes such as loan amount, loan purpose, loan grade, interest rate, borrower’s annual income, employment status, FICO® score, debt-to-income ratio, and account delinquencies. In addition, it describes the current status of the loan, whether it is fully paid, current, late, default or charged off. We use this information for training the models to predict default probability. The second source of the data includes information on the monthly payments made by the borrowers on the loans, including the monthly principal, interest, and late fee payments received from borrowers. We make use of the monthly payments information to calculate the IRR on the loans to be used in the expected-return calculations.

In the data, only loans that were either fully paid or charged off were utilized for the study as we do not know the final status of ongoing loans. Based on this, the final data set includes 383,735 loans, 80% of which are fully paid and 20% are charged off, creating an imbalanced data set. Table 1 displays the average summary statistics of the loan attributes for each loan grade. From Table 1, we can observe that the majority of the loans belong to grades B and C. Loans in grade A have comparatively low default rates, and the default rates are high for loans in grades E–G, compensated by high interest rates.

5 | RESULTS AND DISCUSSION

In this section, we present the results of the empirical analysis. We start with the selection of machine-learning model and the calculation of the components of the expected-value framework. This is then used as the input for the similarity calculations and the optimization

| TABLE 2 | Models comparison |
| --- | --- |
| Model | AUC | PR-AUC | log(loss) |
| Logistic regression | 0.7374 | 0.4172 | 0.5272 |
| Random forest | 0.7325 | 0.4060 | 0.5373 |
| XGBoost | 0.7423 | 0.4221 | 0.5774 |

AUC: area under receiver operating curve; PR-AUC: area under precision and recall curve.
The best values are highlighted in bold.

| TABLE 3 | Optimization results for segments |
| --- | --- | --- |
| Segment | Default probability | Expected value |
| 1 | 0.0545 | 0.0866 |
| 2 | 0.0759 | 0.103 |
| 3 | 0.0789 | 0.0723 |
| 4 | 0.0555 | 0.0715 |
| 5 | 0.0557 | 0.142 |
| 6 | 0.0558 | 0.109 |
| 7 | 0.0580 | 0.129 |
| 8 | 0.0497 | 0.0711 |
| 9 | 0.0520 | 0.106 |
| 10 | 0.0643 | 0.0825 |
| Average | 0.0600 | 0.0973 |
| Variance | 0.00991 | 0.0248 |

FIGURE 4 Features importance from XGBoost model

BYANJANKAR ET AL.
model. The data analysis was performed with the programming languages Python and R, making use of packages offering implementation of machine-learning algorithms and nonlinear optimization procedures.

### 5.1 Credit-risk model

We trained several credit-risk models to predict default probability of the loans with the three widely used machine-learning models introduced earlier herein. For training the models, the training and test sets were created based on the loan issue dates within the year 2015 considered in the experiment. The loans issued in the last 3 months of the year (October, November, and December) were used as the test set and the rest were used as the training set. This approach for creating training and test sets was applied as we aim at predicting default probability and return based on similar loans from the past and want to simulate a realistic scenario faced by an investor. This setting also provides a realistic scenario for solving the optimization models in the next stage, as we can find similar loans from the first three quarters of the years for loans issued in the last quarter. The features for training the models were selected using the permutation feature importance method (Breiman, 2001). The final set of features selected for training the models are as follows:

- **Int_rate**: Interest rate on the loan.
- **Term**: The number of payments on the loan, that can be either 36 months or 60 months.
- **Dti**: Debt-to-income ratio.
- **Loan_amnt**: The listed amount of the loan applied for by the borrower.
- **Avg_cur_bal**: Average current balance of all accounts.
- **Emp_length**: Employment length in years. Possible values are between 0 and 10, where 0 means less than 1 year and 10 means 10 years or more.
- **Fico_score**: FICO score of borrower.
- **Home_ownership**: The home ownership status provided by the borrower during registration or obtained from the credit report; the values can be RENT, OWN, MORTGAGE, OTHER.
- **Grade**: Grade assigned to the loan by Lending Club.
- **mort_acc**: Number of mortgage accounts.
- **Revol_util**: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
- **Revol_bal**: Total credit revolving balance.

### TABLE 4 Portfolio optimization with budget parameter

| Budget | Default probability | Expected value |
|--------|---------------------|----------------|
| 3,000  | 0.0575              | 0.1046         |
| 5,000  | 0.0582              | 0.1035         |
| 10,000 | 0.05935             | 0.1106         |
| 15,000 | 0.0600              | 0.0973         |
| 20,000 | 0.0605              | 0.0903         |
| 30,000 | 0.0616              | 0.1014         |

### TABLE 5 Sensitivity analysis with budgets and required return

| Budget | Required return | Default probability | Expected value |
|--------|-----------------|---------------------|----------------|
| 10,000 | 0.05            | 0.0593              | 0.111          |
| 10,000 | 0.055           | 0.0593              | 0.111          |
| 15,000 | 0.05            | 0.0615              | 0.0883         |
| 15,000 | 0.055           | 0.0600              | 0.0973         |
| 15,000 | 0.06            | 0.0623              | 0.0927         |
| 25,000 | 0.05            | 0.0609              | 0.102          |
| 25,000 | 0.055           | 0.0609              | 0.102          |
| 25,000 | 0.060           | 0.0635              | 0.0244         |

**Annual_inc** The self-reported annual income provided by the borrower during registration.

**Mths_since_recent_inq** Months since most recent inquiry.

**Year_from_earliest_cr_line** Number of years from the earliest credit account at the time of application.

**Delinq_2yrs** The number of ≥30 days past-due incidences of delinquency in the borrower’s credit file for the past 2 years.

Since the data are imbalanced and that misclassifying defaults as nondefaults has more associated costs, we implemented cost-sensitive classification models with class weights. Taking into consideration the imbalance ratio and applying a heuristic approach, we assigned class weights, where the default class was assigned four times the weight of the nondefault class. Furthermore, one-hot encoding was applied for representing categorical features as numerical, which resulted in 41 features, increased from the original 17. Grid search with cross-validation was applied to search for the best parameters for the models. Finally, the models were evaluated on the test set using three evaluation metrics: area under receiver operating curve, area under precision and recall curve, and log(loss).

The results of the models are not significantly different from each other, as seen in Table 2. The XGBoost model performs slightly better than the rest of the models with respect to most of the evaluation metrics. Hence, treating it as the best-performing model, we use the default probabilities produced by XGBoost as the basis of further modeling, in particular for calculating the components of the expected-value framework. Though it is not the focus of the study, it is an interesting point to look at the loan attributes that are assigned high importance according to the chosen model. We use “gain” criteria to calculate feature importance for the XGBoost model, which ranks the features based on the contribution of a feature in improving classification accuracy across all the trees in the model.

The results from the feature importance calculations show that not all features have significant contribution to model accuracy. The top 15 most important features from the XGBoost model are presented in Figure 4; they account for about 90% of the total contribution to improvement in model accuracy. The model identifies three characteristics (a loan belonging to grade A, interest rate, and having a 36 months period) that have high influence in determining the default probability of a loan. As it can be observed in the figure, these
three variables alone account for a significant part of the model's performance.

5.2 | Evaluating the optimization model

As already discussed, one can use default probability and expected value to calculate different distance matrices for the purpose of return and risk estimation of new loans based on historical performances of similar loans. In the remainder of the analysis, we performed the calculations using a random sample of 20,000 data points from the training set for computational reasons. It is important to note that though we rely on a sample in the analysis, the number of data points is still significantly higher than most of the previous studies, which utilize a few hundred loans as input to the calculations in the optimization process (Chi et al., 2019; Guo et al., 2016). Furthermore, for computational purposes and variability, we randomly created 10 segments of 2,000 data points from the test set. Portfolio optimization is performed separately for each of the 10 segments created. In the analysis, the investment amount is set as $15,000 and the required return is specified to be 5%. Table 3 shows the optimization results on the 10 segments with this setting. The results clearly show that, with the expected-value approach, we achieve a higher expected portfolio return for the majority of the segments (nine out of 10), and the average expected return across the segments is also higher. It is important to note at the same time that the expected-value approach results in IRR with higher variance across the segments compared with the default probability approach.

5.3 | Sensitivity analysis

Lenders in P2P lending, being individual investors, may have different risk–return requirements. Hence, we performed sensitivity analysis from different perspectives and recalculated the optimization model with varying parameters.

Considering the diversity of P2P investors in terms of budget they have available for investment, we performed a sensitivity analysis on the impact of the investment budget. With the given investment budget and the required return of 5%, the optimization is performed with both approaches. Table 4 presents the comparison of the portfolio optimization results for the two approaches for different investment budgets. The results are reported as the average expected return across the 10 test segments for each of the budget parameters. The comparison clearly shows that the expected-value framework approach performs consistently better than the default probability approach. The better results with the expected-value framework approach can be intuitively explained based on the construction of the estimates: Contrary to default probability, the expected-value framework provides an estimate of the return in the presence of the risk. Therefore, an estimate of the return helps in finding profitable loans, whereas default probability only estimates the risk.

For the second part of the sensitivity analysis, we considered the two most important parameters, investment budget and required return, and tested different combinations of them (Guo et al., 2016). The sensitivity analysis is performed with three possible budget values: $10,000, $15,000, and $25,000. Similarly, three values of required return were tested: 5%, 5.5%, and 6%. The sensitivity analysis is performed with three possible budget values: $10,000, $15,000, and $25,000. Similarly, three values of required return were tested: 5%, 5.5%, and 6%. The sensitivity analysis is performed with three possible budget values: $10,000, $15,000, and $25,000. Similarly, three values of required return were tested: 5%, 5.5%, and 6%. The sensitivity

FIGURE 5 Portfolio optimization with weight constraints
analysis reported in Table 5 shows that the optimization based on the expected-value framework approach performs better than the default probability approach for almost all the combinations.

We further extended the sensitivity analysis applying different weights constraints, which would prevent the optimization algorithm from assigning high weights (i.e., high proportion of the available budget to a single loan). Four possible weight values were tested: 0.1, 0.3, 0.6, and 0.9. For instance, a constraint with 0.1 would imply that we are not allowed to allocate more than 10% of the total available budget into one single loan. The weight constraints were combined with six possible different investment budgets when performing the optimization. The results of the portfolio optimization with the multiple combination of weight constraints and budgets are shown in Figure 5. The results show the average return across the 10 test segments. Again, one can clearly see that the optimization based on input from the expected-value approach performs better than the default probability approach.

6 | CONCLUSIONS

In this paper, we proposed a data-driven model for P2P lending decision-making: An optimal portfolio of loans is identified to aid lenders in their investment allocation decisions. The proposed model utilizes an instance-based credit-risk assessment framework, when the return and risk of new loans is predicted using the information of similar borrowers’ historical data. As the main novelty of this paper, we proposed to use the expected-value framework introduced by Provost and Fawcett (2013) as the basis of establishing the similarity weights, combined with traditional kernel estimations discussed by Guo et al. (2016).

The estimated return and risk values are then used in traditional mean-variance optimization models to obtain an optimal portfolio of investments. We conducted several numerical experiments using data from the popular Lending Club P2P lending platform. As the results of the experiments show, the proposed model offers better performance than existing models, particularly models that use rating-based similarity weights constraints, which would prevent the optimization algorithm from assigning high weights (i.e., high proportion of the available budget to a single loan).

Future studies can focus on extending the present work from various aspects. First, additional machine-learning models could be tested in combination of the expected-value framework to further optimize the first part of the modeling process. Second, the model should be tested on data from different P2P platforms across different countries in order to further assess validity.

DATA AVAILABILITY STATEMENT

The data were derived from the following resources available for download from the P2P lending platform after registration: https://www.lendingclub.com/info/statistics.action

REFERENCES

Babaei, G., & Bamdad, S. (2020a). A multi-objective instance-based decision support system for investment recommendation in peer-to-peer lending. Expert Systems with Applications, 150, 113278.

Babaei, G., & Bamdad, S. (2020b). A neural-network-based decision-making model in the peer-to-peer lending market. Intelligent Systems in Accounting, Finance and Management, 7, 142–150.

Bachmann, A., Becker, A., Buercrckner, D., Hilker, M., Kock, F., Lehmann, M., Tiburtius, P., & Funk, B. (2011). Online peer-to-peer lending—A literature review. Journal of Internet Banking and Commerce, 16(2). Retrieved from https://www.licommercecentral.com/open-access/online-peertopeer-lending-a-literature-review.php?id=38238

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

Byanjankar, A., Heikillä, M., & Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach, 2015 IEEE symposium series on computational intelligence, pp. 719–725.

Byanjankar, A., & Viljanen, M. (2019). Predicting expected profit in ongoing peer-to-peer loans with survival. Intelligent decision technologies 2019: Proceedings of the 11th KES international conference on intelligent decision technologies (KES-IDT 2019), volume 1, pp. 15–26.

Chi, G., Ding, S., & Peng, X. (2019). Data-driven robust credit portfolio optimization for investment decisions in P2P lending. Mathematical Problems in Engineering, 2019, 1902970.

Cho, P., Chang, W., & Song, J. W. (2019). Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision. IEEE Access, 7, 16925–16939.

Ding, Y., Cheng, L., & Jiang, C. (2020). Choosing portfolios based on bipartite graph of P2P lending networks. Data Analysis and Knowledge Discovery, 3(12), 76–83.

Dorfliebtn, G., Oswald, E.-M., & Zhang, R. (2019). From credit risk to social impact: On the funding determinants in interest-free peer-to-peer lending. Journal of Business Ethics. https://doi.org/10.1007/s10551-019-04311-8

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. Applied Economics, 47(1), 54–70.

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, 249(2), 417–426.

Han, J.-T., Chen, Q., Liu, J.-G., Luo, X.-L., & Fan, W. (2018). The persuasion of borrowers’ voluntary information in peer to peer lending: An empirical study based on elaboration likelihood model. Computers in Human Behavior, 78, 200–214.

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2009). Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending? AFA 2011 Denver meetings paper. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3D1570115

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Annals of Operations Research, 266(1–2), 511–529.

Klafft, M. (2008). Online peer-to-peer lending: A lenders’ perspective. Proceedings of the international conference on e-learning, e-business, enterprise information systems, and e-government, EEE, pp. 371–375.

Kumar, V., Natarajan, S., Keerthana, S., Chinmayi, K. M., & Lakshmi, N. (2016). Credit risk analysis in peer-to-peer lending system, 2016 IEEE international conference on knowledge engineering and applications (ICKEA), pp. 193–196.

Li, J., Hsu, S., Chen, Z., & Chen, Y. (2016). Risks of P2P lending platforms in China: Modeling failure using a Cox hazard model. The Chinese Economy, 49(3), 161–172.
Li, W., Ding, S., Wang, H., Chen, Y., & Yang, S. (2020). Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China. *World Wide Web, 23*(1), 23–45.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications, 31*, 24–39.

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

Mansini, R., Ogryczak, W., & Speranza, M. G. (2015). Linear and mixed integer programming for portfolio optimization. Cham, Switzerland: Springer International Publishing.

Markowitz, H. (1959). *Portfolio selection: Efficient diversification of investments*. New York, NY: John Wiley & Sons Inc.

Oreski, S., Oreski, D., & Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert Systems with Applications, 39*(16), 12605–12617.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media.

Sasaki, Y. (2007). The truth of the F-measure. Lecture notes. Retrieved from https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems, 89*, 113–122.

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PloS One, 10*(10), e0139427.

Suryono, R. R., Purwandari, B., & Budi, I. (2019). Peer to peer (P2P) lending problems and potential solutions: A systematic literature review. *Procedia Computer Science, 161*, 204–214.

Verstein, A. (2011). The misregulation of person-to-person lending. *University of California, Davis, Law Review, 45*, 445–530.

Wang, H., & Greiner, M. E. (2011). Prosper—The eBay for money in lending 2.0. *Communications of the Association for Information Systems, 29*(1), 13.

Wang, H., Kou, G., & Peng, Y. (2020). Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. *Journal of the Operational Research Society, https://doi.org/10.1080/01605682.2019.1705193*

Xia, Y. (2019). A novel reject inference model using outlier detection and gradient boosting technique in peer-to-peer lending. *IEEE Access, 7*, 92893–92907.

Xia, Y., Liu, C., & Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications, 24*, 30–49.

Yum, H., Lee, B., & Chae, M. (2012). From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms. *Electronic Commerce Research and Applications, 11*(5), 469–483.

---

How to cite this article: Byanjankar A, Mezei J, Heikkilä M. Data-driven optimization of peer-to-peer lending portfolios based on the expected value framework. *Intell Sys Acc Fin Mgmt*. 2021;28:119–129. https://doi.org/10.1002/isaf.1490