RESEARCH

Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets

Štefan Lyócsa[1,2,5*] , Petra Vašaničová[1], Branka Hadji Misheva[3] and Marko Dávid Vateha[4]

*Correspondence:
info@econ.muni.cz
[5] Faculty of Economics
and Administration, Masaryk
University, Lipová 507/41a,
602 00 Brno, Czech Republic
Full list of author information
is available at the end of the
article

Abstract

For the emerging peer-to-peer (P2P) lending markets to survive, they need to employ credit-risk management practices such that an investor base is profitable in the long run. Traditionally, credit-risk management relies on credit scoring that predicts loans' probability of default. In this paper, we use a profit scoring approach that is based on modeling the annualized adjusted internal rate of returns of loans. To validate our profit scoring models with traditional credit scoring models, we use data from a European P2P lending market, Bondora, and also a random sample of loans from the Lending Club P2P lending market. We compare the out-of-sample accuracy and profitability of the credit and profit scoring models within several classes of statistical and machine learning models including the following: logistic and linear regression, lasso, ridge, elastic net, random forest, and neural networks. We found that our approach outperforms standard credit scoring models for Lending Club and Bondora loans. More specifically, as opposed to credit scoring models, returns across all loans are 24.0% (Bondora) and 15.5% (Lending Club) higher, whereas accuracy is 6.7% (Bondora) and 3.1% (Lending Club) higher for the proposed profit scoring models. Moreover, our results are not driven by manual selection as profit scoring models suggest investing in more loans. Finally, even if we consider data sampling bias, we found that the set of superior models consists almost exclusively of profit scoring models. Thus, our results contribute to the literature by suggesting a paradigm shift in modeling credit-risk in the P2P market to prefer profit as opposed to credit-risk scoring models.

Keywords: Profit scoring, Credit scoring, Financial intermediation, P2P, Fintech

Introduction

The peer-to-peer (or person-to-person, P2P) lending market facilitates financial transactions between borrowers and lenders. Considering that services are processed through Internet technologies, P2P lending is considered a financial innovation, a so-called FinTech product (Ahelegbey et al. 2019; Kim and Cho 2019b; Kou et al. 2021a; Allen et al. 2021; for a broader review of FinTech-related research). The P2P lending market offers the possibility for borrowers who would often not be eligible for a loan through bank-offered services to obtain one (e.g., Li et al. 2018). Lenders, often individual investors, are lured to participate by higher interest rates and diversification

potential. The P2P market might represent a substitute and a complement to the traditional bank lending market (Tang 2019; De Roure et al. 2021). The substitution effect might occur during a regulatory or systemic shock to the traditional banking sector (Kou et al. 2021b). On the contrary, complementarity is visible when the P2P lending sector extends credit to the markets that remain underserved within the traditional lending paradigm (see Jagtiani and Lemieux 2018 [for LC dataset]; Jagtiani et al. 2021 [for mortgage market]).

Although different business models exist for P2P lending, the most common and traditional approach is centered on an Internet-based lending platform that facilitates transactions between borrowers and potential lenders (e.g., Lending Club, Bondora, Prosper). Default rates above 10% are not exceptional in P2P markets (e.g., Bondora). Given the informational asymmetry between borrowers and lenders (see Emekter et al. 2015; Serrano-Cinca et al. 2015) and that such loans are usually unsecured, the lender faces considerable credit risks (Kou et al. 2014; Li et al. 2021). As with any new technology, for the P2P lending market to be sustainable, a (growing) customer base is necessary, which essentially means that lenders who will be profitable in the long run are needed. This case requires state-of-the-art credit-risk models, which in turn is the aim of this paper. Balyuk (2019) showed how credit from a P2P market might signal banks' increased creditworthiness of the borrower. The effect is larger for borrowers with short credit history and low credit grades. This information spillover from the P2P to the traditional credit market is somewhat surprising, given that banks have decades of experience in consumer lending. However, Balyuk (2019) suggested that the improved ability of P2P lenders to evaluate credit risk can most likely be attributed to the use of machine learning (ML) algorithms and alternative data sources. Interestingly, Jagtiani and Lemieux (2019) found that the correlation of credit grades for similar loans issued by banks and provided by P2P platforms declined over time as well. This case is an effect that can be attributed to the unique data or advanced scoring methodology used by players in the P2P lending market. Given the recommendations of the Basel Committee for Banking Supervision, the financial industry has developed a plethora of statistical credit-risk evaluation methods that harness information from past loans to assess the credit risk of new loan applicants and thereby aid credit-risk modeling (e.g., Bastani et al. 2019; Giudici and Misheva 2018; Guo et al. 2016; Kim and Cho 2019a; Malekipirbazari and Aksakalli 2015; Serrano-Cinca and Gutiérrez-Nieto 2016; Xia et al. 2017a, b, 2018).

Credit-risk models that are based on predicting the probability of default are usually referred to as credit scoring (CS) models. Most of the advances in P2P credit-risk models fall into this category (e.g., Malekipirbazari and Aksakalli 2015; Xu et al. 2019; Xia et al. 2018; Li et al. 2020). Alternatively, profit scoring (PS) models predict a loan's profitability and not only defaults. Despite very promising early results (e.g., Bastani et al. 2019; Serrano-Cinca and Gutiérrez-Nieto 2016; Xia et al. 2017b), little effort has been made in this domain, and many unanswered questions remain. In this paper, we contribute to this latest strand of the literature, that is, we propose a PS-based model that predicts the annualized adjusted internal rate of return of a loan. We document that in the context of both data sets considered (Bondora and Lending Club market places) and in an out-of-sample framework, our approach outperforms the standard CS models in terms of statistical significance and economic relevance (profitability and total profit).

The remainder of the paper is organized as follows. In the next section, we provide a short review of the most closely related works on PS models in the P2P market. Next, we present the data and specific features of our data sets. Then, we present our PS method, namely, how we estimate a loan's profitability, the statistical models that we use, and our forecasting and evaluation procedures. The presentation of empirical results follows, and the final section summarizes our key findings.

Related works

In the P2P literature, credit-risk models are based on either statistical (e.g., logistic regression (LR); Emekter et al. 2015; Ge et al. 2016; Guo et al. 2016; Serrano-Cinca et al. 2015; Zhang et al. 2017), nonparametric (e.g., decision trees or random forest [RF]; Malekipirbazari and Aksakalli 2015; Zhou et al. 2019), or artificial intelligence methods (e.g., support vector machines or artificial neural networks; Byanjankar et al. 2015; Sjoblom et al. 2019; Xu et al. 2019). Among the credit-risk models, the standard CS framework models the probability of default. Subsequently, the higher the risk of the borrower, the lower the given credit rating grade.[1]

Although the CS approach has proven to be successful, the goal of CS is not necessarily aligned with the investor's long-term goal, which is profit maximization. For example, many of the non-performing loans have a history of payments, suggesting that not all non-performing loans are alike. In some cases, borrowers may have paid off a sum that equals or is even greater than the initial loan amount, whereas in other cases, not a single payment has been made. Clearly, for the lender/investor, the difference between the two non-performing loans is relevant because it leads to different loan returns. Serrano-Cinca and Gutiérrez-Nieto (2016), therefore, suggested using the PS approach for credit-risk models, where the dependent variable is represented by the loans' returns as opposed to an indication of whether the loan defaulted or not.

In the P2P literature, the first PS model was presented by Serrano-Cinca and Gutiérrez-Nieto (2016). In their study, they used a data set from Lending Club (US) and found that CS and PS models represent different aspects of the loan. The reason is that the factors driving the probability of default are different from those driving investors' profitability. Moreover, they report that when using a decision tree approach (CHAID) in a PS framework, the returns are not only above the average but also above those suggested by LR models. Next, Xia et al. (2017b) proposed a cost-sensitive boosted decision tree to evaluate annualized loan return. Using data from Lending Club and We.com (China), they found that their approach outperforms standard methods, and more importantly, PS models that explain annualized returns outperform CS models to explain loan defaults. Bastani et al. (2019) followed the work of Serrano-Cinca and Gutiérrez-Nieto (2016) in that they used data from Lending Club and model the internal rate of return. Their approach is interesting in that it draws from the wide and deep learning algorithm of Cheng et al. (2016), which combines the predictions from the CS and the PS models

---

[1] Motivated by the literature in finance, marketing, and psychology, extensive research suggested that utilizing soft factors can address some of the main short-comings of traditional approaches (e.g., Duarte et al. 2012; Liang and He 2020; Zhang et al. 2020). However, with respect to hard factors, soft factors also seem to have only limited potential to improve discrimination between good and bad loans (Wang et al. 2020). We therefore focus on traditional hard factor ML-based credit risk model analysis, which in turn allow for loan contracts to be more personalized, reflecting the unique features of the specific borrowers.

in two stages. In the first stage, they predicted non-default loans, which are modeled in the second stage, where the predicted internal rate of return is of interest. In a test sample, the proposed two-stage approach resulted in positive returns and fared better than the approach of Serrano-Cinca and Gutiérrez-Nieto (2016). Thus, assembling information from CS and PS models might make sense. Finally, Xia et al. (2017b) and Bastani et al. (2019) also addressed the imbalance problem of the P2P datasets, where bad loans tend to be under-represented. They argued that models that account for the imbalance problem might lead to more accurate predictions.

Surprisingly, the literature on PS on P2P is very limited, given that the PS models seem to clearly outperform their CS counterparts. In this paper, we contribute to the literature on the P2P PS models in several ways. First, we do not model annualized return (Xia et al. 2017b) or the standard internal rate of return (Bastani et al. 2019; Serrano-Cinca and Gutiérrez-Nieto 2016). Instead, we model an adjusted annualized internal rate of return, where the reinvestment rate is based on the performance of previous loans on the market. Second, we propose a statistical framework that is based not only on standard, easy to implement and interpret models (multivariate linear and LRs with regularization constraints) but also on more sophisticated models (RF and neural networks) that are used for CS and PS, thereby facilitating a fair comparison. Third, previous evidence is mostly related to the Lending Club marketplace, whereas Serrano-Cinca and Gutiérrez-Nieto (2016) and Bastani et al (2019) noted that PS models need to be validated on other P2P platforms as well. Are previous positive results of the PS model related only to the Lending Club (Bastani et al. 2019; Serrano-Cinca and Gutiérrez-Nieto 2016; Xia et al. 2017b) or we.com (Xia et al. 2017b) lending markets? Evidence from other markets is missing. Therefore, apart from a random sample of loans from the Lending Club database, we use a sample of loans from a European platform Bondora, which offers short-term risky loans. We found that PS models tend to perform much better than CS models, and therefore, we strengthen the case for PS models in the literature. We also evaluate individual models' performance using absolute profits and returns as a loss function because these are ultimately the main concerns of investors. Contrary to most studies in this field, our evaluation is also based on statistical tests that consider data snooping bias. We found a set of superior models that almost always include models that predict loan returns.

Data

Existing studies predominantly worked with data from the US-based Lending Club (Bastani et al. 2019; Emekter et al. 2015; Guo et al. 2016; Jin and Zhu 2015; Serrano-Cinca and Gutiérrez-Nieto 2016; Serrano-Cinca et al. 2015; Teply and Polena 2020; Xia et al. 2017b; Ye et al. 2018) and Prosper (Guo et al. 2016; Miller 2015; Wang et al. 2018; Zhang and Liu 2012; Zhang and Chen 2017), leaving other P2P market platforms under-represented in the literature. Our primary interest is establishing the validity of the PS models using data from a European lending platform Bondora.[2] However, to establish a

---

[2] As far as we are aware, only Byanjankar et al. (2015) used data from Bondora.

fair comparison across markets and validate our PS models, we also use a random sample of loans from the Lending Club marketplace.

The lending platform Bondora offers a database of loan characteristics and payments. In this paper, we show that credit-risk models can be improved. Instead, of modeling defaults, we model the annualized modified internal rate of returns calculated from the loan payments database. Our first loan starts on 21st February 2009 and ends on 11th November 2016. To focus on short- to mid-term loans, we remove loans that lasted less than 1 month or longer than 5 years.[3] We use only a sample of loans that are issued from Estonia or Finland and that had a Bondora rating version 2 available. After data pre-processing, our sample covers 161 explanatory variables and consists of 10,002 loans. Among which, 8001 were selected for training and 2001 for validation.

To match the size of the Bondora dataset, we used a random sample of loans from the much larger Lending Club database of finished loans. As before, 8001 loans were randomly selected to form the training and 2002 loans to form the testing dataset. The earliest loans were from 1st January 2013. Although all loans had a nominal maturity of 36 months, loans that had a real duration of less than 1 month (very early repayments) were removed from the dataset. After data pre-processing, our sample covers 142 explanatory variables.[4]

We used two versions of the training dataset (for Bondora and Lending Club data). For training PS models, where the internal rate of return is of concern, we used the raw datasets. For training the CS model, we address the imbalance that arises because of the under-representation of defaults in the training dataset (Table 1). Our approach is to use random under-sampling of the majority class (good loans) and random over-sampling (with replacements) of the minority class (non-performing, defaulted loans).

Data for both datasets were pre-processed in two ways. First, several non-negative numerical variables were skewed (to the right), which led us to apply the logarithmic transformation. Moreover, all categorical variables were transformed into dummy variables. Second, both datasets were subject to the following algorithm to address extremes, under-representation of classes and collinearity issues:

- For all numerical variables, the lowest and highest 0.1% were winsorized.
- For each dummy variable, we required at least 1% of event occurrences (i.e., either 1% or more one's or zero's).
- If any two variables had an absolute value of the Spearman's rank correlation coefficient higher than 0.95, one of the two variables was (randomly) removed.
- We checked whether exact linear multi-collinearity exists, and if yes, one of the variables was (randomly) removed.
- We ensure that the range of each variable in the testing dataset falls within the range of the same variable in the training dataset.

---

[3] We observed that using the profit performance measure sometimes leads to extremely high returns, which we traced to the fact that several loans were repaid early, thereby reducing the real maturity of the loan and making the annualized return unreasonably large. This case is also the idea behind removing loans from our sample that had a real duration of less than 1 month.

[4] A detailed list of explanatory variables for datasets and transformation is available upon request.

Table 1 Descriptive statistics of loan performance measures: modified internal rates of return

	Unit	Training dataset (8000 loans)					Testing dataset (2000 loans)				
		Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
Panel A: bondora dataset											
Good (non-defaulted) loans	Annualized %	27.06	27.66	7.49	-1.02	105.7	27.01	27.93	7.47	1.47	76.8
Defaulted loans	Annualized %	−43.98	−39.52	42.71	−100	59.56	−46.55	−41.59	43.42	−100	30.76
# Defaulted loans		# 1803 loans (22.5%)					# 486 (24.3%)				
Panel B: lending club dataset											
Good (non-defaulted) loans	Annualized %	13.08	12.25	4.31	4.68	42.68	13.23	12.39	4.54	4.92	37.65
Defaulted loans	Annualized %	−39.75	−36.31	32.38	−99.52	21.41	−34.78	−28.48	31.4	−99.3	13.41
# Defaulted loans		# 1577 loans (19.7%)					# 389 (19.5%)				

SD, Min, and Max standard deviation, minimum, and maximum of modified internal rates of return

Methodology

Loan performance measures

To distinguish between potentially performing and non-performing loans, the CS literature on the P2P market uses the standard default/not-default credit-risk framework. With that in mind, a loan is considered to perform well if all liabilities originating from the loan are repaid within a given payment schedule—on time (including the grace period). We denote the standard loan performance measure as follows:

$$P_{i,t_i}^{(1)} = \begin{cases} 0 & \text{Performing loan} \\ 1 & \text{Non - performing loan} \end{cases}, \tag{1}$$

where index i denotes the given loan, and $t = 1, 2, \ldots$ is the usual time index; we use t_i to denote the beginning of the ith loan contract on the specific day t. The CS models aim to estimate Eq. (1) for loan evaluation purposes. In this paper, Eq. (1) is based on the status of the loan as reported by the respective P2P lending platform.

Assume that the borrower receives the loan amount in a single payment at the beginning of the period denoted as CO_{i,t_i}, where, as before, index t_i highlights the fact that the loan amount is paid out at the beginning of the period. This case is also the same for all loans in our sample. The loans have different (nominal) maturities m_i (in days), and their real maturities can also differ from the nominal (agreed upon) date because of early repayment by the borrower. Over the given time period, one or more regular or irregular payments are received from the borrower by the investor. If all the payments are made on time, then the investor receives the loan amount plus the profit determined by the interest rate on the loan. As loan maturities differ, we assume that the investor has the possibility to re-invest received payments. In this way, we make loans with different real maturities comparable to each other in terms of their profitability. The future value is as follows:

$$CI_i = \sum_{t \in (t_i, t_i + m_i\rangle} CI_{i,t} \times \left(1 + R_{t_i}\right)^{\frac{t_i + m_i - t}{365}}, \tag{2}$$

where $CI_{i,t}$ are cash inflows over the period $t \in (t_i, t_i + m_i\rangle$, and R_{t_i} is a fixed reinvestment rate assumed to be known at the start of the loan. The investor's annualized return is calculated as follows:

$$P_{i,t_i}^{(2)} = \left(\left(\frac{CI_i}{CO_{i,t_i}} \right)^{\frac{365}{m_i}} - 1 \right) \times 100 [\%]. \tag{3}$$

Equation (3) is our second loan performance measure. The value of the return depends on how the reinvestment rate is estimated. The standard critique of using the reinvestment rate is based on two premises. The first is whether an investor even has the opportunity to re-invest incoming proceeds in investments with similar risks. The second is whether the opportunities offer returns comparable to the assumed return from the evaluated investment/loan. Most established P2P markets (Lending Club, Mintos, and Bondora) have sufficient liquidity to offer many similar loans. Therefore, we consider the reinvestment assumption to be valid. With regard to the value of the reinvestment rate, our approach is empirical and designed not to overestimate the overall return. In this

Lyócsa *et al. Financial Innovation* (2022) 8:32

Page 8 of 21

case, we use a return that was achieved in the past, which proceeds in the following two steps:

1. In the first step, we calculate Eq. (3) for all loans in our sample, assuming that $R_{t_i} = 0$, that is, no reinvestment rate. We denote the resulting returns as $P_{i,t_i}^{(*)}$.
2. In the second step, we calculate Eq. (3) for all loans, but now for each loan, the reinvestment rate R_{t_i} is equal to the following:

$$med\left(P_{j,(t_j+m_j)}^{(*)} : t_i - 365 \le t_j + m_j < t_i\right),$$

that is, the reinvestment rate R_{t_i} is the median value of $P_{j,t_j}^{(*)}$ calculated overall loans that finished $(t_j + m_j)$ in the past 365 days prior to the beginning of the evaluated ith loan. This approach ensures that our reinvestment rate is historical and tracks the improvement or worsening of the economic conditions of the borrowers. The rate is calculated over loans that have concluded and also include defaulted loans. However, this approach cannot be applied to initial loans. Instead of removing such loans, we use a zero reinvestment rate as a reinvestment rate.

Competing models

To show that modeling an investor's rate of return is a meaningful exercise, we perform a statistical and economic evaluation in an out-of-sample forecasting framework. We compare realized returns per loan and total profits of a hypothetical investor who is using either the standard CS model based on default predictions or the PS model based on the loan's return $\left(P_i^{(2)}\right)$ prediction.

The following sections describe the four classes of credit-risk models that we employ in this study: (1) linear regression-based regularization techniques (lasso, ridge, elastic net), (2) logistic-based regularization techniques, (3) RF, and (4) neural networks. We use regularization methods because they can be estimated quickly using conventional processing power and are also easy to interpret. We use RF and neural networks as these are standard ML models used in the P2P lending market literature.

Regularization in linear models

The lasso, ridge, and elastic net model estimates can be expressed as special cases of the following optimization problem (Tibshirani 1996; Zou and Hastie 2005):

$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2 + \lambda \left[\frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|\right], \tag{4}$$

where y_i is the ith loan performance measure, $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are $p \times 1$ column vectors of the *standardized* explanatory variables and coefficients, respectively. Parameter λ controls for the weight of the penalty term, and if $\lambda = 0$, the model breaks down to ordinary least squares. If $\alpha = 1$, the model breaks down to the lasso approach, $\alpha = 0$ leads to the ridge regression, and $0 < \alpha < 1$ is the elastic net approach. The key difference between the three models lies in how they handle correlated regressors. In the case of multiple correlated regressors, lasso tends to select one into the model at the expense of others; ridge

selects and reduces coefficients to a similar size, whereas the elastic net is a compromise between the two approaches. The combination of an $\alpha \in (0.1, 0.2, ...0.8, 0.9)$ and λ parameter is estimated. This case leads to the following four forecasts: LM (linear regression model), $LM^{\lambda_{\min}, \alpha=1}$, $LM^{\lambda_{\min}, \alpha=0}$, and $LM^{\lambda_{\min}, \alpha_{\min}}$.

Regularization in LR models

As before, we use penalization techniques adapted for the LR. The parameter estimates can be expressed as follows:

$$\min_{\beta_0, \boldsymbol{\beta}} - \left[\frac{1}{N} \sum_{i=1}^{N} y_i \left(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} \right) - \log \left(1 + e^{(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta})} \right) \right] + \lambda \left[\frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right]. \tag{5}$$

The only difference now is that y_i represents a default/no-default loan performance measure. Letting $\lambda = 0$ leads to the standard LR model, whereas $\alpha = 1$ leads to the logistic lasso, $\alpha = 0$ to the ridge lasso, and $0 < \alpha < 1$ to the elastic net lasso. Suitable parameters are found via tenfold cross-validation maximizing area under the curve (AUC). We end up with four forecasts: LR, $LR^{\lambda_{\min}, \alpha=1}$, $LR^{\lambda_{\min}, \alpha=0}$, and $LR^{\lambda_{\min}, \alpha_{\min}}$.

Random forest

As the name indicates, a random tree is a randomly constructed tree from a set of possible trees having K random features at each node. More formally, a RF classifier is a combination of tree-structured classifiers $\{h(x, \theta_k), k = 1\}$, where θ_k is the independent identically distributed random vectors (Breiman 2001). Once the trees are created, they vote on the most popular class. The specific steps followed in training the RF model both for the classification (modeling defaults) or regression tasks (modeling returns) in this work are listed below (Friedman et al. 2001):

- for $\{k = 1 : K\}$

  - Select a bootstrap sample Z from the training data set.
  - Build a RF to the sample Z by recursively repeating the following steps for each terminal node of the tree:

    - randomly choose m features from the total input space p;
    - select the best performing features and the best split points;
    - split the node.

  - Output: $\{T_k\}^K$

Having trained out the RF algorithm, we proceed with making a prediction for a new loan contract, x:

- Modeling the returns: $f_{rf}^K(x) = \frac{1}{K} \sum_{k=1}^{K} T_k(x)$.
- Modeling the defaults: Let $C_k(x)$ be the predicted loan status of the k tree. Then, $C_{rf}^K = $ majority vote.

The RF models need to be tuned using data from the training data set. Specifically, suitable values for *maximum tree depth* (3, 6, 9, and 12), *number of trees* (500, 1500, and 3000), and the *number of variables to possibly split* at in each node (5, 10, 15, 20, 25, 30, and 40) need to be determined. We use tenfold cross-validation and a grid search, where optimum parameters are those that minimized mean squared error (regression) or maximized AUC.

Neural networks

In addition to the RF, we also train feed-forward neural networks with a single hidden layer. Feed-forward networks have units that are one-way connected to our units, and they can be labeled from inputs to outputs so that each unit is only connected to units with higher numbers. A generic feed-forward network with one hidden layer can be represented by the following function (Ripley 2007):

$$y_k = f_k \left(\alpha_k + \sum_{j \to k} W_{jk} f_j \left(\alpha_j + \sum_{i \to j} W_{ij} X_i \right) \right). \tag{6}$$

Namely, to form the total input x_i, each unit summarizes its input and adds the bias. Consequently, to obtain the output y_i, we apply a function f_i to x_i. The connections from i to j have weights, $w_{i,j}$ which multiply the signal passing through the units. The inputs, on the other hand, have $f = 1$ as they only distribute the input. The neural network-based credit and PS system consists of two main steps: (1) data normalization and (2) model training and validation. In the first step, we re-scale the numerical variables into a range of [0,1]—process necessary for the neural network training and classification/evaluation. In the second phase, to specify the two hyperparameters, *size* (i.e., the number of units in the hidden layer) and *decay* (i.e., the regularization parameter to avoid overfitting), we employ tenfold cross-validation before using data from the training dataset.[5]

Notably, the literature offered many studies that aimed to classify loan applicants into creditworthy or not-creditworthy using artificial neural networks (Byanjankar et al. 2015; Moscato et al. 2021; Plawiak et al. 2020; Turiel and Aste 2019). However, in practice, this methodology is not used extensively. One highly relevant barrier for wider adoption of such ML models in CS in practice is related to the concept of explainability (Arrieta et al. 2020; Arya et al. 2019). Namely, ML solutions, such as neural networks, are often referred to as black boxes because, typically, tracing the steps that the algorithm took to arrive at its decision is difficult. This challenge is particularly relevant for P2P platforms, which in the attempt to offer cheap administration of loans through automatized scoring, are subjected to the General Data Protection Regulation (GDPR). GDPR provides a right to explanation, thereby enabling users to ask for an explanation as to the decision-making processes affecting them.

---

[5] For the Bondora dataset, the optimum size parameter is 5 (CS model) and 1 (PS model), whereas the decay parameter in both cases was found to be 0.1. For the Lending Club dataset, the optimum size parameter was again 5 (CS model) and 1 (PS model), with decay set to 0 and 10^{-4}.
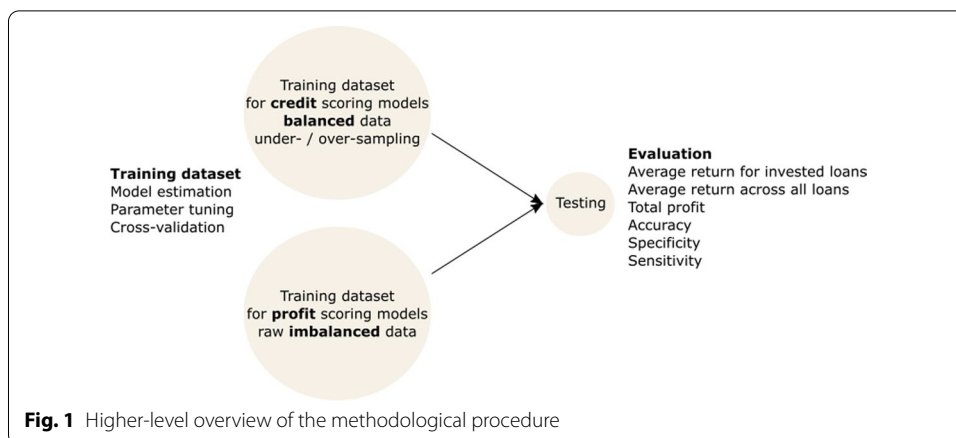
Fig. 1 Higher-level overview of the methodological procedure

Forecasting procedure

Our forecasting procedure follows a standard procedure found in the P2P literature as we randomly divide our sample of loans into training (80%) and testing (20%) datasets. In the first step, using all loans from the training dataset, we estimate and fine-tune (via cross-validation) predictive models. In the second step, given estimated model parameters and characteristics of loans in the training dataset, we predict those loans' performance measures $\left(P_{i,t_i,r}^{*\{1,2\}}\right)$. Figure 1 shows the procedure.[6]

Simply having predicted loan performance measures is not enough to decide whether to invest or not in the given loan. For example, if the LR model for the *i*th loan estimates the probability of default to be $P_{i,t_i,r}^{*\,(1)} = 0.234$, should the investor invest? A similar question arises for models explaining loan returns. For example, if the LM model for the *i*th loan estimates the return to be $P_{i,t_i,r}^{*\,(2)} = 15.24\%$, should the investor make an investment? For both types of predictions, suitable threshold values are needed. For CS models, our threshold is $TR_{r,i}^{CS} = 0.50$ as we are using a balanced dataset (see the "Data" section), that is, $P_{i,t_i,r}^{*\,(1)} > TR_{r,i}^{CS} = 0.5$ are predicted to default (non-performing). For PS models, we use the raw dataset, and our threshold is naturally $TR_{r,i}^{PS} = 0.00\%$, that is, loans with a negative predicted modified internal rate of returns $P_{i,t_i,r}^{*\,(2)} < TR_{r,i}^{PS} = 0\%$ are predicted to default (non-performing).

Performance evaluation

Returns and profit measures

Our choice to prefer returns and profits as performance measures is motivated by the fact that CS and PS models cannot be compared via the ROC curve and AUC measure. Moreover, in any real-life scenario, P2P platform operators and investors need to evaluate the credit-risk model by using a single threshold as opposed to a range of possible values. Average returns and total profit are economic measures that provide a direct and fair comparison between CS and PS models. To assess the performance of each model *r* and loan *i*, we use the return:

---

[6] Bastani et al. (2019) presented an alternative approach by combining the CS and PS models into a two-stage sequential approach. We do not opt for this approach as the one-stage PS models lead to higher average returns and total profits. However, their approach deserves attention in future studies.

$$R_{r,i} = S(r,i) \times P_{i,t_i}^{(2)}, \tag{7}$$

where in the case of PS models, $S(r,i)$ is the signaling function:

$$S(r,i) = \begin{cases} 1 & \text{if } TR_{r,i}^{PS} \leq P_{i,t_i,r}^{*\,(2)} \\ 0 & \text{if } TR_{r,i}^{PS} > P_{i,t_i,r}^{*\,(2)} \end{cases}, \tag{8}$$

which returns 1 if the predicted performance measure exceeds the optimum threshold (the case for modeling defaults, that is, using $TR_{r,i}^{CS}$ is analogous) and 0 if otherwise. The mean return over all loans is as follows:

$$R_r^{all} = NF^{-1} \sum_{i=1}^{NF} R_{r,i}. \tag{9}$$

Irrespective of the predictive model *r*, we use the realized return $P_{i,t_i,r}^{(2)}$ to evaluate the loan. Very conservative models might be highly successful in predicting positive returns correctly. However, from another aspect, these models might recommend investing in only a handful of loans. To remedy this effect, the mean return, as defined above, gives 0% return for loans, for which investment is not recommended.

Although the average return across all loans is of interest to the P2P platform provider, regulators, and supervisory institutions, investors might be interested in the realized returns across invested loans. We, therefore, report mean return only over-invested loans:

$$R_r^{inv} = |\{\forall i : S(r,i) = 1\}|^{-1} \sum_{\{\forall i : S(r,i)=1\}} R_{r,i}, \tag{10}$$

where |.| denotes the cardinality of a set.[7]

Next, we report the standard deviation of the returns as follows:

$$SD_r^{all} = \left[NF^{-1} \sum_{i=1}^{NF} \left(R_{r,i} - R_r^{all} \right)^2 \right]^{0.5}, \tag{11}$$

SD_r^{inv} is defined accordingly.

We also report the total nominal profit, that is, the difference between the sums of money inflows minus loan amount without assuming any reinvestment, which is realized across only invested loans $\left(R_r^{profit} \right)$. In this way, we directly consider the importance of the prediction, which should be higher in the case of larger loans because the consequences of imprecise predictions are larger. More specifically:

$$R_r^{profit} = \sum_{\{\forall i : S(r,i)=1\}} S(r,i) \times \left(CI_i - CO_{i,t_i} \right). \tag{12}$$

---

[7] Interestingly, establishing what type of returns is usually reported in the literature is difficult, which is surprising given that the values tend to be quite different.

Statistical evaluation

The model that leads to the highest returns might not always be superior to other models. Differences across the performance of models might be just an artifact of the inherent high uncertainty regarding the outcomes of P2P loans. Therefore, we formally compare the realized return [Eq. (9)] of all 12 models: six models that predict returns and six models that predict the probability of default. To account for multiple pairwise comparisons and possible data snooping, we use the model confidence set (MCS) of Hansen et al. (2011).

In our setting, we define the loss function for a given model and a loan to be $l_{i,r} = -R_{i,r}$, that is, the higher the return, the lower the loss of a given model. Next, the difference between the losses of models *m,n* is as follows:

$$d_{m,n} = l_{i,m} - l_{i,n}; \quad m, n = 1, 2, \ldots, \ i = 1, 2, \ldots, NF. \tag{13}$$

The equal predictive ability (EPA) hypothesis is as follows:

$$H_0 : E[d_{m,n}] = 0; \quad \forall m, n \ and \ H_1 : E[d_{m,n}] \neq 0; \quad \text{for some } m, n.$$

We use the T_{MAX} statistics of Hansen et al. (2011) to test the above hypothesis, where the distribution under the null hypothesis is derived using a bootstrapping procedure with 5000 bootstrap samples. As indicated above, the MCS procedure is a sequence of EPA tests, where we start with a set of all 12 models and perform the above test. If the null is not rejected, the superior set consists of all models. If the null is rejected, we remove the worst performing model and continue with the EPA test on the remaining 11 models. The procedure continues until the null is not rejected or only one model is left. The α that indicated the confidence level is set to 0.10. The higher the α, the lower the confidence level, and more models tend to be selected in the superior set of models.

Data pre-processing, sampling (under- and over-sampling), statistical model estimation, and evaluation is performed via program R (Wickham et al. 2021; Wallig et al. 2020a, 2020b; Bernardi and Catania 2018; Gorman 2018; Friedman et al. 2010; Kuhn et al. 2008; Ripley, 2007; Wright and Ziegler 2015). Scripts performing the predictions and evaluation of loans are available as Additional file 1.

Results

Overview of performance loan measures

In Fig. 2 (Bondora left, Lending Club right panels), we highlight loans with negative (modified internal rates of returns, returns henceforth) returns ($P^{(2)}$). Similar figures are reported by Serrano-Cinca and Gutiérrez-Nieto (2016) and Bastani et al (2019). The loan returns are skewed to the left, as negative returns are possible and returns near -100% occur often. This case also leads to a large variation in returns. In Table 1 (training dataset), we report the default rate of 22.5% (Bondora) and 19.7% (Lending Club) and the average annualized returns stratified for profitable and nonprofitable returns.

The results show that fishing for good loans might be very lucrative as the return is 27.06% for profitable loans on Bondora and 13.08% for Lending Club. Given the low interest rate environment in respective countries over the given sample period, these numbers suggest why investing in the P2P market is attractive for many investors. On
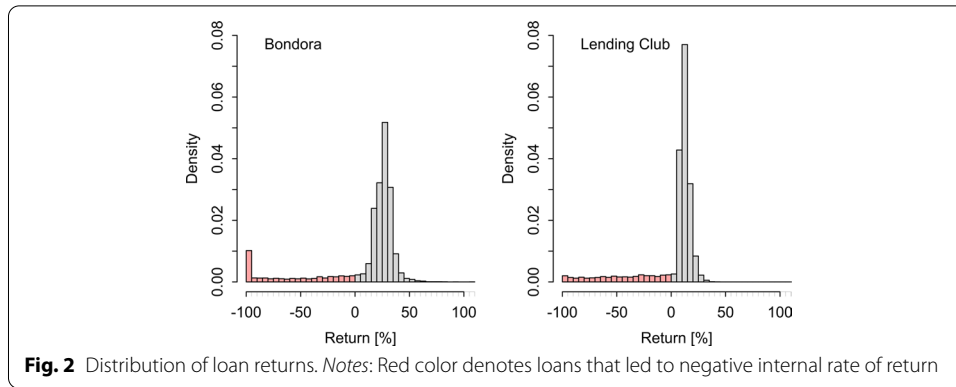
Fig. 2 Distribution of loan returns. *Notes*: Red color denotes loans that led to negative internal rate of return

Table 2 Out-of-sample performance of credit and profit scoring models—Bondora

Model	% invested loans	Performance across invested loans		Performance across all loans		Total profit in mil. EUR
		Average return	SD	Average return	SD	
Panel A: Credit scoring models						
LR	59.8	19.57	33.06	11.70	27.30	0.53
$LR^{\lambda_{min},\alpha=1}$	59.9	19.72	32.98	11.82	27.30	0.54
$LR^{\lambda_{min},\alpha=0}$	59.3	19.69	32.96	11.69	27.17	0.53
$LR^{\lambda_{min},\alpha_{min}}$	59.7	19.58	33.17	11.70	27.37	0.53
RFC	75.8	19.49	38.48	14.78	34.53[†]	0.69
NNC	60.6	17.61	36.85	10.66	29.93	0.47
Panel B: Profit scoring models						
LM	75.7	19.34	39.76	14.65	35.58	0.66
$LM^{\lambda_{min},\alpha=1}$	76.3	19.66	39.54	15.00	35.53[†]	0.70
$LM^{\lambda_{min},\alpha=0}$	76.2	19.69	39.29	15.02	35.32[†]	0.70
$LM^{\lambda_{min},\alpha_{min}}$	76.2	19.67	39.55	15.00	35.53[†]	0.70
RFR	75.4	20.55	38.80	15.49	34.83[†]	0.72
NNR	71.8	20.35	38.82	14.62	34.15[†]	0.68

$LR^{\lambda_{min},\alpha=1}$, $LR^{\lambda_{min},\alpha=0}$, $LR^{\lambda_{min},\alpha_{min}}$ are lasso, ridge and elastic net versions of logistic regression, $LM^{\lambda_{min},\alpha=1}$, $LM^{\lambda_{min},\alpha=0}$, $LM^{\lambda_{min},\alpha_{min}}$ are lasso, ridge and elastic net versions of linear regression

SD standard deviation, *LR* logistic regression-based models, *LM* linear regression-based models, *RFC* random forest classification, *NNC* neural network classification, *RRR* denotes random forest regression, *NNR* neural network regression

[†] Denotes a model that belongs to the set of superior models

the other hand, for nonprofitable loans, the return is -43.98% (Bondora) and -39.75% (Lending Club). Comparing loans from the two markets shows that our sample of loans from Bondora is riskier. That is, the sample offers higher returns for profitable loans but also lower returns for non-performing loans. Moreover, the standard deviations are larger for returns realized on Bondora as opposed to Lending Club loans, for performing (7.49% vs. 4.31%) and non-performing loans (41.71% vs. 32.38%).

Credit or profit scoring?

Tables 2, 3, 4 and 5 show the results from the individual CS and PS models. We report results for the Bondora loans in Tables 2 and 3 and results for the Lending Club loans in

Table 3 Loan default classification comparison of credit and profit scoring models—Bondora

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)
Panel A: Credit scoring models			
LR	71.2	73.5	70.5
$LR^{\lambda_{min},\alpha=1}$	71.4	73.5	70.7
$LR^{\lambda_{min},\alpha=0}$	70.9	73.7	70.0
$LR^{\lambda_{min},\alpha_{min}}$	71.0	73.0	70.3
RFC	78.8	56.2	86.1
NNC	67.9	65.0	68.8
Panel B: Profit scoring models			
LM	76.4	51.4	84.5
$LM^{\lambda_{min},\alpha=1}$	76.9	51.2	85.2
$LM^{\lambda_{min},\alpha=0}$	77.1	51.9	85.3
$LM^{\lambda_{min},\alpha_{min}}$	76.8	51.2	85.1
RFR	77.4	54.1	84.9
NNR	75.5	57.6	81.3

$LR^{\lambda_{min},\alpha=1}$, $LR^{\lambda_{min},\alpha=0}$, $LR^{\lambda_{min},\alpha_{min}}$ are lasso, ridge and elastic net version of logistic regression, $LM^{\lambda_{min},\alpha=1}$, $LM^{\lambda_{min},\alpha=0}$, $LM^{\lambda_{min},\alpha_{min}}$ are lasso, ridge and elastic net version of linear regression

RFC random forest classification, *NNC* neural network classification, *RRR* random forest regression, *NNR* neural network regression

Table 4 Out-of-sample performance of credit and profit scoring models—lending club

Model	% invested loans	Performance across invested loans		Performance across all loans			Total profit in mil. USD
		Average return	SD	Average return	SD	MCS	
Panel A: Credit scoring models							
LR	60.8	8.66	18.20	5.27	14.81		1.42
$LR^{\lambda_{min},\alpha=1}$	61.0	8.53	18.28	5.20	14.87		1.39
$LR^{\lambda_{min},\alpha=0}$	61.4	8.57	18.15	5.26	14.81		1.40
$LR^{\lambda_{min},\alpha_{min}}$	61.2	8.49	18.38	5.19	14.96		1.39
RFC	78.5	7.46	21.46	5.86	19.26		1.62
NNC	70.1	8.21	20.31	5.75	17.42		1.58
Panel B: Profit scoring models							
LM	68.6	8.52	20.63	5.84	17.54		1.61
$LM^{\lambda_{min},\alpha=1}$	72.5	8.94	20.43	6.48	17.85[†]		1.84
$LM^{\lambda_{min},\alpha=0}$	72.2	8.69	21.12	6.27	18.35[†]		1.79
$LM^{\lambda_{min},\alpha_{min}}$	72.8	8.93	20.42	6.50	17.87[†]		1.84
RFR	73.3	8.47	21.69	6.21	18.95[†]		1.82
NNR	80.1	7.83	22.76	6.27	20.61[†]		1.79

$LR^{\lambda_{min},\alpha=1}$, $LR^{\lambda_{min},\alpha=0}$, $LR^{\lambda_{min},\alpha_{min}}$ are lasso, ridge and elastic net version of logistic regression, $LM^{\lambda_{min},\alpha=1}$, $LM^{\lambda_{min},\alpha=0}$, $LM^{\lambda_{min},\alpha_{min}}$ are lasso, ridge and elastic net version of linear regression

SD standard deviation, *LR* logistic regression-based models, *LM* linear regression-based models, *RFC* random forest classification, *NNC* neural network classification, *RRR* random forest regression, *NNR* neural network regression

[†] Denotes a model that belongs to the set of superior models

Table 5 Loan default classification comparison of credit and profit scoring models—lending club

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)
Panel A: Credit scoring models			
LR	67.5	67.2	67.6
$LR^{\lambda_{\min},\alpha=1}$	67.5	66.7	67.7
$LR^{\lambda_{\min},\alpha=0}$	67.9	66.7	68.1
$LR^{\lambda_{\min},\alpha_{\min}}$	67.6	66.4	67.9
RFC	75.7	42.6	83.7
NNC	72.1	55.1	76.2
Panel B: Profit scoring models			
LM	69.1	51.3	73.4
$LM^{\lambda_{\min},\alpha=1}$	72.4	49.7	77.9
$LM^{\lambda_{\min},\alpha=0}$	71.4	47.9	77.0
$LM^{\lambda_{\min},\alpha_{\min}}$	72.6	49.5	78.1
RFR	72.0	46.4	78.1
NNR	73.8	33.8	83.5

$LR^{\lambda_{\min},\alpha=1}$, $LR^{\lambda_{\min},\alpha=0}$, $LR^{\lambda_{\min},\alpha_{\min}}$ are lasso, ridge and elastic net version of logistic regression, $LM^{\lambda_{\min},\alpha=1}$, $LM^{\lambda_{\min},\alpha=0}$, $LM^{\lambda_{\min},\alpha_{\min}}$ are lasso, ridge and elastic net version of linear regression
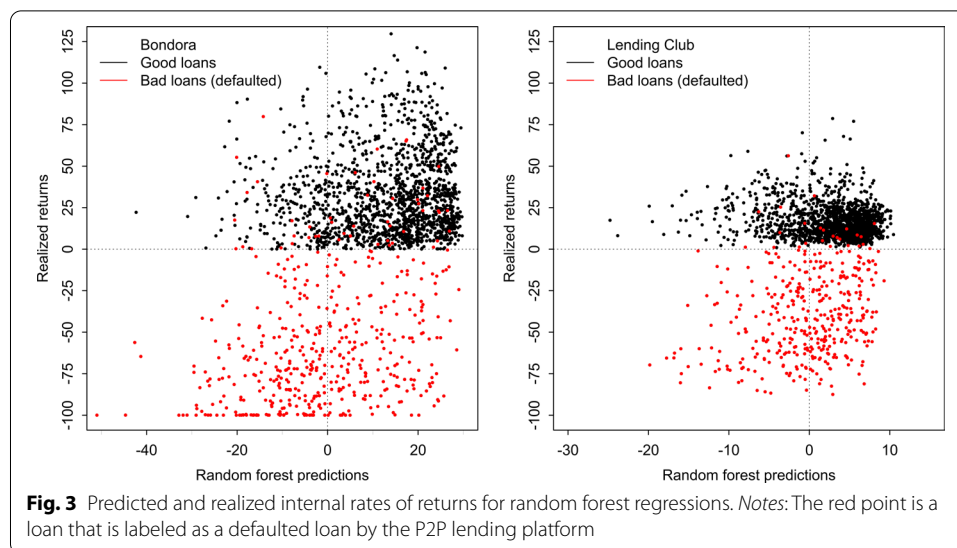
RFC random forest classification, *NNC* neural network classification, *RRR* random forest regression, *NNR* neural network regression.

Tables 4 and 5. For example, the value of 59.8% in the first row of Table 2 corresponds to the percentage of loans, which were identified by the forecasting model in the given row as one where an investment can be made, that is, $S(r, i) = 1$. The average annualized return per invested loan using LR is 19.57%, with a considerable standard deviation of 33.06. Notably, for PS models, the returns are very similar but they have a higher standard deviation. Although this case might suggest superiority of CS models, this comes at a price of a much lower percentage of invested loans, which for PS jumps to more than 75%. An exception is the RF CS model that performs similarly to the PS RF model. This model is the only CS model that also leads to returns that belong to the superior set of models, as indicated by the Hansen et al.'s (2011) test.

As CS models tend to overestimate risks (which leads to the low percentage of invested loans), unsurprisingly, the average return across all loans is much lower, as opposed to the PS models, from around 11% for CS models to around 15% for PS models. Opposed to CS models, returns are only 3.1% higher for PS models across invested loans but much higher, by 24.0%, across all loans. Is the higher number of invested loans worth the effort? The total profit measure[8] suggests that it is as it leads to absolute profits (the last column in Table 2) that are approximately 26.7% higher for PS models.

In Table 3, we report accuracy, specificity, and sensitivity across CS and PS models for the Bondora loans. An interesting observation is that default scoring models tend to be more accurate at predicting good loans (specificity) at the expense of predicting non-performing loans (sensitivity). The overall accuracy of PS is however higher, on average by 6.7%, and ranges from 75.5% (neural network model) to 77.4% (RF). For CS, the range starts from 67.9% (neural network model) to 78.8% (RF). As before, among CS models,

---

[8] Which are non-reinvested interest payments.

Fig. 3 Predicted and realized internal rates of returns for random forest regressions. *Notes*: The red point is a loan that is labeled as a defaulted loan by the P2P lending platform

RF clearly stands out, matching the RF regression, the best PS model. However, prior to any analysis, which model will perform the best is unclear. Given this model choice uncertainty, our results suggest that, in general, PS is overwhelmingly the preferred choice in modeling P2P credit risk on the Bondora loan market. The reason is that only one among CS models is able to match results from individual PS models.

Having established our key results for the Bondora dataset, we report performance results for the Lending Club dataset in Tables 4 and 5. Previous research (e.g., Bastani et al. 2019; Serrano-Cinca and Gutiérrez-Nieto 2016) already established the superiority of PS models for Lending Club loans. However, whether these results hold for regularization methods, RF, and neural network models is unclear. In several ways, results in Tables 4 and 5 are similar to those for the Bondora dataset. The average return across invested loans is in the range of 7.46% (RF) to 8.66% (LR) for CS models, whereas it ranges from 7.83% (neural networks) to 8.94% (lasso regression) for PS models. On average, PS outperforms CS models by 2.9%—close to what we observed with the Bondora dataset. When we turn our attention to average returns across all loans, the gap between CS (from 5.19% for an elastic net to 5.86% for RF) and PS (from 5.84% linear regression to 6.50% elastic net) models widens to 15.5% in favor of PS models. These results are also supported via the MCS that includes only five PS models (all except linear regression) and total profit that is 21.5% higher for PS models.

A closer inspection of our results shows (Tables 3 and 5) that the gains from PS are achieved in the ability of PS models to better identify bad loans, that is, they have higher specificity. On the contrary, CS models are better at identifying good loans. The overall accuracy is 3.1% higher for PS models. Moreover, notably, as before, RF stands out among CS models achieving a high level of accuracy that even surpasses that of PS models. To visualize the accuracy of the RF model's accuracy (generally best performing model), Fig. 3 plots predicted and realized returns. In Fig. 3, we can also observe that several loans are labeled as defaulted (red dot), whereas their realized return was positive. This case happens if the borrower has not met his or her obligations although he or she has paid off most of his or her loans and interest.

Investors' perspective typically differs from that of a lending platform as the former is solely focused on profitable investments. To mimic the cherry-picking behavior of an investor, Serrano-Cinca and Gutiérrez-Nieto (2016) and Bastani et al (2019) reported internal rates of return for loans with the highest 100 (former) or 30 (later) predicted scores (the Lending Club loan market). In the former case, the linear regression led to a return of 11.92%, whereas the CHAID[9] led to a 5.98% return across 100 loans. In the latter case, the one-stage methodology produced returns in a range of 9.4% (wide learning) to 13.4$ (wide and deep learning). Moreover, the two-stage methodology led to a range of 12.8% (wide learning) to 16.4% (wide and deep learning) across 30 loans. Our results are similar as we achieved 10.92% for top 100 loans and 12.91% for top 30 loans on the Lending Club loan market using the best PS approach—RF regression. With the CS approach, the most that we could hope for was an 8.05% and 7.84% return with the RF classification algorithm. The same strategy on the Bondora loan market would lead to higher returns for the investor. The best CS (RF classification) led to 9.65% for the top 30 and 10.95% for the top 100 loans. PS models fared better here as well, with much higher returns at 28.95% for the top 100 and 28.80% for top 30 loans (RF regression). These results show that PS models are specifically suited for budget-constrained risk-maximizing investors who have to select a certain number of loans.

An important aspect of investing is the trading activity governed by the given credit-risk model, which is of concern to investors and P2P market providers alike. As already noted before, sizable differences exist, as PS models suggest investment in 71.8% (neural networks) to 76.3% (lasso) of cases. On the contrary, CS models in only 59.3% (ridge) to 75.8% (an exception of RF) of cases. On average, a difference of 20.4% exists in favor of PS models. A similar pattern is observed for the Lending Club loan market with an increase in trading activity by 11.8% for PS models.

To summarize, the benefits of using PS models are higher overall returns, accuracy, and trading activity, whereas returns across invested loans are similar to those of CS models. These results hold for the Bondora and Lending Club loan markets.

Conclusion

In the past decade, the emergence of novel P2P lending has led to new challenges for investors, risk managers, and regulators. For the industry to thrive, its credit-risk models should be improved. The technology can serve as an intermediary between the lender and borrower in a market of consumer loans. In this paper, we present empirical results from PS models that help decision-makers, investors, and operators of P2P platforms to manage risky loans better. In doing so, we provide new and supporting evidence that PS models tend to outperform default scoring models.

We use data on loan and loan payments from Bondora, a European P2P platform that facilitates short-term risky loans between borrowers and lenders, including data from the well-known Lending Club marketplace. Using regularization methods (lasso, ridge, elastic net) in linear and LRs and RF and neural networks, our empirical results suggest that modeling the adjusted internal rate of returns leads to much higher returns

---

[9] Chi-square automatic interaction detection algorithm.

(across all loans) and profits compared with modeling loan defaults. Our results contribute to the existing literature on credit-risk models for P2P markets by showing how to significantly improve the risk management of P2P loans. Consequently, the improved risk management of P2P loans might fuel the growth of the P2P market. However, to be able to use PS models in the first place, P2P platforms should strive for transparency in providing data on loan payments for past and existing loans, a practice that is still not an industry standard.

Abbreviations
CS: Credit scoring; EN: Elastic net; EPA: Equal predictive ability; IRR: Internal rate of return; MCS: Model confidence set; LR: Logistic regression model; MCMC: Markow chain Monte Carlo; OLS: Ordinary least squares; P2P: Peer-to-peer or person-to-person; PS: Profit scoring; US: United states; RF: Random forest; RFC: Random forest classification; RFR: Random forest regression; SD: Standard deviation; NNC: Neural network classification; NNR: Neural network regression; ML: Machine learning; GDPR: General data protection regulation; AUC: Area under the curve; ROC: Receiver operating characteristic.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s40854-022-00338-5.

> **Additional file 1.** Data and scripts performing the predictions and evaluation of loans in R.

Author's contribution
All authors contributed equally, read and approved the final manuscript.

Declaration

Competing interests
All authors declare no competing interest.

Author details
[1]Faculty of Management and Business, University of Prešov, Prešov, Slovak Republic. [2]Institute of Economic Research, Slovak Academy of Sciences, Bratislava, Slovak Republic. [3]ZHAW School of Engineering, Winterthur, Switzerland. [4]Department of Banking and International Finance, Faculty of National Economy, University of Economics in Bratislava, Bratislava, Slovak Republic. [5]Faculty of Economics and Administration, Masaryk University, Lipová 507/41a, 602 00 Brno, Czech Republic.

References
Ahelegbey DF, Giudici P, Hadji-Misheva B (2019) Factorial network models to improve P2P credit risk management. Available at SSRN 3349001
Allen F, Gu X, Jagtiani J (2021) A survey of fintech research and policy discussion. Review of Corporate Finance. Forthcoming
Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115
Arya V, Bellamy RK, Chen PY, Dhurandhar A, Hind M, Hoffman SC, Houde S, Liao QV, Luss R, Mojsilović A et al (2019) One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. arXiv preprint https://arxiv.org/abs/1909.03012
Balyuk T (2019) Financial innovation and borrowers: evidence from peer-to-peer lending. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2802220
Bastani K, Asgari E, Namavari H (2019) Wide and deep learning for peer-to-peer lending. Expert Syst Appl 134:209–224
Bernardi M, Catania L (2018) The model confidence set package for r. Int J Comput Econ Econom 8(2):144–158
Breiman L (2001) Random forests. Mach Learn 45(1):5–32
Byanjankar A, Heikkilä M, Mezei J (2015) Predicting credit risk in peer-to-peer lending: a neural network approach. In: 2015 IEEE symposium series on computational intelligence, vol 57, no 5. pp 719–725

Cheng HT, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, Anderson G, Corrado G, Chai W, Ispir M et al (2016) Wide and deep learning for recommender systems. In: Proceedings of the 1st workshop on deep learning for recommender systems. ACM, pp 7–10

De Roure C, Pelizzon L, Thakor AV (2021) P2P lenders versus banks: cream skimming or bottom fishing? Available at SSRN https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3174632

Duarte J, Siegel S, Young L (2012) Trust and credit: the role of appearance in peer-to-peer lending. Rev Financ Stud 25(8):2455–2484

Emekter R, Tu Y, Jirasakuldech B, Lu M (2015) Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. Appl Econ 47(1):54–70

Friedman J, Hastie T, Tibshirani R et al (2001) The elements of statistical learning, vol 1. Springer series in statistics. Springer, New York

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1

Ge R, Feng J, Gu B (2016) Borrower's default and self-disclosure of social media information in P2P lending. Financ Innov 2(1):30

Giudici P, Misheva BH (2018) P2P lending scoring models: Do they predict default? J Digit Bank 2(4):353–368

Gorman B (2018) mltools: Machine learning tools. https://cran.r-project.org/web/packages/mltools/index.html Accessed 11 July 2021

Guo Y, Zhou W, Luo C, Liu C, Xiong H (2016) Instance-based credit risk assessment for investment decisions in P2P lending. Eur J Oper Res 249(2):417–426

Hansen PR, Lunde A, Nason JM (2011) The model confidence set. Econometrica 79(2):453–497

Jagtiani J, Lemieux C (2018) Do fintech lenders penetrate areas that are underserved by traditional banks? J Econ Bus 100:43–54

Jagtiani J, Lemieux C (2019) The roles of alternative data and machine learning in fintech lending: evidence from the lending club consumer platform. Financ Manag 48(4):1009–1029

Jagtiani J, Lambie-Hanson L, Lambie-Hanson T (2021) Fintech lending and mortgage credit access. J FinTech 1(01):2050004

Jin Y, Zhu Y (2015) A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending. In: 2015 Fifth international conference on communication systems and network technologies. IEEE, pp 609–613

Kim A, Cho SB (2019a) An ensemble semi-supervised learning method for predicting defaults in social lending. Eng Appl Artif Intell 81:193–199

Kim JY, Cho SB (2019b) Predicting repayment of borrows in peer-to-peer social lending with deep dense convolutional network. Expert Syst 36:e12403

Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Inf Sci 275:1–12

Kou G, Akdeniz ÖO, Dinçer H, Yüksel S (2021a) Fintech investments in European banks: a hybrid IT2 fuzzy multidimensional decision-making approach. Financ Innov 7(1):1–28

Kou G, Xu Y, Peng Y, Shen F, Chen Y, Chang K, Kou S (2021b) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decis Support Syst 140(113):429

Kuhn M et al (2008) Building predictive models in r using the caret package. J Stat Softw 28(5):1–26

Li W, Ding S, Chen Y, Yang S (2018) Heterogeneous ensemble for default prediction of peer-to-peer lending in china. IEEE Access 6:54396–54406

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

Li T, Kou G, Peng Y, Philip SY (2021) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2021.3109066

Liang K, He J (2020) Analyzing credit risk among Chinese P2P-lending businesses by integrating text-related soft information. Electron Commer Res Appl 40(100):947

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

Miller S (2015) Information and default in consumer credit markets: evidence from a natural experiment. J Financ Intermed 24(1):45–70

Moscato V, Picariello A, Sperlí G (2021) A benchmark of machine learning approaches for credit score prediction. Expert Syst Appl 165:113986

Pławiak P, Abdar M, Pławiak J, Makarenkov V, Acharya UR (2020) Dghnl: a new deep genetic hierarchical network of learners for prediction of credit scoring. Inf Sci 516:401–418

Ripley BD (2007) Pattern recognition and neural networks. Cambridge University Press, Cambridge

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. Decis Support Syst 89:113–122

Serrano-Cinca C, Gutiérrez-Nieto B, López-Palacios L (2015) Determinants of default in P2P lending. PLoS ONE 10(10):e0139427

Sjoblom M, Castello A, Gadzinski G et al (2019) Profitability vs. credit score models—a new approach to short term credit in the UK. Theor Econ Lett 9(04):1183

Tang H (2019) Peer-to-peer lenders versus banks: Substitutes or complements? Rev Financ Stud 32(5):1900–1938

Teply P, Polena M (2020) Best classification algorithms in peer-to-peer lending. N Am J Econ Finance 51(100):904

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (methodol) 58(1):267–288

Turiel JD, Aste T (2019) P2P loan acceptance and default prediction with artificial intelligence. arXiv preprint https://arxiv.org/abs/1907.01800

Wallig M, Microsoft, Weston S (2020a) Foreach: provides foreach looping construct. https://cran.r-project.org/web/packages/foreach/ Accessed 11 July 2021

Lyócsa *et al. Financial Innovation*   (2022) 8:32   Page 21 of 21

Wallig M, Microsoft Corporation, Weston S, Tenenbaum D (2020b) doParallel: Foreach Parallel adaptor for the 'parallel' package. https://cran.r-project.org/web/packages/doParallel/index.html. Accessed 11 July 2021

Wang Z, Jiang C, Ding Y, Lyu X, Liu Y (2018) A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. Electron Commer Res Appl 27:74–82

Wang Z, Jiang C, Zhao H, Ding Y (2020) Mining semantic soft factors for credit risk evaluation in peer-to-peer lending. J Manag Inf Syst 37(1):282–308

Wickham H, François R, Henry L, Müller K (2021) dplyr: a grammar of data manipulation. https://cran.r-project.org/web/packages/dplyr/ Accessed 11 July 2021

Wright MN, Ziegler A (2015) Ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint https://arxiv.org/abs/1508.04409

Xia Y, Liu C, Li Y, Liu N (2017a) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl 78:225–241

Xia Y, Liu C, Liu N (2017b) Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electron Commer Res Appl 24:30–49

Xia Y, Liu C, Da B, Xie F (2018) A novel heterogeneous ensemble credit scoring model based on bstacking approach. Expert Syst Appl 93:182–199

Xu D, Zhang X, Feng H (2019) Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model. Int J Finance Econ 24(2):903–921

Ye X, La D, Ma D (2018) Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score. Electron Commer Res Appl 32:23–36

Zhang K, Chen X (2017) Herding in a P2P lending market: Rational inference or irrational trust? Electron Commer Res Appl 23:45–53

Zhang J, Liu P (2012) Rational herding in microloan markets. Manag Sci 58(5):892–912

Zhang Y, Li H, Hai M, Li J, Li A (2017) Determinants of loan funded successful in online P2P lending. Procedia Comput Sci 122:896–901

Zhang W, Wang C, Zhang Y, Wang J (2020) Credit risk evaluation model with textual features from loan descriptions for P2P lending. Electron Commer Res Appl 42(100):989

Zhou J, Li W, Wang J, Ding S, Xia C (2019) Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A Stat Mech Appl 534(122):370

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B (stat Methodol) 67(2):301–320

Publisher's Note