Predicting Expected Profit in Ongoing Peer-to-Peer Loans with Survival Analysis-Based Profit Scoring

Byanjankar, Ajay; Viljanen, Markus

Published in:
Intelligent Decision Technologies 2019

DOI:
https://dx.doi.org/10.1007/978-981-13-8311-3_2

Published: 01/01/2019

Document Version
Accepted author manuscript

Document License
Publisher rights policy

Predicting Expected Profit in On-Going Peer-to-Peer Loans with Survival Analysis Based Profit Scoring

Ajay Byanjankar[1] and Markus Viljanen[2]

[1] Åbo Akademi University, Turku, Finland
ajay.byanjankar@abo.fi
[2] University of Turku, Turku, Finland

Abstract. The growing popularity of P2P lending has attracted more borrowers and lenders to the sector. With the growth in the popularity of P2P lending there have been many studies focusing on analyzing credit risk in P2P lending. However, the credit risk is only a part of the story. The higher interest rates are allocated to the riskier loans, and the higher interest rates may or may not in fact compensate for the defaults expected. Therefore, the profit of a loan depends on both the interest rate and the default probability. Since investors are ultimately concerned with return on investment, models should help investors to predict the profit as accurately as possible. We develop a model that predicts the expected profit of a loan using survival analysis based monthly default probability. Our approach extends previous profit scoring approaches, since it can be applied to any loan data set, including current data sets with many on-going loans.

Keywords: P2P Lending, Credit risk, Survival analysis, Profit Scoring

1 Introduction

Peer-to-peer (P2P) lending is an online micro financing solution that helps to match lenders and borrowers without any financial intermediaries and collateral. The whole process of lending and borrowing take place over the internet facilitated by a P2P lending platform. Borrowers make loan application with their financial and demographic information. The approved borrowers' list with their information is then made available to the lenders for investment. Lenders can then select borrowers from the list to invest and spread the investment to multiple borrowers.

Several benefits have been proclaimed to favor P2P lending. The cost and ease of borrowing could potentially be reduced by the automation of the lending process, the unbundling of unnecessary services and the disintermediation of financial institutions [1, 2]. The borrowers get attracted to the platforms for easy and quick access to credit, which has contributed to the rapid growth in P2P lending. The lenders are motivated towards P2P lending due to higher return compared to similar traditional investments.

However, the return may not always be as high as advertised since P2P lending is equally exposed to financial risk. The credit risk associated to P2P lending is mostly focused on lenders, as many P2P lending platforms only act as intermediaries [3]. The

main source of credit risk in P2P lending being the absence of collateral, is further increased by difficulties in risk evaluation as most lenders are not professional investors [4]. The accurate real time assessment of credit risk is a significant challenge because there is limited historical data and many loans are still on-going.

Investors use credit scoring methods to classify loans into different risk categories, aiming to distinguish between high risk and low risk loans, but additional steps are needed to predict the profit of the investment. The approach of calculating profit over a customer's life time is known as profit scoring [5, 6]. There are few studies that attempt to predict the profit of a loan in P2P lending, and they are limited to data sets with complete loans [7]. Excluding ongoing loans would create bias in the analysis because it selectively removes loans more likely to survive. To be fair, one can create a smaller unbiased data set by taking only those loans that have had the possibility of being fully observed, i.e. loans from four years ago if the maximum loan duration is four years. However, models trained on historical data may not accurately predict the current profits, since the market place has gone through significant changes in interest rates and credit ratings in the meantime.

Our research extends the literature on P2P lending by developing a profit scoring model using survival analysis that takes into account all loans, no matter how recent. Both repaid and on-going loans are used to analyze the credit risk. We use survival analysis to predict the credit risk as the monthly default probability. A simple formula then calculates the expected profit given the interest rate and the default probability.

2 Literature Review

The focus of P2P lending studies has been to analyze the borrowers' features impact on the credit risk. Many studies have applied statistical methods and machine learning techniques to develop credit scoring model for analyzing the credit risk in P2P lending.

Klafft [8] derives few simple rules from the study of US P2P lending platform Prosper, invest on loans that have no delinquencies, debt to income rate below a certain value and no credit inquires. Emerkter et al. [9] applied non-parametric test to identify the significance of borrowers' characteristics on probability of default and modeled the default risk with a binary logit regression. They further examined the relation between default probability and loan duration using Cox Proportional Hazard model. Similarly, Lin et al. [10] built a credit risk model for a P2P lending platform in China with logistic regression and identified features affecting default risk. Byanjankar et al. [11] developed an artificial neural network which outperformed logistic regression in classifying loans into defaults and non-defaults. Malekipirbazari and Aksakalli [12] performed a comparative study of machine learning methods with random forest, logistic regression, support vector machine, and k-nearest neighbor classifier.

In addition to standard credit scoring models, there have been attempts to apply survival analysis for modeling the credit risk in P2P lending. Survival analysis complements traditional credit scoring model in the sense that it has the ability to incorporate ongoing loans for credit risk modeling that are ignored by traditional credit scoring model. However, the applications have been limited to only analyzing the

relation between borrowers' features and credit risk and are not applied for predicting the future risk. Cinca et al. [1] analyzed credit risk in P2P lending platform Lending Club, where univariate tests and survival analysis were applied to identify features explaining loan defaults. The analysis reveal the factors explaining defaults to be loan grade, interest rate, loan purpose income, credit history and borrowers' indebtedness. Secondly, a logistic regression model was developed for predicting default that identified loan grade to be most significant determinant of default. Durovic [13] applied non-parametric survival analysis to find evidence of relationship between loan characteristics and default probability of loans in P2P lending platform 'Lending Club'.

The profit scoring approach extends the credit risk analysis, because it focuses on estimating the profit, rather than just predicting the credit risk. Cinca and Nieto [7] developed a profit scoring model for P2P lending applying multivariate linear regression and decision tree. The results of profit scoring were compared to logistic regression credit scoring, where the profit scoring outperformed credit scoring in identifying profitable investments. Their approach is able to select most profitable loans, but, the drawback of their study is the exclusion of ongoing loans.

3 Model

We now define our model, which consists of two parts. The first part uses survival analysis to predict a monthly default probability, and the second part has a formula that predicts the profit using this probability. The loan definition is quite straightforward. In Bondora, each loan is defined by the loan amount M, the monthly interest rate I and the number of monthly payments n. The loan amortization schedule consists of monthly payments, given by the annuity formula $P = MI(1+I)^n/((1+I)^n - 1)$. The simplest model would assume that the loan schedule stays fixed, so that the loan consists of monthly payments $P(t) = P$ until time T when the loan has been censored, repaid or the borrower defaults. This loan behavior is illustrated in Fig. 1. However, borrowers may also reschedule to repay loans early, extend their loans or fall back on their payments. We assume that in this case the loan scheme is fair to an investor: each payment reduces the outstanding principal which is multiplied by the interest rate to obtain the outstanding principal in the next period. The loan schedule is defined by a single constraint that the loan should be repaid so that the principal goes to zero.

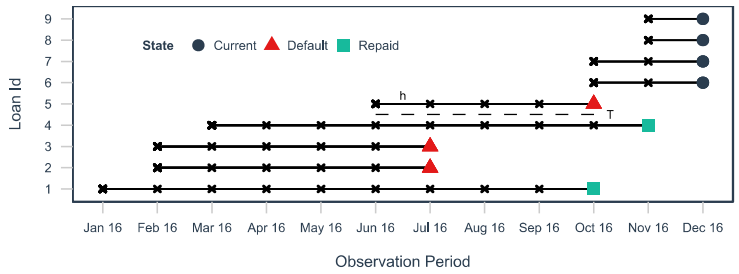

Fig. 1. Loan behavior with a loan follow-up T and the monthly default probability h.

3.1 Default Model

Because borrowers default on their payments, an investor may receive fewer than n payments. Denote the loan follow-up time as T and the loan default status as $I = \mathbb{I}$(loan defaulted). Let $C(t) = \mathbb{I}(T > t)P(t)$ be the actual loan payments at each month t. From the investor's perspective, we have payments $C(t) = P(t)$ for $t < T$ and $C(t) = 0$ for $t \geq T$ after a loan has defaulted. The continuous default time then implies the number monthly payments $n \in \{0,1,2 \dots\}$ and therefore the loan profit. We use survival analysis to model the distribution of the random variable T, using the convention that a loan with no monthly payments had a default in $T \in [0,1)$. The probability of a loan surviving to time t is given by the survival function:

$$S(t) = \mathbb{P}(T > t) = \exp\big(-H(t)\big) \tag{1}$$

where $H(t)$ is the cumulative hazard, defined as the integral of the loan default hazard:

$$H(t) = \int_0^t h(u)du \text{ where } h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2}$$

We define a monthly default probability as the probability of defaulting in an interval between loan payments, given that the loan survives to the interval:

$$h = \mathbb{P}(t \leq T < t + 1 | T \geq t) = 1 - \exp(-[H(t + 1) - H(t)]) \tag{3}$$

Covariates of a single loan $\overline{x_i} = (x_{i1}, \dots, x_{ip})'$ and the corresponding parameters $\overline{\beta} = (\beta_1, \dots, \beta_p)'$, can be incorporated into the default hazard using the popular Cox proportional hazards model, for example:

$$h_i(t) = h_0(t) \exp\big(\overline{\beta}'\overline{x_i}\big) \tag{4}$$

where $h_0(t)$ is a baseline hazard multiplied up (higher risk) or down (lower risk).

3.2 Profit Model

To calculate the value of the loan, we need to take into account the time value of money using discounted cash flow (DCF) analysis. Each monthly payment is discounted by the investor's monthly profit requirement i to arrive at the present value of the loan. For fully observed historical loans, we can compute the present value from to give us the actual profit. However, the profit of each new loan is unknown and the previous loans were made without knowledge of the actual monthly payments that were to be realized in the future. We therefore take expectation to get the expected present value of a loan:

$$C_{DCF}(i) = \sum_{t=1}^{\infty} \frac{\mathbb{I}(T > t)P(t)}{(1 + i)^t} \quad \Longrightarrow \mathbb{E}[C_{DCF}(i)] = \sum_{t=1}^{\infty} \frac{S(t)P(t)}{(1 + i)^t}$$

If an investor is willing to lend or purchase a non-random loan for C, then $C > C_{DCF}(i)$ implies they had a lower profit requirement and $C < C_{DCF}(i)$ implies they had a higher profit requirement. In the later case, our hypothetical investor should be willing to

invest in the loan. In an efficient market loans with the same amount of risk should have the same implicit profit requirement. We can obtain the implicit profit requirement by solving the equation $C = C_{DCF}(i)$ for i. This is the amount an investor with profit requirement i should be willing to pay for the loan, assuming that they know the actual payments. For a new loan we can solve $C = \mathbb{E}[C_{DCF}(i)]$ for the expected profit i, which is implied by the survival function $S(t)$. The expected profit corresponds to the present value of a portfolio with infinitely many such loans. In the formula above, we assumed for simplicity that the loan schedule $P(t)$ is fixed. Otherwise, we would need to also model the random loan schedule $\{P(t)\}_{t\geq 1,2\dots}$ to compute the expected value.

3.3 Special Case: Constant Default Rate

So far the discussion is applicable to any survival model. We consider a constant default hazard model in the experiments. This is known as the exponential model:

$$h(t) = \lambda, \qquad H(t) = \lambda t, \qquad S(t) = \exp(-\lambda t) \text{ and } h = 1 - \exp(-\lambda)$$

We assume a constant hazard because it allows three significant simplifications:

1) A simple formula can be derived for the expected profit. We do not need to numerically solve $C = \mathbb{E}[C_{DCF}(i)]$ for i.
2) The expected profit is independent of the loan schedule. The solution i of the equation $C = \mathbb{E}[C_{DCF}(i)]$ does not depend on $P(t)$, implying that repayments, extensions and late loans have no impact on the profit.
3) Each monthly profit is an unbiased estimate of the loan profit. Given observed monthly profits C_{it} (defined in 5.1), we have $\mathbb{E}[C_{it}] = \arg_i[C = \mathbb{E}[C_{DCF}(i)]]$ implying that their mean value is very close to the true profit of a portfolio.

For reasons of space, we defer the straightforward proofs of these statements to a subsequent article. The expected profit given the loan interest rate I, the loan monthly default rate h and the loss given default D can be calculated from a profit formula:

$$C = \mathbb{E}[C_{DCF}(i)] \Longrightarrow i = (1 - h)I + hD \tag{5}$$

This formula uses a generic loss rate D for the present value of a defaulted loan, in the case that default does imply a total loss of principal, meaning 100% loss ($D = -1$). Bondora states this quantity for each loan, and some platforms sell the defaulted loans to collection agencies for a given percentage of the remaining principal.

4 Data

The data for the research was obtained from Bondora[1], a leading P2P platform in Europe, which currently operates in Estonia, Finland, Spain and Slovakia. There are

[1] https://www.bondora.com/en/public-reports

altogether 65675 loans described by 112 features that mostly include demographic and financial information on borrowers. Our snapshot included loans that were issued between 28 February 2009 and 4 October 2018. The rating system was applied by Bondora from 2013 and hence there are no ratings assigned to loans issued before 2013. We therefore include loans issued from 2013 onwards. The data also describes the current state of the loans and their payment behavior. Each loan is either in the state of repaid, current or late. A loan is considered to be in default if the loan is more than 60 days past its due payment.

The data consists of 36.5% of defaulted loans, 41.8% current loans and 21.7% of repaid loans. The high default rates indicate the high risk of the loans. Table 1 shows the number of loans across the year in different loan status. As seen from the table, the vast majority of recent loans are current loans. In addition, the base interest rate on the loans has been decreasing in the recent years, which implies a significant difference between older and newer loans.

Table 1. Loans over the years, where Interest refers to A-rated loans as the baseline

	Year					
	2013	**2014**	**2015**	**2016**	**2017**	**2018**
Current	182	775	1436	2533	8199	14200
Repaid	1447	2716	2139	2539	2568	1081
Default	846	3942	4471	5441	7166	1209
Interest	24.95%	24.77%	16.22%	12.40%	11.84%	11.76%

The borrowers are classified into 8 different risk groups based on Rating levels assigned to them. Table 2 shows the distribution of ratings in the data and the average interest rates and current status across the ratings. It is evident that the interest rate increases as the credit risk increases to compensate for the losses due to defaults.

Table 2. Ratings with their average annual interest rate % and current status

Ratings	***AA***	***A***	***B***	***C***	***D***	***E***	***F***	***HR***
Interest (%)	11.9	14.6	16.8	22.1	27.9	33.8	45.0	80.0
% Defaulted	9.60	13.6	17.6	24.9	33.3	38.4	42.7	70.4
% Repaid	22.8	26.1	21.1	21.7	19.3	16.3	18.7	18.8
% Current	67.6	60.3	61.3	53.3	47.4	45.3	38.6	10.6

For the purpose of our study, we use a subset of the 112 features. We do not consider features with missing values or post loan application behavior. Since our interest is to analyze borrowers at the time of application, we include a subset of relevant features available at that time. We consider literature in P2P lending and perform careful analysis of features in selecting the features for the modeling. We used the Bondora data set columns NewCreditCustomer, VerificationType, Age, Gender, AppliedAmount, Interest, LoanDuration, UseOfLoan, MaritalStatus, EmploymentStatus, IncomeTotal, Rating and Status. The current state of the loan is given by the Status column (default/repaid/late/current).

5 Experiments

The main objective of the research is to predict the profit of loans. We need to be able to calculate the actual profit in censored loans to measure the accuracy of our model. We explain how this is done in the first chapter. We then evaluate the accuracy of the model in the second chapter. In the third chapter, we select a portfolio of most profitable loans, comparing our model to credit scoring and rating based selection.

5.1 Default and Profit Model Evaluation

We consider each monthly period separately to obtain unbiased estimates of profit in the presence of censoring. This idea works as follows. Each loan consists of monthly intervals $[t, t+1)$ defined by months $t = 0,1,2,\ldots$. For every monthly interval up to the default time T, a loan either defaults $D_{it} = \mathbb{I}(t \leq T < t+1)$ or survives $1 - D_{it} = \mathbb{I}(T \geq t+1)$. Given the monthly default probability h_{it} for person i and month t, the monthly default is a Bernoulli trial with probabilities h_{it} and $1 - h_{it}$ of obtaining 1 or 0, respectively. The outcome values are D_{it} and the predicted probabilities are h_{it}.The profits are defined similarly. If the loan defaults in the interval we lose D_i and if the loan survives we obtain I_i on the remaining principal. Each monthly profit is then either D_i or I_i percent, where we denote $C_{it} \in \{D_i, I_i\}$ as the realized monthly profit for person i in month t. This is also a Bernoulli trial with probabilities h_{it} and $1 - h_{it}$ of obtaining D_i or I_i, respectively. The outcome values are C_{it} and the predicted profits are given by the profit formula $Y_{it} = (1 - h_{it})I_i + h_{it}D_i$.

We have observed the monthly outcomes $D_{it} \in \{0,1\}$ and $C_{it} \in \{I_i, D_i\}$ which we compare to the predicted values h_{it} and Y_{it}. We have a single model for the defaults and the expected profits, because the default model implies the profit model. We evaluate accuracy using the Mean Squared Error (MSE):

$$\text{MSE}_{\text{Loan}} = \frac{1}{\sum_{i=1}^{N} T_i} \sum_{i=1}^{N} \sum_{t=1}^{T_i} (D_{it} - h_{it})^2 \qquad \text{MSE}_{\text{Profit}} = \frac{1}{\sum_{i=1}^{N} T_i} \sum_{i=1}^{N} \sum_{t=1}^{T_i} (C_{it} - Y_{it})^2$$

5.2 Accuracy of The Model

We use exponential survival model to model the monthly default probability. We partition the data into train and test set in the ratio 80:20. The train set is used to train the model and we evaluate the model with the test set. Applying the hazard function in the presence of covariates as shown in equation 4 we obtain the monthly default probabilities. We evaluate the default model using the mean squared error (MSE) and the area under the ROC curve (AUC). After we get the predicted default probability, we predict the monthly profit from the loan using the profit formula in equation 5 and evaluate the profit error with MSE.

The results are shown in Table 3. The low default error and profit error show that our model is accurate at identifying the exact default probability and monthly profits. The AUC score for the default prediction also states a reasonable discrimination

performance of the model. Hence, the model looks promising in predicting the expected profit of loans, and possibly selecting loans with excess profits.

Table 3. Results on train and test set

	Default Error	***Profit Error***	***AUC (default)***
Train set	0.0288	0.0170	0.7114
Test set	0.0292	0.0167	0.7098

5.3 Profitability of Loans

To illustrate the difficulty of selecting loans, Fig. 2 shows the comparison of the interest rates and the predicted profits in the test data. The interest rates are the profits that lenders would obtain if there were no defaults at all. Most interest rates tend to lie between 25% and 35%. However, the majority of the predicted profits are between 5% and 15% and there are considerable amount of loans with significant loss. The profit in P2P loans does not seem to be as high as it might be expected from the interest rates alone when we take into account the risk of default and principal lost in default.

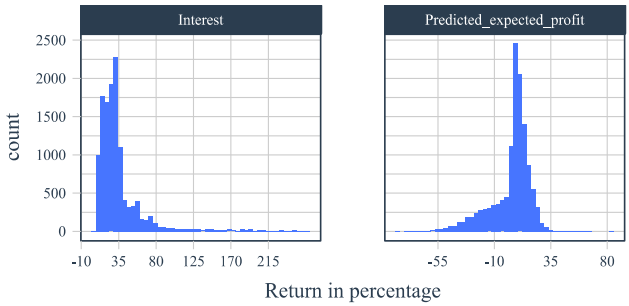

Fig. 2. Histogram of interest and estimated profit rates on test set

This difference is explained by the profit formula, which shows that there should be a direct relation between the default risk and the interest rate. In Fig. 3, we plot the predicted default hazards against the interest rates. For a nicer visualization, we assume loss given default of 0.9; the median value for all the loans. The black curve represents the loans having zero profit; the green curve is the loans with 20% profit and red curve with 20% loss. The interest rates clearly correlate with the default risk to compensate for the defaults. Investors therefore use their own models or intuition to estimate this trade-off. Since there is some variation in the interest rates for a given predicted default probability, our model does not fully agree with the investors decisions and this suggests it may also be able to generate excess returns.

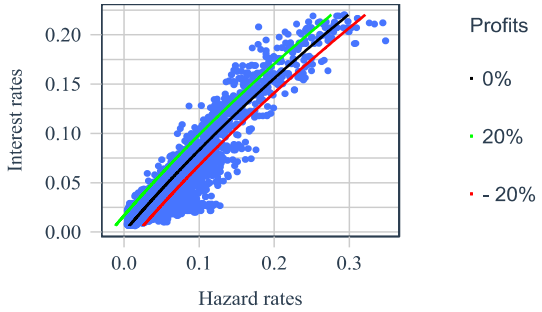

Fig. 3. Relation between interest and hazard rates with levels of profitability

The selection of loans can be exploited to build a portfolio with higher or lower profits. Fig. 4 depicts an experiment, where we have picked top X% of loans based on different selection criteria. The actual return was estimated on a test, using the approach in the first chapter. As seen in the figure, applying our model results yields a higher return compared to selecting by their ratings and hazard rates. Investing only on top 10 percent of loans based on their predicted profit would have received an average profit of 20%, which is higher than 12% based on hazard rates and 10% based on Ratings. Knowing the default risk gives an improvement over selection by rating, but incorporating the default risk in the profit formula gives even better results.

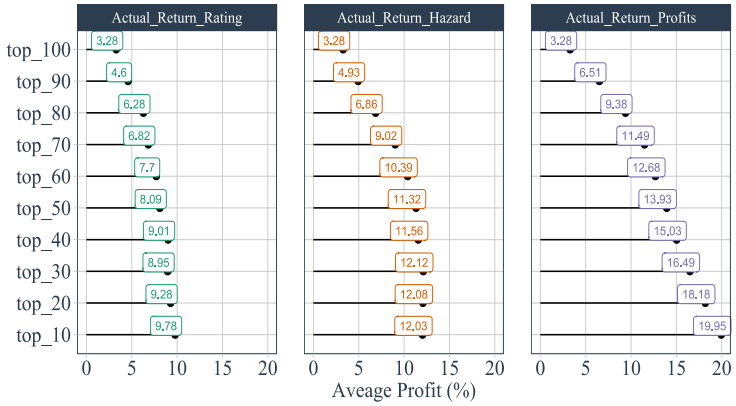

Fig. 4. Average portfolio return on loans

6 Conclusion

P2P lending is a growing field of micro finance that operates online and has gained popularity as an alternative to traditional banking. Along with the growth, there are also challenges in P2P lending, where credit risk is one of the major concerns for the lenders. We argue that it is important to predict not just the credit risk, but also the profit in the

loans. The analysis should incorporate recent loans that are still on-going to accurately predict the profits in the changing economic conditions of P2P lending.

Hence, we applied survival analysis to model the monthly default probability in P2P lending that incorporates on-going, repaid and defaulted loans. We extended this model by a formula that predicts the profit of a loan given an interest rate, the predicted default probability and the loss given default. Our results reveal that the loans in P2P lending may also result in significant future losses. Model evaluation shows that our model had good performance in predicting the default risk of a loan and estimating the profit. In addition, selecting loans based on the estimated profits computed by our model yields a higher return compared to relying only on the ratings of the loans or the default risk.

We simply used logistic regression, as the main idea was to show the utility of our profit extension. Our approach and the profit framework can be directly extended with time-varying and more sophisticated models for the monthly default probability.

References

1. Serrano-Cinca, C., Gutierrez-Nieto, B. and López-Palacios, L.: Determinants of default in P2P lending. PloS one, 10(10), p.e0139427 (2015)
2. Demyanyk, Y.S., Loutskina, E. and Kolliner, D.: Federal Reserve Bank of Cleveland (2014)
3. Li, J., Hsu, S., Chen, Z. and Chen, Y.: Risks of p2p lending platforms in china: Modeling failure using a cox hazard model. The Chinese Economy, 49(3), 161-172 (2016)
4. Lee, E. and Lee, B.: Herding behavior in online P2P lending: An empirical investigation. Electronic Commerce Research and Applications, 11(5), 495-503 (2012)
5. Louzada, F., Cancho, V., de Oliveira Jr, M. and Bao, Y.: Modeling time to default on a personal loan portfolio in presence of disproportionate hazard rates. J Stat App Pro., 3(3), 295-305 (2014)
6. Andreeva, G.: European generic scoring models using survival analysis. Journal of the Operational research Society, 57(10), 1180-1187 (2006)
7. Serrano-Cinca, C. and Gutiérrez-Nieto, B.: The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decision Support Systems, 89, 113-122 (2016)
8. Klafft, M.: "Online peer-to-peer lending: A lenders' perspective", Proceedings of the International Conference on ELearning, E-Business, Enterprise Information Systems, and EGovernment, EEE, 371-375 (2008)
9. Emekter, R., Tu, Y., Jirasakuldech, B. and Lu, M.: Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Applied Economics, 47(1), 54-70 (2015)
10. Lin, X., Li, X. and Zheng, Z.: Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. Applied Economics, 49(35), 3538-3545 (2017)
11. Byanjankar, A., Heikkilä, M. and Mezei, J.: "Predicting Credit Risk Levels in Peer-to-Peer Lending: A Neural Network Approach", IEEE Symposium Series on Computational Intelligence, SSCI, pp. 719-725. Cape Town (2015)
12. Malekipirbazari, M. and Aksakalli, V.: Risk assessment in social lending via random forests. Expert Systems with Applications, 42(10), 4621-4631 (2015)
13. Đurović, A.: Estimating Probability of Default on Peer to Peer Market–Survival Analysis Approach. Journal of Central Banking Theory and Practice, 6(2), 149-167 (2017)