Analyzing Peer-to-Peer Lending Secondary Market: What Determines the Successful Trade of a Loan Note?

Byanjankar, Ajay; Mezei, Jozsef; Wang, Xiaolu

Published in:
Trends and Innovations in Information Systems and Technologies - Volume 2, WorldCIST 2020

DOI:
10.1007/978-3-030-45691-7_44

Published: 01/01/2020

Document Version
Accepted author manuscript

Document License
All rights reserved

Link to publication

Please cite the original version:
Byanjankar, A., Mezei, J., & Wang, X. (2020). Analyzing Peer-to-Peer Lending Secondary Market: What Determines the Successful Trade of a Loan Note? In Á. Rocha, H. Adeli, L. P. Reis, S. Costanzo, I. Orovic, & F. Moreira (Eds.), *Trends and Innovations in Information Systems and Technologies - Volume 2, WorldCIST 2020* (pp. 471-481). (Advances in Intelligent Systems and Computing; Vol. 1160 AISC). Springer, Cham. https://doi.org/10.1007/978-3-030-45691-7_44

Analyzing Peer-to-Peer Lending Secondary Market: What Determines the Successful Trade of a Loan Note?

Ajay Byanjankar(✉), József Mezei, and Xiaolu Wang

Åbo Akademi University, Turku, Finland
ajay.byanjankar@abo.fi

Abstract. Predicting loan default in peer-to-peer (P2P) lending has been a widely researched topic in recent years. While one can identify a large number of contributions predicting loan default on primary market of P2P platforms, there is a lack of research regarding the assessment of analytical methods on secondary market transactions. Reselling investments offers a valuable alternative to investors in P2P market to increase their profit and to diversify. In this article, we apply machine learning algorithms to build classification models that can predict the success of secondary market offers. Using data from a leading European P2P platform, we found that random forests offer the best classification performance. The empirical analysis revealed that in particular two variables have significant impact on success in the secondary market: (i) discount rate and (ii) the number of days the loan had been in debt when it was put on the secondary market.

Keywords: Machine learning · Binary classification · Peer-to-Peer lending · Secondary market

1 Introduction

Peer-to-Peer (P2P) lending is a micro finance service operating online to connect borrowers and lenders for loan transactions [1]. P2P platforms allow for easy and quick loan processing for borrowers due to automated handling and less cost because of the absence of traditional financial intermediaries. Following from the premise, it provide higher return to lenders compared to similar traditional investments. In recent years, the service has been gaining popularity and growth as a result of quick and easy access to loan for borrowers [2]. However, there are risks associated to lending, the primary being the lack of collateral. Additional risk of investment loss arises from lack of analytical skills of investors and information asymmetry from online services.

Investors fund borrowers in P2P lending platforms based on the information provided in the loan application. There is an automated service provided by P2P platforms that assists lenders to select the best options. This automated

Á. Rocha et al. (Eds.): WorldCIST 2020, AISC 1160, pp. 471–481, 2020.
https://doi.org/10.1007/978-3-030-45691-7_44

service is used by most of the non-professional investors. Due to the use of automated service, there is very little opportunity for investors to screen the loan applications from the primary market for investments. P2P lending also has a secondary market, where investors can sell their loan holdings and this service is less automated compared to primary market allowing investors to analyse and select investments. Loan holders can split their investment to create several loan notes and put them in the secondary market for sell with either discount rates or premium depending on the present status of the loan [3].

Primary market being mostly controlled by automated service provides less information to study lenders' behavior and hence secondary market could be a good place to analyse lenders' investment behavior. Secondary market is more risky than primary market as loan holders put their holdings to sell when they see problem in loan recovery. Finally, most P2P platforms adopt an agent model that relies primarily on retaining and attracting a stable investor base to generate fee revenue. We believe our study could not only help them improve the selection of borrower characteristics and hence the investor-borrower matching, but also assist in the provision of liquidity of the secondary market. In this study, we analyze the investment behavior of lenders in P2P lending secondary market in understanding a loan note being successfully traded in the market. We study the relation between a loan note being successfully traded and the loan features to establish an understanding on lenders' selection behavior of loan notes. Finally, with the help of classification models we verify our understanding of lenders' selection behavior of loan notes from the secondary market.

2 Literature Review

In this section we will briefly summarize related literature. A search in academic databases shows that contributions considering secondary markets of peer-to-peer lending are scarce. For this reason, we mainly focus on the general use of machine learning in P2P lending.

2.1 Machine Learning-Based Approaches in P2P Literature

Recent years have seen a continuously increasing number of academic contributions utilizing various machine learning algorithms in the context of P2P lending. This interest can largely be explained by the fact that researchers can have easy and direct access to loan data, which is typically not available from traditional financial institutions. This resulted in numerous publications focusing on different markets from around the globe, such as Bondora in Europe [4], Lending Club in the United States [2] and HouBank in China [5]). The most important issue modeled is the probability of loan defaults. There is a wide variety of machine learning algorithms that were utilized to construct optimal models.

The most frequently used machine learning models used in P2P studies include logistic regression, random forests, gradient boosting machine and neural networks; these are the methods tested in the empirical part of this article.

As the typical baseline classification method in classification tasks in finance, logistic regression, is used in [6]. By analysing data from Lending Club, the authors identify using binary logistic regression that credit grade, debt-to-income ratio, FICO score and revolving line utilization are the most important factors that determine loan default. Another widely used machine learning model appearing in P2P literature is random forests. In, [7] a random forest based model is proposed to predict loan default and is found to outperform FICO credit scores. Gradient boosting machine (GBM) is based on the idea of combining several weak classifiers in an ensemble model. [8] combines extreme gradient boosting with misclassification cost and at the same time proposes to evaluate models based on annualized rate of return. Lastly, neural networks, probably the most widely used machine learning model across all domains, has been tested several times in P2P lending [4].

It is important to mention here, that additionally to traditional numerical data, some recent contributions make use of unstructured data, typically texts. In [9], topic modelling is utilized to extract features from loan descriptions with the most predictive power. Using data from the Eloan Chinese P2P platform, and by combining the extracted features with hard information in some traditional machine learning algorithms, the authors show that classification performance can increase by up to 5%.

2.2 P2P Secondary Markets

All the above mentioned articles concern the use of data from the primary market of P2P platforms. However, the option to resell (parts of) purchased loans to interested investors in a secondary market is present in various platforms; in the platform Bondora considered in the empirical analysis, secondary market became available already in 2013. Still, one can identify a very small number of contributions analysing data from secondary markets, although platforms, such as Bondora, typically make it also available. This lack of academic interest is especially staggering when we consider that according to an extensive report presented in [10] on the globally largest P2P market, already in 2017, more than 50% of the Chinese P2P platforms provided the possibility to perform secondary market transactions.

Among the very few related contributions, one can highlight the study presented in [3] in which the authors study mispricing in the secondary market of Bondora analyzing more than 51000 loans. Utilizing least absolute shrinkage and selection operator (LASSO), the mispricing is explained as the consequence of mistaken perceived loan values. In another study [11], the author, analysing data from the LendingClub platform's secondary market, collected data in a three-month period, and found that the liquidity of the secondary market is very low, with less than 0.5% of listings resulting in a trade. The author concluded that it would be impossible to find the fair value for the pool of identical loan notes. In light of the importance of secondary markets in P2P lending and the small number of related contributions, in the following we address this gap in the literature in an empirical study on the secondary market of the Bondora platform.

3 Data and Exploratory Analysis

The data for the study was collected from a leading European P2P platform, Bondora. The data includes loan notes traded in the secondary market of the platform until July 2019. Bondora had launched the secondary market in March of 2013, four years after Bondora was established. The secondary market offers the opportunity to investors to resell outstanding loans from the primary market. Loan holders can split the outstanding principal on their loan holdings to create several loan notes and put them in the secondary market for sell with either discount rates or premium depending on the present status of the loan. In the analysis section a negative sign(-) is used to just differentiate discount rate from premium keeping the magnitude of the value intact. A loan listing is allowed to be placed in the secondary market for a maximum period of 30 days and if it is traded within the time limit it is given a status of 'Successful' else it is removed automatically and given the status 'Failed'.

The data includes the information related to transactions of the loan notes, such as start and end date of the transaction, amount, discount or premium rate, number of days that the loan has been in debt, and the result indicating whether the loan note was sold or not. Each loan note is connected to its original loan id and based on this some relevant demographic and financial information on the loan note are extracted from the main loan database for additional features. After performing data preprocessing on the raw data, there are around 7.3 million records of loan notes. This large number (in contrast to the less than 100,000 loans) is the result of loan holders being able to split their holdings into multiple loan notes. The large number of loan notes provide investors with plenty of options to select for investments and also makes it more computationally challenging to apply sophisticated machine learning models requiring a large amount of computation.

3.1 Exploratory Analysis

In the following, we discuss the basic characteristics of the data. First, we can observe that the majority of the loan notes, 61% failed (not sold or cancelled before expiry) and 39% were successfully sold in the secondary market. The higher number of failed loan notes in the secondary market is expected as loan holders typically sell out loans on which they have difficulties in recovering the payments. Figure 1 illustrates the time in relative to the Loan Duration that has passed from the loan issue date to the time the loan was listed in the secondary market. From Fig. 1 we can see that majority of loans appear in secondary market after they have crossed 10% to 30% of their Loan Duration as usually borrowers tend to fail in their payments after few initial payments. In addition, investors in the secondary market are interested in purchasing loan notes which are in the early stage, as they behave more like new loans and hence have possibility of making higher profits through interest collections.The loan notes in the category '>100' signifies that the loan notes have crossed their initial Loan Duration to a large extent.

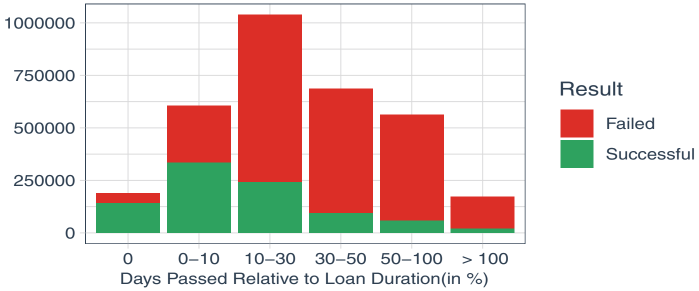

Fig. 1. Time since loan issue

In Fig. 2, the distribution of the number of days the loans were in debt at the time of listing and the discount rates are depicted. Majority of the loan notes have low debt days, between 0 and 10, and high probability of being sold out. It illustrates that loan holders in general have less patience and low risk tolerance and try to immediately resell the loans as soon as payments are late by a few days. Loan notes with more than 10 days in debt show very low probability to be sold on the secondary market, with the exception of the cases when higher discount rates are offered.

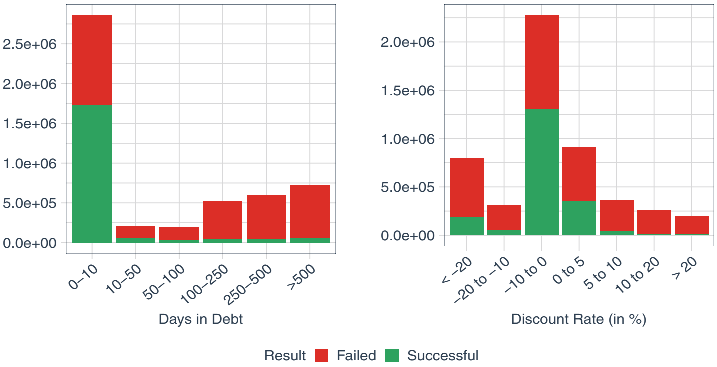

Fig. 2. Debtdays and DiscountRate distribution

The effect of DiscountRate on the success can be seen in the right in Fig. 2. The bins in x-axis are arbitrarily created to better represent the distribution and the lower bound is excluded while the upper bound is included in the bins. Loan notes that were listed with DiscountRate between 10 and 0 constitute most of the notes, where loan notes with 0 DiscountRate is around 38% of the total loan notes and we can observe high average likelihood of success for these notes. However, loan notes with very high discount rates had low success rate as they might be perceived risky by investors. Similarly, we can observe the trend that the higher the offered premium is, the less likely it is that the transaction will succeed. Figure 2 shows that the features DebtDays and DiscountRate tend to have high impact on a loan note being successful in the secondary market. Therefore, we further investigate these two features and also analyse their interaction.

Based on the presented figures, high success rate can be observed together with low discount rates and number of days in debt. For this reason, first we look at the loan notes that correspond to these criteria. Figure 3 depicts the comparison of the effect of the two features on the success rate when they take on the value 0. The left part of the figure shows the case when each feature is considered individually without any interaction included. According to this, 0 DiscountRate or 0 DebtDays does not have significant impact on result: the proportion of failed and successful loan notes are almost equal in both the cases.

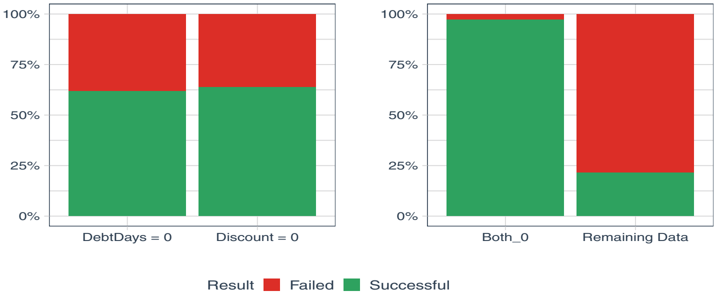

Fig. 3. Relation of debt days and discount with result

When we take the interaction into account, we can see a completely different result in the right part of Fig. 3. The loan notes that have both DiscountRate and DebtDays 0 are likely to be successful in the secondary market almost all the time with very few failed loan notes. Importantly, this group of loans account for approx. 23% of all loan notes. For the remaining data we see the opposite behavior: most loan notes not having 0 DebtDays and DiscountRate have failed in the secondary market. Further analysing the relation between DiscoutRate and DebtDays for the remaining data, a more refined picture on the relation between DiscoutRate and DebtDays for loan notes is presented in Fig. 4.

The heat map in Fig. 4 shows the success rate of loan notes for different combination of DiscountRate and DebtDays where both DiscountRate and DebtDays are not 0. From Fig. 4, we can conclude that for the majority of combinations the success rate is close to 0. This shows a clear relation between DiscountRate, DebtDays and the success rate. As an exceptional group, loan notes with low DebtDays and higher DiscountRate have high success rate. The success rate decreases for high premium loans and is near to zero for loan notes listed at premium with high number of DebtDays. Similarly, success rate decreases as number of days in debt increases. As a summary, we can conclude that the investors choose loan notes that have combination of lower DebtDays and higher DiscountRate and very clearly neglect loan notes with premium. The two features are likely to be very predictive of a loan note being successful in the secondary market.

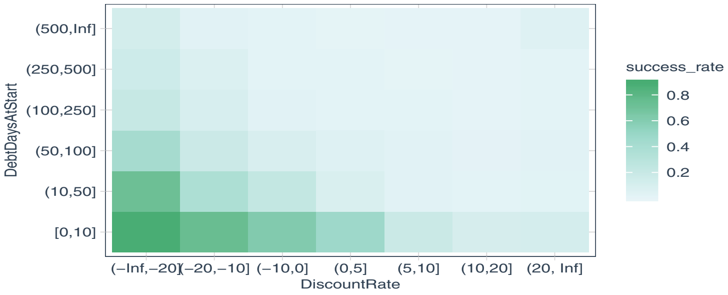

Fig. 4. Success rate at different levels

4 Classification Models and Results

According to the presented descriptive analysis, DebtDays and DiscountRate are very informative regarding the success of a loan note in the secondary market. Investors are very likely to make their investment selection decisions mostly based on these two features. To further validate our assumption, we trained several machine learning models to classify successful and failed loan notes and evaluate the importance of the features. Loan notes having both DebtDays and DiscountRate as 0 are almost always likely to be successful with a success rate of around 97% as seen in Fig. 3. With a reasonable amount of loan notes (around 23%) showing such behavior, it is safe to derive a simple rule that such loan notes will be successful in the secondary market. For this reason, we reduce our focus to the remaining data, where it is further split into train and test set for training and evaluating classification models to identify successful and failed loans.We treat being successful as the positive class.

Multiple machine learning models, such as Logistic Regression with Lasso penalty (LR), Random Forest (RF), Gradient Boosting Model (GBM) and Neural Network (NN) were applied to train the classification models. The models were optimized through hyper-parameter search and applying feature selection with the feature importance from the model. AUC score (Area under the Receiver Operating Curve) was used as the primary evaluation metric. The results of the final models on the test data are presented in Table 2. The F1 score is reported as the maximum possible score and Accuracy is reported at the threshold for the maximum F1 score. Among the models, RF model performs the best with very high Accuracy, AUC and F1 score and low Logloss. The final set of features for the best performing model Random Forest can be seen in Table 1.

Table 1. Features set for random forest

Features	Description
DiscountRate	Discount or Premium offered
DebtDays	Number of days the loan was in debt
start_day	Day of month the loan was listed in secondary market
start_hour	Hour the loan was listed in secondary market
start_month	Month the loan was listed in secondary market
start_weekday	Weekday the loan was listed in secondary market
days_passed	Days passed relative to original loan duration
Interest	Interest rate on loan
Probability of Default	Probability of default within a year
Rating	Rating assigned to the loan

Table 2. Classification results with all features

Models	Accuracy	AUC	Logloss	F1
LR	0.745	0.800	0.446	0.571
RF	**0.925**	**0.969**	**0.189**	**0.840**
GBM	0.886	0.931	0.268	0.756
NN	0.894	0.940	0.252	0.774

The feature importance from all the models are illustrated in Fig. 5, that shows the top 8 features with the features' name presented in their abbreviation form. All the models identify DebtDays(DDAS) and DiscountRate(DsCR) as the top two features, which matched our assumption from the earlier section. In addition, for the best performing model Random Forest, the importance of the two features is very high compared to rest of the features. Hence, the results show that the investors highly rely on the two features when deciding to select the loan notes from the secondary market.

To further validate the importance and effect of the two features, we trained the classification model with only two of the features DebtDays and DiscountRate. The models with only these two features still achieved good results as seen in Table 3. The results indicate the strong predictive power of the two features.

The partial dependence plots for the features DebtDays and DiscountRate in Fig. 6 show the mean effect of the features along with the variance. For DebtDays, mean success rate rapidly decreases as number of DebtDays goes above 0 and remains almost constant after about 100 days. This shows that most of the investments are made on loan notes with low DebtDays; for higher DebtDays there seems to be no investment pattern as shown by the constant and low success rate. For the DiscountRate, the mean success rate is higher at higher DiscountRate but decreases with the DiscountRate and is almost the same from −25 to 0. For the loan notes with premium the success rates decrease rapidly

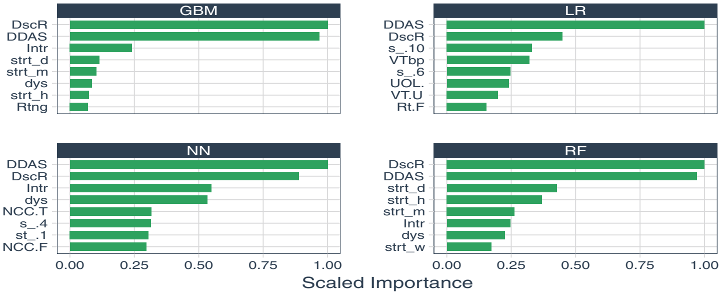

Fig. 5. Feature importance

Table 3. Classification results with two features

Models	Accuracy	AUC	Logloss	F1
LR	0.788	0.797	0.463	0.60
RF	**0.831**	**0.866**	**0.362**	**0.645**
GBM	0.827	0.859	0.370	0.638
NN	0.833	0.848	0.386	0.631

with higher premium. The high variance seen for the cases with higher DiscountRate hints that higher discount rates may alone not be the deciding factor, but is the combination of both DiscountRate and DebtDays.

5 Conclusion

An increasingly established way for P2P platforms to provide increased means of liquidity to investors is a secondary market. While academic research is widely available to understanding P2P platform processes, evaluation of platforms and models to estimate loan default to assist investors, research contributions focusing on secondary markets are scarce. In this article, we consider the problem of predicting the success of a posting in a P2P secondary market. Several widely used machine learning models are tested with data used from the leading European P2P platform Bondora. Based on evaluating the performance of models on more than 7 million postings, we found that most of the algorithms, except for logistic regression, perform very similarly, with AUC value as high as 0.925 is achievable. Furthermore, we found that success of a posting is largely determined by two specific variables: (a) the number of days since the loan has been in debt at the beginning of the posting and (b) the discount rate. As it is shown in the article, using only these two variables, one can construct models with high performance.

In this work, we have done one of the first steps in trying to understand secondary markets of P2P platforms by utilizing machine learning algorithms. In the future, the study can be extended by incorporating other variables on

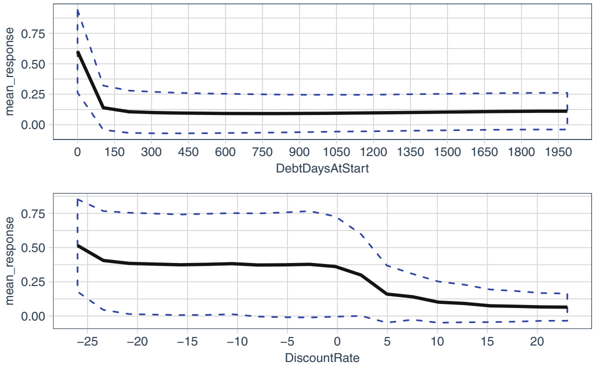

Fig. 6. Partial dependence plots

different loan characteristics that may impact the success of a transaction. Furthermore, in lack of identifier for investors in our dataset, it is not possible to assess any kind of network effect potentially present in the platform; this could be analysed with appropriate data available. Finally, our results reflect the data from a European P2P platform, we cannot claim that similar models would definitely show same performance when used to data from different areas, such as China, the largest P2P market; this would require further investigation.

References

1. Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., Tiburtius, P., Funk, B.: Online peer-to-peer lending-a literature review. J. Internet Bank. Commer. **16**(2), 1 (2011)
2. Kumar, V., Natarajan, S., Keerthana, S., Chinmayi, KM., Lakshmi, N.: Credit risk analysis in peer-to-peer lending system. In: 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA), pp. 193-196 (2016)
3. Caglayan, M., Pham, T., Talavera, O., Xiong, X.: Asset mispricing in loan secondary market. Technical Report Discussion Papers 19-07. Department of Economics, University of Birmingham (2019)
4. Byanjankar, A., Heikkilä, M., Mezei, J.: Predicting credit risk in peer-to-peer lending: a neural network approach. In: 2015 IEEE Symposium Series on Computational Intelligence, pp. 719-725 (2015)
5. Guo, W.: Credit scoring in peer-to-peer lending with macro variables and machine learning as feature selection methods. In: 2019 Americas Conference on Information Systems(2019)
6. Emekter, R., Tu, Y., Jirasakuldech, B., Lu, M.: Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Appl. Econ. **47**(1), 54–70 (2015)
7. Malekipirbazari, M., Aksakalli, V.: Risk assessment in social lending via random forests. Expert Syst. Appl. **42**(10), 4621–4631 (2015)
8. Xia, Y., Liu, C., Liu, N.: Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electron. Commer. Res. Appl. **24**, 30–49 (2017)

9. Jiang, C., Wang, Z., Wang, R., Ding, Y.: Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Ann. Oper. Res. **266**(12), 511–529 (2018)
10. Yin, H.: P2P lending industry in China. Int. J. Ind. Bus. Manage. **1**(4), 0001–0013 (2017)
11. Harvey, S.: Lending Club's Note Trading Platform Facade: An Examination of Peer-to-Peer (P2P) Lending Secondary Market Inefficiency. University of Dayton Honors Thesis (2018)