Journal of Industrial and Systems Engineering
Vol. 13, No. 4, pp. 288-306
Autumn (November) 2021

A data driven model for credit scoring of loan applicants within a crowdfunding scenario in a P2P lending platform in Iran

Neda Hodjat Panah[1], Mohammad Reza Rasouli[1*]

[1]*School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran*

nhodjatpanah@ymail.com, rasouli@iust.ac.ir

Abstract

Crowdfunding is a fundraising tool to solicit many small amounts of capital from a large number of potential investors. Peer to peer lending is known as a main type of crowdfunding in which lenders and borrowers can interact directly through an online platform. By eliminating the intermediaries and therefore reducing operating expenses, P2P platforms can provide a win-win situation for both borrowers and lenders. However, the absence of intermediaries –such as banks- increases the risk of loan repayment fraud. To avoid such losses, credit scoring methods help lenders to decide on a specific loan by assessing corresponding credit risk.

This paper proposes a credit scoring model on a P2P lending platform in Iran. Although data-driven approaches have increasingly used to enhance credit scoring within financial domains, there is a lack of research on assessing the usability of these approaches within P2P crowdfunding scenarios. This research focuses on developing a novel data-driven model that can enhance P2P credit scoring within crowdfunding scenarios. To do so, on the basis of data from an Iranian P2P lending platform, five different tree-based classifiers were developed, among which Random Forest resulted in the best accuracy (97.80%). Lenders in the used platform are businesses, each having a different risk tolerance threshold. A default probability was computed for each loan request to help lenders make decisions based on their own risk tolerance. The results clearly demonstrate how novel data analytics approaches can enhance intelligent decision making about P2P funding within P2P lending platforms.

Keywords: P2P Lending platforms, credit scoring, Iranian lending platform, random forest algorithm

1-Introduction

Crowdfunding is a rapidly growing phenomenon that allows fundraisers to solicit funds from a potentially large pool of investors (Polena and Regner, 2018). Based on the earned benefits, crowdfunding falls into four categories: donation, reward, equity, and lending. In the lending-based crowdfunding, peer-to-peer lending- also referred to as P2P- is one of the relatively new models, experiencing dramatic growth around the world (Belleflamme et al., 2015). The first P2P lending platform, Zopa, was established in 2005 in the UK. Since then, the P2P lending market has grown significantly. According to Valuates Reports, the global P2P lending market size was estimated at USD 67.93 Billion in 2019 and is expected to reach USD 558.91 Billion by 2027, with a 29 percent compound annual growth rate.

*Corresponding author

ISSN: 1735-8272, Copyright c 2021 JISE. All rights reserved

P2P lending refers to the act of connecting lenders and borrowers through a lending platform. It allows individuals -real or legal- to lend or borrow funds directly, without the intervention of any financial intermediaries such as traditional banks (Klafft, 2008). Bringing together borrowers and lenders as well as setting regulations for their participation in the lending process are the two main tasks of the platforms. Compared to traditional banks, P2P lending platforms have lower operating costs, which enables them to offer rapid access to capital, a competitive interest rate, and a speedy lending process to borrowers as well as a higher rate of return to lenders (Patwardhan, 2018).

Although the elimination of financial intermediaries makes loan offers more attractive for both lenders and borrowers, there are also drawbacks. Unlike traditional systems, in lending platforms, the bank does not compensate for repayment fraudulence, and the borrowers directly bear the losses. Therefore, it is essential to properly decide whether to accept or reject a loan application to minimize the damage. Credit scoring methods are helpful in this regard. Credit scoring refers to data-oriented approaches that help lending organizations decide whether to loan or not (Feng *et al.*, 2018). Models used for these approaches are mainly classification.

While requesting a loan in the platform, applicants must provide information about loan purposes, their personal and financial information. Using the data available about the loan and borrower's characteristics, credit scoring methods give predictions on the ability of an applicant to repay the loan. Both lenders and platforms need to distinguish the success and default drivers of a loan request to experience a safe investment. Not all features are of the same importance in training the model. Focusing on the most significant factors can improve the model performance, using less time and energy. It also helps business owners gain a better perception of their users' behavioral patterns.

Crowdfunding is a broad field of study. Comparing the two financially profitable categories (equity-based and lending-based crowdfunding), we can say that the evaluation of actors' behavior in lending market has not received enough attention in the literature. Furthermore, the use of credit scoring approaches is mainly directed towards B2C and B2B scenarios. Few studies have applied credit scoring techniques in lending-based crowdfunding. The insufficiency of previous works addressing credit scoring methods in lending-based crowdfunding scenarios represents the research gap. Accounting for 98 percent of global activity, China, The United States, and the United Kingdom are known as leaders in the P2P lending marketplace (Suryono, *et al.*, 2019). Most studies in this field have used data from well-known, rich data sets such as Lending Club, Prosper, and several datasets belonging to Chinese platforms. However, the lending market is experiencing a growing trend in other countries, including countries in Europe and Asia. In the last few years, instances of P2P lending platforms have been emerged in Iran. Although, no academic research has studied actors' behavior in Iranian lending platforms so far.

To fill the gap, we raised the questions of how can we use data-oriented approaches to predict applicants' loan repayment default? What features should we focus on? To answer these questions, we proposed a credit scoring model on a new data set from a P2P lending platform in Iran. Data used in this report has not yet appeared in any other academic study. Feature selection, data cleaning, and preparation are among the key activities performed in this study. Extracting and adding new features (feature engineering) improved the accuracy of the classification model by almost 20 percent. We identified the 4 most significant factors in our prediction model. Lenders in this platform are businesses, each having a different risk tolerance threshold. A default probability is computed for each loan request to help lenders make decisions based on their own risk tolerance.

The remainder of this paper is organized as follows: In section 2 we review the related concepts and literature. In section 3 we describe the methodology of this research and the data used to establish our credit scoring model. In section 4 we provide our proposed credit scoring model and suggest a bidding system for decision making. In section 5 we refer to the significance of this work. In section 6 we represent a brief description of the results of this study.

2-Related literature

As mentioned, crowdfunding is divided into four general categories: donation, reward, equity, and lending. Below we review the research done in each section. A brief overview of the credit scoring literature is then provided to illustrate the research gap.

2-1-Donation-based crowdfunding

In this type of crowdfunding, donates are collected to fund charitable projects expecting no financial gain in return (Xu, 2018). Berliner and Kenworthy mentioned that donation-based crowdfunding is known as a popular strategy to address extravagant healthcare expenses. their studies showed that arranging narrative literacies in a way that illustrates the deservingness of charity can encourage individuals to make donations (Berliner and Kenworthy, 2017). In a similar survey, Majumdar and Bose found that rational and credible charity request messages receive more donations (Majumdar and Bose, 2018). Xu investigated the effect of using media modalities in appeals descriptions. He found that posting videos and pictures generally lead to more donations, but the effects differ depending on the category of projects (Xu, 2018). Robiady, et al. showed that there is a positive relation between proper storytelling and donors' engagement in donation performance (Robiady, et al., 2020).

2-2-Reward-based crowdfunding

In this type of crowdfunding, donors expect to receive non-financial rewards (e.g., goods and services) with a delay after their donation (Raab et al., 2020).

Using data from the KissKissBankBank (KKBB) –crowdfunding platform in France, specialized in non-monetary rewards- Petitjean examined factors attracting financial support for reward-based campaigns. He found that the presence of a video, previous success rate of the project category, early contributions, and the number of comments left by the backers are among the success factors (Petitjean, 2018). Anglin *et al.* used data from Kickstarter and suggested that an individual or organization's positive psychological capital (i.e., hope, optimism, confidence) language can help gain more supports (Anglin *et al.*, 2018). Wang and Yang recognized that product innovativeness, perceived product quality, creator ability, and webpage visual design can positively influence backers' intentions to support projects (Wang and Yang, 2019). Raab *et al.* raised the question of how facial expressions can affect funding decisions. Their investigation showed that sad and happy facial expressions positively influence backers, while intense emotional expressions discourage them (Raab *et al.*, 2020).

2-3-Equity-based crowdfunding

This type of crowdfunding contains the sale of shares to investors as a way of raising funds (Lukkarinen et al., 2016). Cordova, et al. explored the correlation between factors related to fundraising success. They concluded that project success is positively correlated with the project duration, while it shows a negative correlation with the expected raised funds (Cordova, et al., 2015). In 2015, Hörisch compared the success rate of environmental-oriented projects with other than that. He claimed that surprisingly no positive correlation is observed between being environmentally sustainable and crowdfunding success of the project (Hörisch, 2015). In contrast, in a 2020 survey, Hörisch and Tenner found that higher levels of environmental orientation increase the likelihood of successful fundraising, especially in the United States. However, they did not find any significant effect of social orientation on boosting chances of funding (Hörisch and Tenner, 2020). Lukkarinen et al. highlighted the features of pre-selected campaigns and the usage of public and private networks as being related factors in the success of campaigns (Lukkarinen et al., 2016). Yuan, *et al.* mentioned that previous works have mainly addressed project goals, durations, and categories. Yuan, *et al.* implemented sentiment analysis to extract the linguistic sentiments underlying the textual descriptions of fundraising projects (Yuan, et al., 2016). Jiang *et al.* investigated the impact of both hard information and soft information on fundraising performance. They found that the hard information such as the creator's experience, comments quantity, the number of backers, positively affect crowdfunding success. To access soft information, they applied Latent Dirichlet Allocation (LDA) to projects' descriptions, which extracted project topics, and they used sentiment analysis on comments to extract investors' perception. they showed that positive sentiment in comments is correlated with crowdfunding success (Jiang et al., 2020). Johan and Zhang suggested that providing detailed qualitative information about business models, strategies, marketing, and milestones is attractive to investors. However, sharing too much description discourages funders (Johan and Zhang, 2020). Similarly, Thapa tried to find the best amount of information sharing (pictures, videos, and text) to maximize investors' contribution and found that the amount of information has a curvilinear relationship with the probability of success (Thapa, 2020). Again, in a similar way, Liu, *et al*. found that the amount of information published on social networks under a certain threshold has a positive effect on capital attraction, while higher volumes negatively affect funders. They also reported a positive effect of an entrepreneur's reputation and project-sharing cascades on fundraising (Liu, et al., 2021).

2-4-Lending-based crowdfunding

This type of crowdfunding contains lending money as a way of investment, expecting to get paid back with interest (Byanjankar, Heikkila and Mezei, 2015). Zhang et al. conducted an empirical study on Paipaidai, the largest lending platform in China. They examined the determining factors of accepting a loan on this platform. They found that annual interest rate, repayment period, description, credit grade, successful loan number, failed loan number, gender, and borrowed credit score are the most significant factors (Zhang et al., 2017). Nieto and Cinca raised the issue of profit scoring versus credit scoring. They explained that credit scoring systems estimate the probability of loan failure, while profit scoring focuses on the expected profitability prediction, measured by the internal rate of return. They concluded that the selection of borrowers based on the profit scoring system using multivariate regression shows better performance than the credit scoring system based on logistic regression. Lending Club data was used to conduct this study (Serrano-Cinca and Gutiérrez-Nieto, 2016). Xia et al. pointed out that far from reality, most conventional loan evaluation models assume a balanced misclassification cost. Combining cost-effective learning and XG Boost, they proposed a cost-sensitive boosted tree loan evaluation model that assumes different misclassification costs for distinct classes. Instead of accuracy based evaluation, this study evaluated the model based on the expected profitability on two databases- Lending Club and wee.com. (Xia et al. , 2017). Ge *et al.* examined the predictive power of self-disclosed social media information on borrowers' failure in P2P lending. They identified social deterrence as a fundamental mechanism that reduces customers' failure rate and also increases the likelihood of repayment after disclosure. The study used a combined database consisting a P2P lending data and data published on a popular social network in China (Ge et al., 2017). Alomari and Fingerman applied association rule mining on Lending Club data aiming to find meaningful rules between different characteristics of applicants in P2P lending (Alomari and Fingerman, 2017). Polena and Regner (Polena and Regner, 2018) studied the determinants of borrower failure on a P2P platform on Lending Club data. They defined four risk classes and tested the significance of the determinants within each risk class. They found that for most variables, the determinant significance depends on the data class, and only a small number of variables are of great decisive importance in all risk classes. Kozodoi *et al.* extended the use of profit measures to feature selection and developed a multi-objective framework to address both profitability and comprehensibility. Applying on nine datasets- including Lending Club data- they concluded that their proposed approach outperforms previous feature selection strategies while using fewer features (Kozodoi et al., 2019).

2-5-Credit scoring

Credit scoring methods help to properly decide whether to accept or reject a loan application to minimize the losses. Most research on credit scoring has been focused on the B2C market: In a 2007 study, Huang, *et al.* Used three strategies to build credit scoring models based on hybrid support vector machines to evaluate an applicant's credit score through his or her characteristics (Huang, et al., 2007). Martens *et al.* noted the incomprehensibility of the support vector machine method due to its complex mathematical functions and used a rule extraction technique to maintain comprehensibility (Martens et al., 2007). Bekhet and Eletter proposed two credit scoring models to help Jordan's bank lending decisions (Bekhet and Eletter, 2014). Alborzi and Khanbabaei introduced a new hybrid model of behavioral scoring and credit scoring based on neural networks in the banking sector (Alborzi and Khanbabaei, 2016). Bhatia *et al.* Applied a wide range of statistical methods to credit scoring, including linear discriminant analysis, random forest, logistic regression, and XG Boost (Bhatia et al., 2017). Chopra and Bhilare conducted an experimental investigation to study bank customers loan failure using the decision tree as a base learner and compare it with ensemble tree learning methods such as bagging, boosting, and random forests (Chopra and Bhilare, 2018). Feng *et al.* introduced a new dynamic ensemble classification method based on soft probabilities for credit scoring, considering the related costs of type 1 and type 2 error. (Feng et al., 2018). Moradi and Mokhatab Rafiei developed a dynamic model that can model political-economic fluctuations (Moradi and Mokhatab Rafiei, 2019). Lucas *et al.* Considered credit card transactions as a sequence rather than single events (Lucas et al., 2020). Some studies have also applied credit scoring in the B2B market: Moon and Sohn proposed a new technology evaluation model taking into account both technology-related characteristics and environmental circumstances such as firm-specific characteristics and economic conditions (Moon and Sohn, 2010). Liang, et al. examined the effect of feature selection on prediction models using different classifiers (Liang, et al., 2015). Ju, et al., provided a framework for applying the time-varying Cox Hazard Proportional model to technology-based credit scoring (Ju, et al., 2015).

Comparing the two financially profitable categories (equity-based and lending-based crowdfunding), we can say that the evaluation of actors' behavior in the lending market has not received enough attention in the literature. Although different methods have been developed in recent years to address actor's behavior in the context of marketing and finance, applying these methods to enrich lending-based crowdfunding scenarios seems to be neglected. Furthermore, the use of credit scoring approaches is mainly directed towards B2C and B2B scenarios, and a few studies have employed credit scoring techniques in lending-based crowdfunding. In other words, applying credit scoring approaches within emerging business platforms, which facilitate actor-to-actor collaborations, has not been addressed sufficiently. In this way, in our research, we focus on the issue that how credit scoring approaches can be applied within lending-based scenarios in actor-to-actor business platforms.

3-Methods and data

We applied the CRISP-DM[1] methodology to form stages of this study with a standard and structured approach. This framework outlines common procedures used by data specialists and is the most widely-used analytics model. CRISP-Dm helps to simplify complex and large projects by taking simple data mining tasks. It provides guidelines to best practices, documentation and also allows the research to be replicated (*Why using CRISP-DM will make you a better Data Scientist?*, 2020). Figure 1 shows the steps of this research process, mapped on the CRISP-DM model. Evaluation metrics are chosen according to table 1 (Sokolova and Lapalme, 2009).
This study is coded with Python 3.7.3 in Jupyter Notebook as an IDE and presentation tool.

Table1. Evaluation measures

Measure	Calculation	Focus
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	Effectiveness of a classifier
Precision	$\frac{TP}{TP+FP}$	Class agreement of the data labels with the positive labels
Recall	$\frac{TP}{TP+FN}$	Effectiveness of a classifier to identify positive labels
F-measure	$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$	Relations between positive labels and those given by a classifier
AUC	$1/2(\frac{TP}{TP+FN}+\frac{TN}{TN+FP})$	The model’s ability to distinguish between classes

[1] Cross-industry process for data mining

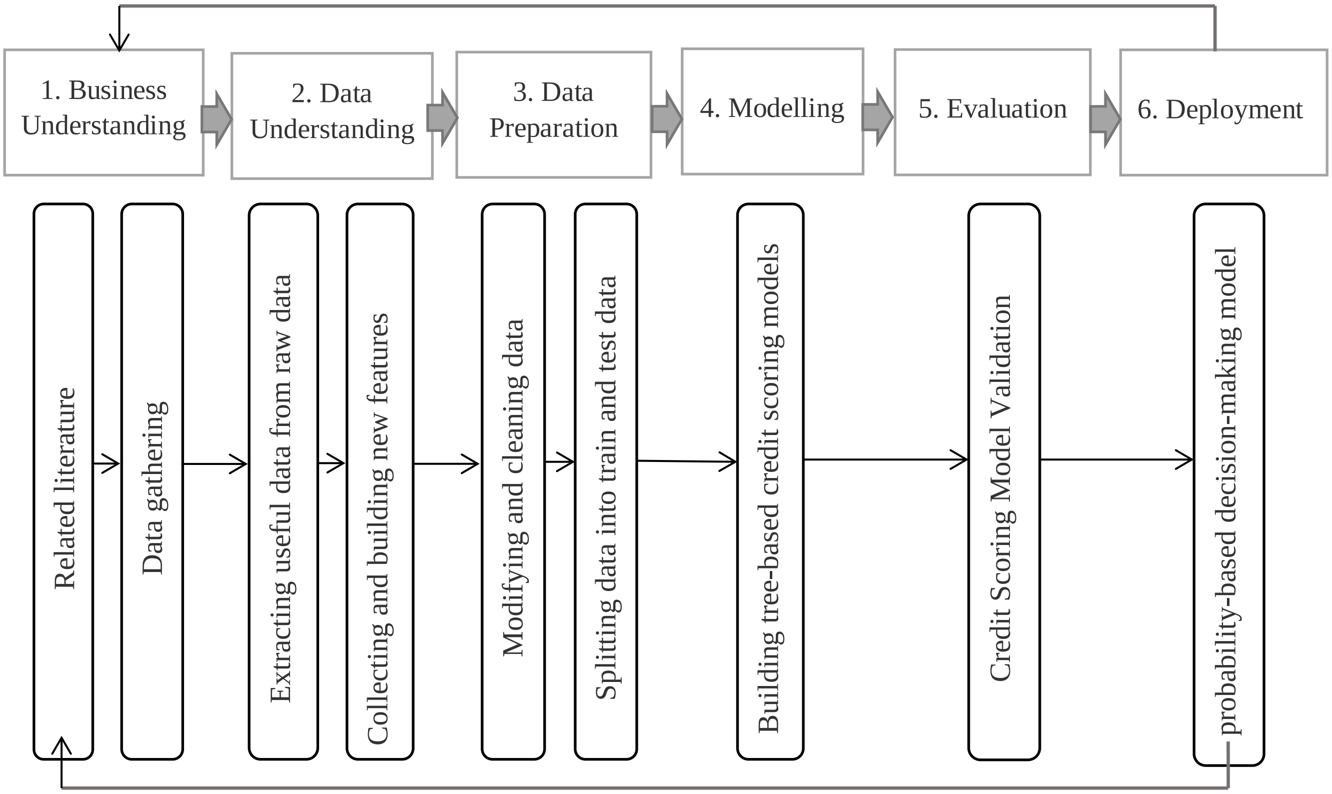

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment
Related literature
Data gathering
Extracting useful data from raw data
Collecting and building new features
Modifying and cleaning data
Splitting data into train and test data
Building tree-based credit scoring models
Credit Scoring Model Validation
probability-based decision-making model

Fig1. Research process

Related literature was discussed at the previous section.

Data gathering: Data from an Iranian P2P lending platform was obtained and used in this research. The name of the platform is not divulged to preserve the requirements of the company. In this platform, lenders are legal individuals, and borrowers are real people. Credit is allocated to borrowers for the purpose of goods' purchasing. Features of this database include demographic characteristics of borrowers -including age, gender, income, occupation, place of residence- traits of lenders -including service categorization- loan characteristics -such as loan amount, the number of loan installments, the amount of each installment, invoice date, payment date. Some of the features of this database are continuous numeric, and the rest are nominal. This categorizes our database into mixed-type data.

Extracting useful data from raw data: To extract useful data from raw data, the installment table was selected as the base table. Each of the records in this table represents an installment's characteristics and payment status. Records in the canceled loan orders table were identified and deleted in the base table using their order ID. Related features from different database tables were added to the base table and integrated with one another. To determine the final status of installment payments, only the installments whose invoice-date was before 2020-01-15 were chosen. Installments that had not been paid until then, were considered unpaid.

Collecting and building new features: Very little information was available about lenders. Information about each lender's income and size of their businesses was extracted from the Alexa website and added to the database. The difference between the invoice date and the payment date was added as a new feature. According to this feature, a label was created for each record. Label 1 introduces the installments paid on time, and label 0 represents the installments paid with a delay. Other features added to the database include the number of loans taken from each lender, the average amount borrowed from each lender, the average delay per borrower-if they had a previous history of borrowing in the system, the average amount of loans taken by the borrower, the average delay per lender, the invoice month.

Modifying and cleaning data: Due to the interdependence of some features, a number of features were expected to be correlated. Therefore, a correlation heat-map was drawn taking into account all the features. As can be seen in figure 2, some features were highly correlated. To remove strongly correlated features, one of the two attributes having a correlation greater than |0.7| was deleted. After removing the correlated features, the correlation heat-map was obtained as figure 3. The number of features was reduced from 52 to 21 after removal. There were missing values in a few features. Incomplete records were deleted and the number of records decreased from 148509 to 147938.

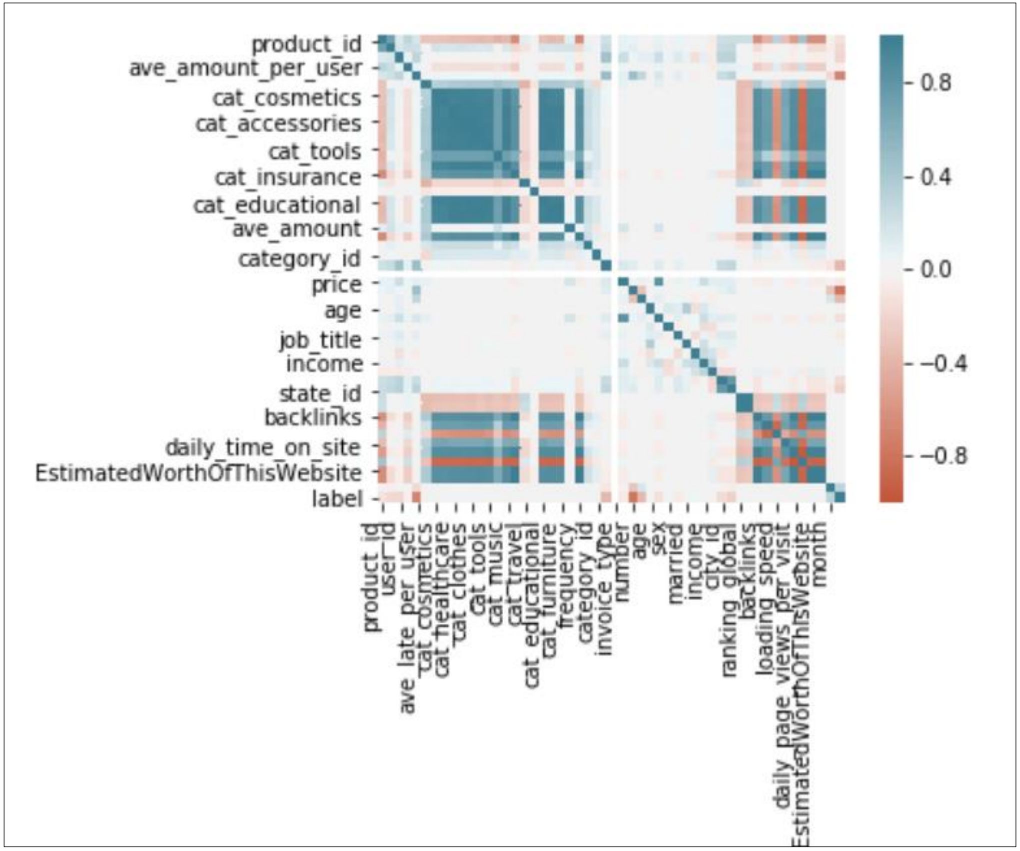


Fig 2. Correlation heat-map before removal of correlated features

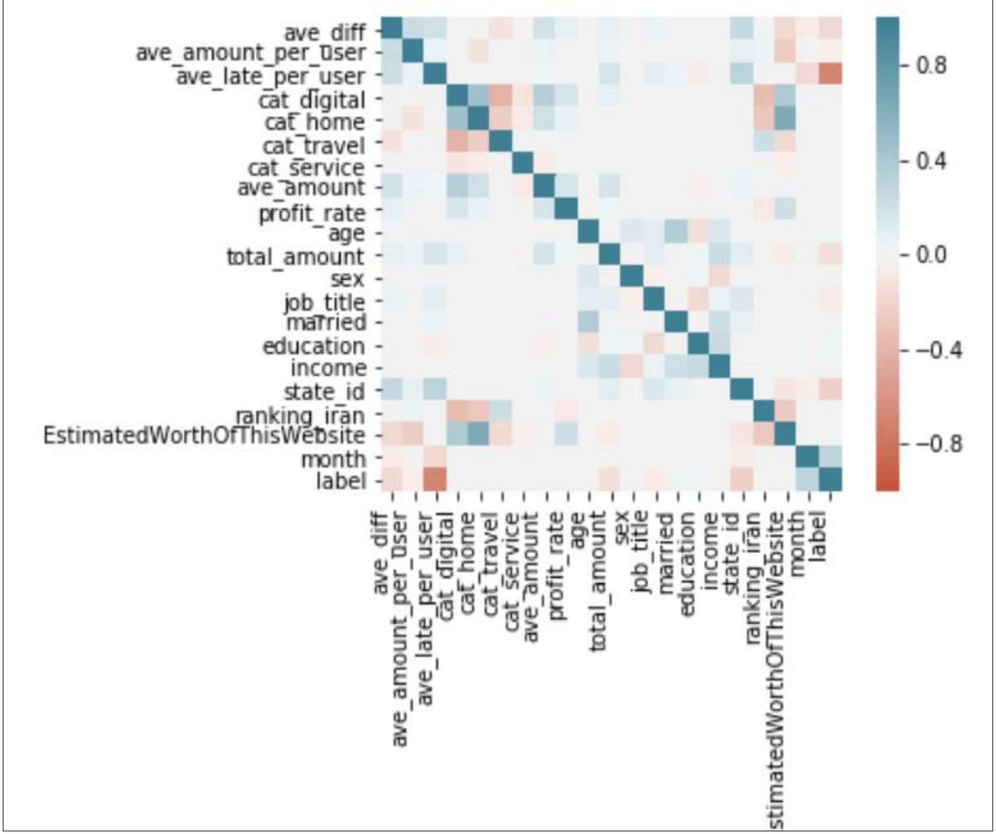


Fig 3. Correlation heat-map before removal of correlated features

Splitting data into train and test data: The prepared database was then randomly divided into two parts, with a ratio of 30% for testing and 70% for training. Five tree-based classification methods were implemented on the data, and the results were analyzed: Decision Tree, Random Forests, XGBoost, AdaBoost, and Extra Trees.

Note: Appendix B provides accordance of features names in figures with the text.

4-Results

Building tree-based credit scoring models: Features in the final dataset include average delay per lender, average amount of loans taken by the borrower, average delay per borrower, average amount borrowed from each lender, interest rate, lender’s income, service category, loan amount, invoice month, borrower's age, borrower's gender, borrower's job, borrower's marital status, borrower's education, borrower's location. Trials showed that tree-based classifiers give better results than other classification methods. This phenomenon is explained due to the mixed nature of the data. Five tree-based classification methods were implemented on the data: Decision Tree, Random Forests, XGBoost, AdaBoost, and Extra Trees.

The Decision Tree classifier has close proximity to the human brain decision-making system. With a flowchart-like structure representing a series of questions and their possible answers, decision trees divide the original dataset into subsets, each purer than their parent set (Haponik, 2020). The purest subsets at ending nodes determine the route's label. Generally, Tree-based models can handle numerical, categorical, and mixed-type data and therefore are proper to use in a wide range of problems. They also perform well and fast on large datasets with a low effort needed for data preparation during data pre-processing, as scaling and normalization. (Dhiraj, 2019).

Though simple decision trees are well known and easy to interpret, they suffer from instability and overfitting in some cases (Random forest: many are better than one, 2017). To address these issues, ensemble models -bagging and boosting- are introduced. In bagging, several same weak predictors are trained independently, and the final result is obtained by merging all predictions with equal weights. Differently in boosting, each weak predictor is affected by its previous trained model, and to combine, each predictor receives a weight according to its performance (Vadapalli, 2020).

Random Forest (RF) and Extra Trees (ET) are two algorithms of bagging ensemble models. In RF, each tree is trained on a bootstrap sample, which subsamples instances with replacement, and the best distinguishing feature is selected from a random subspace of all attributes. In contrast, ET takes the whole dataset to train each tree and chooses the cut points randomly from features space. Both algorithms use a majority voting system, though having lower computational costs, makes ET faster. AdaBoost and XGBoost are among boosting ensemble models. Adaboost algorithm grows weak learners -trees- and sequentially assigns higher weights to misclassified samples. Each predictor's worth is calculated based on modified weights, and the class label having the vote of the most worthy predictors is chosen. While Adaboost has a few hyperparameters and is easy to visualize, XGBoost benefits from being considerably faster. The higher computational speed is achieved by narrowing the search space of the best distinguishing attribute, using several hyperparameters. Generally, tree-based ensemble algorithms are robust against overfitting, noise, and outliers, are relatively computationally inexpensive, and usually perform better than their weak base learners (Nikulski, 2020).

4-1-Validation measures

Credit Scoring Model Validation: Validation measures of all five models were calculated and compared before and after the addition of added features. The most often evaluation measures that can be calculated according to the confusion matrix are introduced in table 1.

Figure 4 shows the accuracy criteria for the implemented classification models. The accuracy of all models has increased by almost 20% after the addition of added features. As can be seen, the model made using Random Forests gives the best accuracy. The other evaluation measures in table 1 are provided in Appendix A.

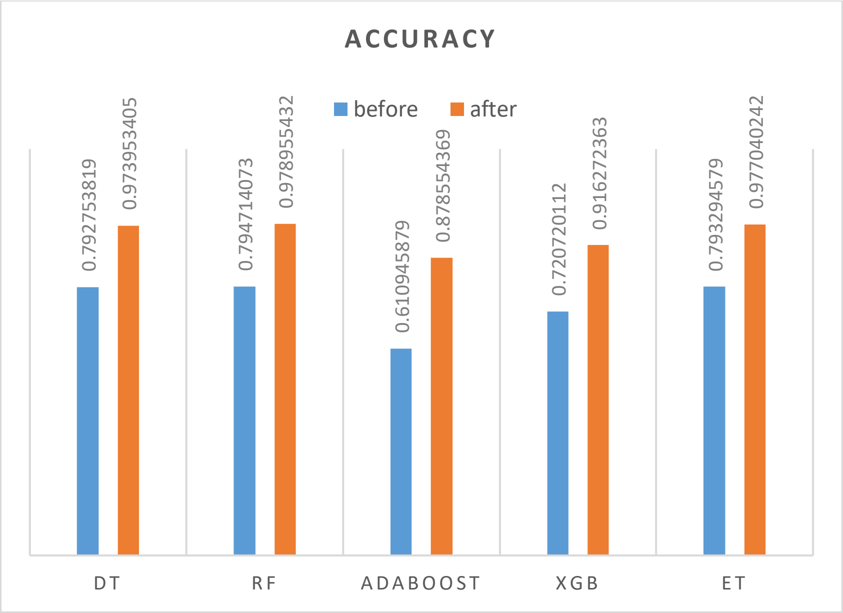


Fig 4. Comparison of the classification models accuracy before and after adding new features

4-2- Feature importance

To discover the significance of each predictor variable, the Feature Importance diagram was plotted based on our best classifier, Random Forests. Knowing the importance of predictive features in predictive models has the following advantages (Chandrashekar and Sahin, 2014):

- With a better understanding of model logic, one can focus only on important variables to improve the model.

- A number of variables that are not important can be omitted and a similar or even better performance of the prediction model can be observed in a shorter time and at a lower cost.

- It helps business owners interpret data.

According to figure 5, “average delay per borrower “and "invoice month" are respectively by far the most significant features in determining repayment default. Thus, loan repayment depends firstly on the borrowers' payment history and secondly on the economic and social conditions at the loan maturity date. After that, "loan amount", "the average amount of loans taken by the borrower" and "borrower’s age" are the most interpretative attributes.

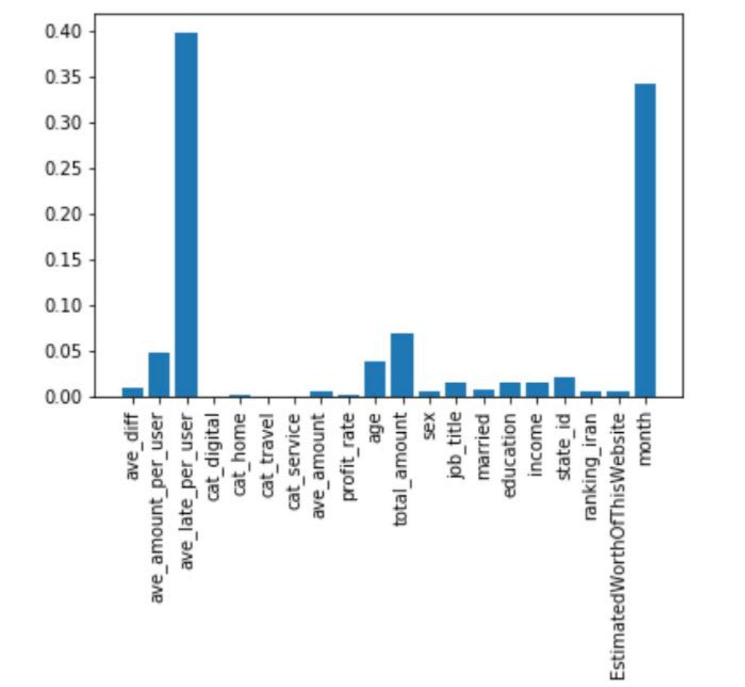
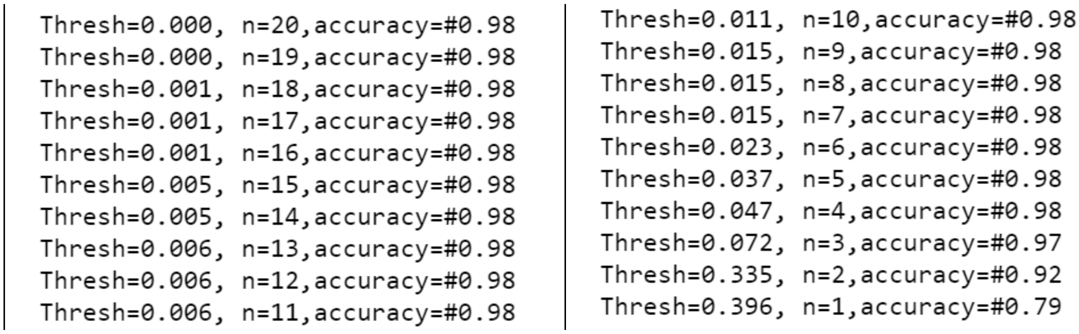


Fig 5. Feature Importance

4-3- Feature selection

Some of the features used did not have any effect on increasing the model accuracy. Table 2 shows three parameters: (1) threshold, (2) number of features, (3) accuracy. The threshold parameter indicates a boundary value that "n" number of attributes have importance higher than this value. The accuracy parameter depicts the accuracy of the model with an "n" number of features.

Table 2. Feature selection



```
Thresh=0.000, n=20,accuracy=#0.98    Thresh=0.011, n=10,accuracy=#0.98
Thresh=0.000, n=19,accuracy=#0.98    Thresh=0.015, n=9,accuracy=#0.98
Thresh=0.001, n=18,accuracy=#0.98    Thresh=0.015, n=8,accuracy=#0.98
Thresh=0.001, n=17,accuracy=#0.98    Thresh=0.015, n=7,accuracy=#0.98
Thresh=0.001, n=16,accuracy=#0.98    Thresh=0.023, n=6,accuracy=#0.98
Thresh=0.005, n=15,accuracy=#0.98    Thresh=0.037, n=5,accuracy=#0.98
Thresh=0.005, n=14,accuracy=#0.98    Thresh=0.047, n=4,accuracy=#0.98
Thresh=0.006, n=13,accuracy=#0.98    Thresh=0.072, n=3,accuracy=#0.97
Thresh=0.006, n=12,accuracy=#0.98    Thresh=0.335, n=2,accuracy=#0.92
Thresh=0.006, n=11,accuracy=#0.98    Thresh=0.396, n=1,accuracy=#0.79
```

According to table 2, with only four features, the prediction accuracy remains the same at 98%. Therefore, only by having the most important four determining features, namely (1) average delay per borrower, (2) invoice month, (3) loan amount, and (4) the average amount of loans taken by the borrower, the same accuracy may be achieved.

4-4-Possibility of loan repayment

Probability-based decision-making model: The Random Forests model was used to decide whether to accept or reject a particular loan order. Due to the high accuracy of this classifier, the trained model was implemented on the whole dataset (not just test data). This time, instead of taking the output as a binary label, the model output was calculated probabilistically. An average repayment probability was calculated for all installments related to each loan.

This probability indicates the risk of lending. As mentioned, the lenders in this platform are businesses. The proposed probability helps them make better decisions according to their own risk tolerance. Businesses may also consider their set objectives in the decisional process. This objective can vary from minimizing investment risk to an extent of risk accepting aiming at maximizing profitability.

To the best of our knowledge, this is the first time that lenders' data on a P2P lending platform is being involved in the loan default prediction. Due to our limitations in obtaining lenders' data, this approach may be considered in future studies more effectively.

5-Discussion

5-1- Practical implication

Applying proper decision-making techniques has always been an issue in all types of crowdfunding. Especially in reward-based and loan-based crowdfunding, which go along with financial benefits, the issue of decision-making becomes more significant. Looking at the literature on crowdfunding, we found that the work done in the context of loan-based crowdfunding -the P2P lending market- is not rich enough. Thus, by combining the concepts of crowdfunding and credit scoring in various market scenarios, we addressed the application of credit scoring methods in a loan-based crowdfunding scenario.

The use of intelligent decision-making systems in lending platforms leads to the growth and improvement of the P2P lending market. Credit scoring systems help lenders experience a safe investment. Although platforms are not responsible for compensating lenders, the occurrence of any loan fraud tarnishes their reputation. Due to the recent emergence of P2P lending in Iran, achieving an accurate credit scoring system is a shared goal of both platforms and lenders.

In credit risk evaluation, having access to relevant and up-to-date information is one of the main concerns. In this study, our main effort was to extract, build and collect useful information from the raw dataset. It was shown that the added features increased the accuracy of our prediction model by about 20%.

The platform considered in this research and any similar crowdfunding system will benefit from the results of this research. This platform can inform its lenders of the possible risk of loans non-repayment to help them make the best decision according to their risk tolerance. In this study, four variables are identified as the most important predictors: average delay per borrower, invoice month, loan amount, and the average amount of loans taken by the borrower. By focusing on gathering accurate data on the mentioned attributes, the platform can make accurate predictions in a shorter time and at a lower cost.

5-2- Scholarly implication

This study also adds to the growing stream of research on the role of credit scoring in P2P lending platforms. To the best of our knowledge, this is the first time that lenders' data are being involved in making predictions on loan default in a P2P lending platform. Due to our limitations in obtaining lenders' data, this approach may be considered in future studies more effectively.

6-Conclusions

This paper proposes a data-driven credit scoring model for evaluating loan applicants in a loan-based crowdfunding scenario in Iran. Most studies in the field of loan-based crowdfunding have used publicly available datasets such as Landing Club and Prosper. In this respect, this study is one of the first studies conducted on an Iranian P2P lending platform. Collecting data, pre-processing and feature engineering have been the main challenges of this research. Several new features are extracted and added to the original database to improve the model's performance. The classification accuracy increased by about 20% using new added features. As expected, due to the mixed nature of data, tree-based classifiers obtained the best results amongst the models we tried. Among the five tree-based implemented classifiers, Random Forest gives the best classification accuracy with a value of 97.8%. Four attributes are identified as the most important defining features, three of which are extracted during pre-processing phase: Average delay per borrower, invoice month, loan amount, and the average amount of loans taken by the borrower. Most of the credit scoring literature has analyzed this problem in a binary format. As mentioned, lenders in our dataset are corporations (versus individuals). Each of these corporations has its own risk tolerance threshold, hence accepts or rejects loans accordingly. To help them make the right decisions based on their risk tolerance, we proposed probability measures instead of using binary outputs.

Information asymmetry is one of the limitations of this research, which provides directions for future research. Little information is available about platform lender firms. Having access to data about lending organizations can help us calculate lenders' risk tolerance thresholds and provide more precise decisional suggestions.

References

Alborzi, M. and Khanbabaei, M. (2016) ‘Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method’, *International Journal of Business Information Systems*, 23(1), pp. 1–22. doi: 10.1504/IJBIS.2016.078020.

Alomari, Z. and Fingerman, D. (2017) ‘Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications’, *New Zealand Journal of Computer-Human Interaction*.

Anglin, A. H. *et al.* (2018) ‘The power of positivity? The influence of positive psychological capital language on crowdfunding performance’, *Journal of Business Venturing*, 33(4), pp. 470–492. doi: 10.1016/j.jbusvent.2018.03.003.

Bekhet, H. A. and Eletter, S. F. K. (2014) ‘Credit risk assessment model for Jordanian commercial banks: Neural scoring approach’, *Review of Development Finance*, 4(1), pp. 20–28. doi: 10.1016/j.rdf.2014.03.002.

Belleflamme, P., Omrani, N. and Peitz, M. (2015) ‘The economics of crowdfunding platforms’, *Information Economics and Policy*, 33, pp. 11–28. doi: 10.1016/j.infoecopol.2015.08.003.

Berliner, L. S. and Kenworthy, N. J. (2017) ‘Producing a worthy illness: Personal crowdfunding amidst financial crisis’, *Social Science and Medicine*, 187, pp. 233–242. doi: 10.1016/j.socscimed.2017.02.008.

Bhatia, S. *et al.* (2017) ‘Credit Scoring using Machine Learning Techniques’, *International Journal of Computer Applications*, 161(11), pp. 1–4. doi: 10.5120/ijca2017912893.

Byanjankar, A., Heikkila, M. and Mezei, J. (2015) ‘Predicting credit risk in peer-to-peer lending: A neural network approach’, *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, pp. 719–725. doi: 10.1109/SSCI.2015.109.

Chandrashekar, G. and Sahin, F. (2014) ‘A survey on feature selection methods’, *Computers and Electrical Engineering*, 40(1), pp. 16–28. doi: 10.1016/j.compeleceng.2013.11.024.

Chopra, A. and Bhilare, P. (2018) ‘Application of Ensemble Models in Credit Scoring Models’, *Business Perspectives and Research*, 6(2), pp. 129–141. doi: 10.1177/2278533718765531.

Cordova, A., Dolci, J. and Gianfrate, G. (2015) 'The Determinants of Crowdfunding Success: Evidence from Technology Projects', *Procedia - Social and Behavioral Sciences*, 181, pp. 115–124. doi: 10.1016/j.sbspro.2015.04.872.

Dhiraj, K. (2019) *Top 5 advantages and disadvantages of Decision Tree Algorithm, medium*. Available at: https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a (Accessed: 5 June 2021).

Feng, X. *et al.* (2018) 'Dynamic ensemble classification for credit scoring using soft probability', *Applied Soft Computing Journal*, 65, pp. 139–151. doi: 10.1016/j.asoc.2018.01.021.

Ge, R. *et al.* (2017) 'Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending', *Journal of Management Information Systems*, 34(2), pp. 401–424. doi: 10.1080/07421222.2017.1334472.

Haponik, A. (2020) *Decision Tree Machine Learning Model*, *addepto*. Available at: https://addepto.com/decision-tree-machine-learning-model/ (Accessed: 5 June 2021).

Hörisch, J. (2015) 'Crowdfunding for environmental ventures: An empirical analysis of the influence of environmental orientation on the success of crowdfunding initiatives', *Journal of Cleaner Production*, 107, pp. 636–645. doi: 10.1016/j.jclepro.2015.05.046.

Hörisch, J. and Tenner, I. (2020) 'How environmental and social orientations influence the funding success of investment-based crowdfunding: The mediating role of the number of funders and the average funding amount', *Technological Forecasting and Social Change*, 161(April), p. 120311. doi: 10.1016/j.techfore.2020.120311.

Huang, C. L., Chen, M. C. and Wang, C. J. (2007) 'Credit scoring with a data mining approach based on support vector machines', *Expert Systems with Applications*, 33(4), pp. 847–856. doi: 10.1016/j.eswa.2006.07.007.

Jiang, C. *et al.* (2020) 'The impact of soft information extracted from descriptive text on crowdfunding performance', *Electronic Commerce Research and Applications*, 43(June), p. 101002. doi: 10.1016/j.elerap.2020.101002.

Johan, S. and Zhang, Y. (2020) 'Quality revealing versus overstating in equity crowdfunding', *Journal of Corporate Finance*, 65(September), p. 101741. doi: 10.1016/j.jcorpfin.2020.101741.

Ju, Y., Jeon, S. Y. and Sohn, S. Y. (2015) 'Behavioral technology credit scoring model with time-dependent covariates for stress test', *European Journal of Operational Research*, 242(3), pp. 910–919. doi: 10.1016/j.ejor.2014.10.054.

Klafft, M. (2008) 'Online peer-to-peer lending: A lenders' perspective', *Proceedings of the 2008 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government, EEE 2008*, (July), pp. 371–375. doi: 10.2139/ssrn.1352352.

Kozodoi, N. *et al.* (2019) 'A multi-objective approach for profit-driven feature selection in credit scoring', *Decision Support Systems*, 120(March), pp. 106–117. doi: 10.1016/j.dss.2019.03.011.

Liang, D., Tsai, C. F. and Wu, H. T. (2015) 'The effect of feature selection on financial distress prediction', *Knowledge-Based Systems*, 73(1), pp. 289–297. doi: 10.1016/j.knosys.2014.10.010.

Liu, Y., Chen, Y. and Fan, Z. P. (2021) 'Do social network crowds help fundraising campaigns? Effects of social influence on crowdfunding performance', *Journal of Business Research*, 122(February 2019), pp. 97–108. doi: 10.1016/j.jbusres.2020.08.052.

Lucas, Y. *et al.* (2020) 'Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs', *Future Generation Computer Systems*, 102, pp. 393–402. doi: 10.1016/j.future.2019.08.029.

Lukkarinen, A. *et al.* (2016) 'Success drivers of online equity crowdfunding campaigns', *Decision Support Systems*, 87, pp. 26–38. doi: 10.1016/j.dss.2016.04.006.

Majumdar, A. and Bose, I. (2018) 'My words for your pizza: An analysis of persuasive narratives in online crowdfunding', *Information and Management*, 55(6), pp. 781–794. doi: 10.1016/j.im.2018.03.007.

Martens, D. *et al.* (2007) 'Comprehensible credit scoring models using rule extraction from support vector machines', *European Journal of Operational Research*, 183(3), pp. 1466–1476. doi: 10.1016/j.ejor.2006.04.051.

Moon, T. H. and Sohn, S. Y. (2010) 'Technology credit scoring model considering both SME characteristics and economic conditions: The Korean case', *Journal of the Operational Research Society*, 61(4), pp. 666–675. doi: 10.1057/jors.2009.7.

Moradi, S. and Mokhatab Rafiei, F. (2019) 'A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks', *Financial Innovation*, 5(1). doi: 10.1186/s40854-019-0121-9.

Nikulski, J. (2020) *The Ultimate Guide to AdaBoost, random forests and XGBoost*, *Towards Data Science*. Available at: https://towardsdatascience.com/the-ultimate-guide-to-adaboost-random-forests-and-xgboost-7f9327061c4f (Accessed: 5 June 2021).

Patwardhan, A. (2018) *Peer-To-Peer Lending*, *Handbook of Blockchain, Digital Finance, and Inclusion, Volume 1: Cryptocurrency, FinTech, InsurTech, and Regulation*. Elsevier Inc. doi: 10.1016/B978-0-12-810441-5.00018-X.

Petitjean, M. (2018) 'What explains the success of reward-based crowdfunding campaigns as they unfold? Evidence from the French crowdfunding platform KissKissBankBank', *Finance Research Letters*, 26, pp. 9–14. doi: 10.1016/j.frl.2017.11.005.

Polena, M. and Regner, T. (2018) 'Determinants of borrowers' default in P2P lending under consideration of the loan risk class', *Games*, 9(4), pp. 1–17. doi: 10.3390/g9040082.

Raab, M. *et al.* (2020) 'More than a feeling: Investigating the contagious effect of facial emotional expressions on investment decisions in reward-based crowdfunding', *Decision Support Systems*, 135. doi: 10.1016/j.dss.2020.113326.

Random forest: many are better than one (2017) *QuantDare*. Available at: https://quantdare.com/random-forest-many-are-better-than-one/ (Accessed: 5 June 2021).

Robiady, N. D., Windasari, N. A. and Nita, A. (2020) 'Customer engagement in online social crowdfunding: The influence of storytelling technique on donation performance', *International Journal of Research in Marketing*, (xxxx), pp. 1–9. doi: 10.1016/j.ijresmar.2020.03.001.

Serrano-Cinca, C. and Gutiérrez-Nieto, B. (2016) 'The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending', *Decision Support Systems*, 89, pp. 113–122. doi: 10.1016/j.dss.2016.06.014.

Suryono, R. R., Purwandari, B. and Budi, I. (2019) 'Peer to peer (P2P) lending problems and potential solutions: A systematic literature review', *Procedia Computer Science*, 161, pp. 204–214. doi: 10.1016/j.procs.2019.11.116.

Thapa, N. (2020) 'Being cognizant of the amount of information: Curvilinear relationship between total-information and funding-success of crowdfunding campaigns', *Journal of Business Venturing Insights*, 14. doi: 10.1016/j.jbvi.2020.e00195.

Vadapalli, P. (2020) *Bagging vs Boosting in Machine Learning: Difference Between Bagging and Boosting, UpGrad Blog*. Available at: https://www.upgrad.com/blog/bagging-vs-boosting/.

Wang, Z. and Yang, X. (2019) 'Understanding backers' funding intention in reward crowdfunding: An elaboration likelihood perspective', *Technology in Society*, 58. doi: 10.1016/j.techsoc.2019.101149.

Why using CRISP-DM will make you a better Data Scientist? (2020) *Great Learning*. Available at: https://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist/ (Accessed: 6 June 2021).

Xia, Y., Liu, C. and Liu, N. (2017) 'Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending', *Electronic Commerce Research and Applications*, 24, pp. 30–49. doi: 10.1016/j.elerap.2017.06.004.

Xu, L. Z. (2018) 'Will a digital camera cure your sick puppy? Modality and category effects in donation-based crowdfunding', *Telematics and Informatics*, 35(7), pp. 1914–1924. doi: 10.1016/j.tele.2018.06.004.

Yuan, H., Lau, R. Y. K. and Xu, W. (2016) 'The determinants of crowdfunding success: A semantic text analytics approach', *Decision Support Systems*, 91, pp. 67–76. doi: 10.1016/j.dss.2016.08.001.

Zhang, Y. *et al.* (2017) 'Determinants of loan funded successful in online P2P Lending', *Procedia Computer Science*, 122, pp. 896–901. doi: 10.1016/j.procs.2017.11.452.

Appendix A. Comparing evaluation measures for five classifiers

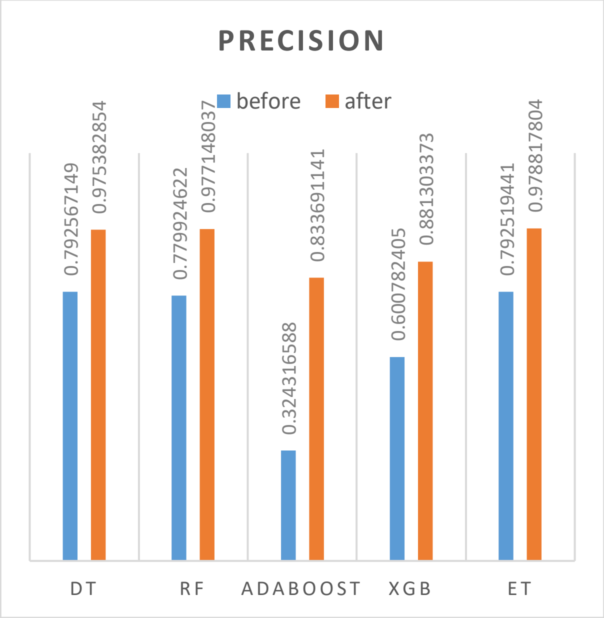


Fig 6. Comparison of the classification models precision before and after adding new features

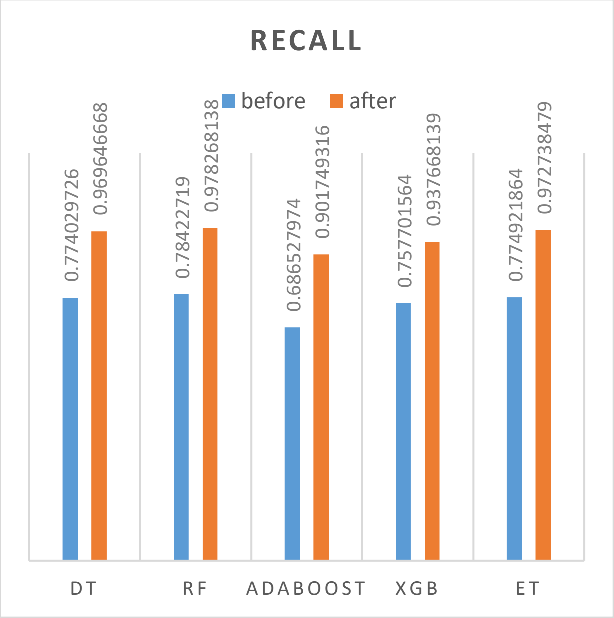


Fig 7. Comparison of the classification models recall before and after adding new features

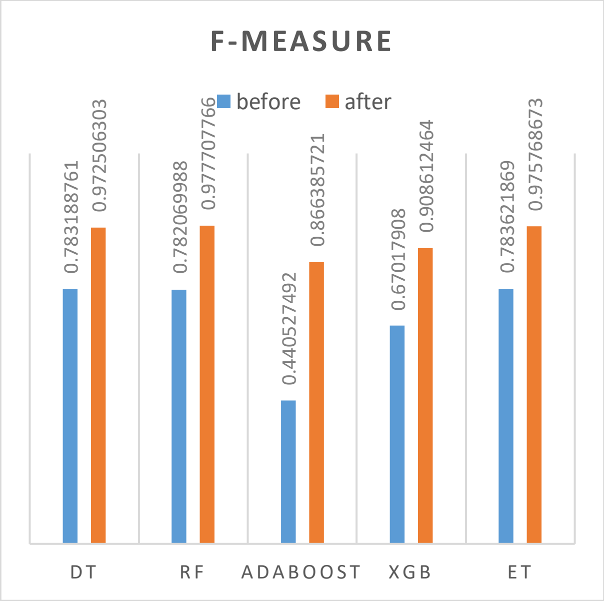


Fig 8. Comparison of the classification models F-measure before and after adding new features

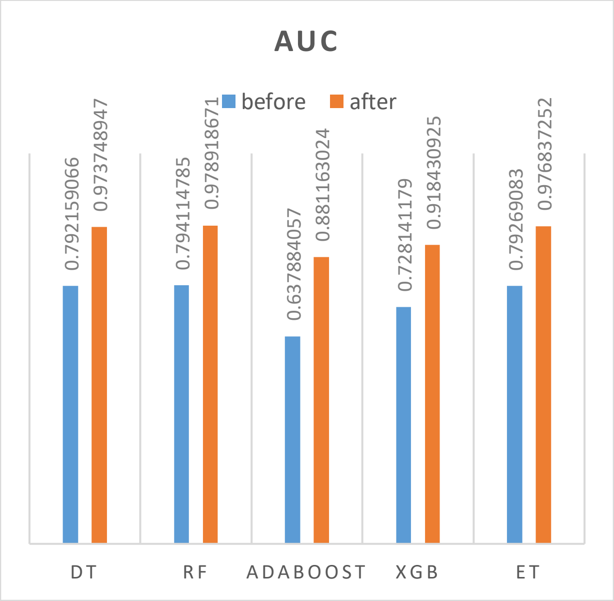


Fig 9. Comparison of the classification models AUC before and after adding new features

Appendix B. Description for attributes' names of figures in accordance to the text

Figure	Text
ave_amount_per_user	the average amount of loans taken by the borrower
ave_amount	the average amount borrowed from each lender
EstimatedWorthOfThisWebsite	lender's income
month	invoice month
age	Borrower's age
sex	borrower's gender
married	Borrower's marital status
education	Borrower's education
state_id	Borrower's location
job_title	Borrower's job
ave_diff	average delay per lender
ave_late_per_user	average delay per borrower
profit_rate	interest rate
cat_digital	service category
cat_home	
cat_travel	
cat_service	
total_amount	loan amount