ISSN 2714-7533

Predicting Peer to Peer Lending Loan Risk Using Classification Approach

Fahmi Zulfikri [a,1], Dendy Tryanda [a,2], Allevia Syarif [a,3], Harry Patria [a,4,*]

[a] Department of Accounting, Faculty of Economic and Business, University of Indonesia, Depok, Indonesia
[1] fahmi.zulfikri@ui.ac.id, [2] dendy.tryanda@ui.ac.id, [3]allevia.syarif @ui.ac.id, [4]harry.patria@sbm-itb.ac.id
* corresponding author

ARTICLE INFO

Article history
Received June 10, 2021
Revised July 7, 2021
Accepted August 28, 2021

Keywords
Borrower profiling
classification
P2P lending
random forest
decision tree

ABSTRACT

Technological innovations have affected all sectors of life, especially, the financial sector with the emergence of financial technology. One of them is marked by the emergence of Peer-to-Peer Lending ("P2P Lending). Credit Risk Management is essential to P2P Lending as it directly affects business results, therefore it is important for P2P Lending to predict borrowers with the highest probability to become good or bad loans based on their profile or characteristics. In the experiments, five classification algorithms are used, which are Gradient Boosted Trees, Naïve Bayes, Random Forest, Decision Tree and Logistic Regression. The result is two modelling performed well that is Random Forest with accuracy 93.38% and Decision Tree with 92.35%.

This is an open access article under the CC–BY-SA license.

1. Introduction

Technological innovations have affected all sectors of life, especially the financial sector, with the emergence of financial technology. One of them is marked by the emergence of Peer-to-Peer Lending (P2P Lending). Peer-to-peer (P2P) lending is a type of micro-financing activity conducted through an online platform, by matching people who have money to invest with people who are looking for a loan. P2P lending typically involves a service provider, through an online platform or a mobile application, acting as a middleman to link investors (or lenders) with borrowers looking for capital. United States P2P platforms issued loans worth $889 million in 2012, $2.9 billion in 2013 and $6.6 billion in 2014. By 2016, P2P lending in the US reached $32.8 billion. In the financial services industry, P2P first emerged in 2005 with a focus on lending and borrowing.

Lending Club is a fintech company that provides a range of financial products and services through a technology-driven platform in the United States. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market. Lending Club was the world's largest peer-to-peer lending platform before abandoning the peer-to-peer lending model in the fall of 2020. With Total assets: $4.641 billion (2017) and Total equity: $922.5 million (2017).

Credit Risk Management is essential to P2P lending as it directly affects business results; therefore, it is essential for P2P lending to predict or assess whether an individual would be capable of repaying the loan. The loan granting decisions consider various factors, including the character of borrower's, the capacity of the borrower to repay a loan, the conditions of the loan, such as the interest rate and the amount of principal. The dataset research is collected from Lending Club, with 7,60% of all the

borrowers have been marked as bad loans, which means that the investors had lost $516,981,729 and potentially would increase to $991,386,147.

The proposed work performs the classification task by distinguishing whether the potential borrower will become a bad loan or good loan. The classification task is performed using algorithms like Gradient Boosted Trees, Random Forest, Naive Bayes, Decision Trees, and Logistic Regression, and the proposed system we use is KNIME (Konstanz Information Miner).

2. Data

2.1. Data Description

Our research dataset is collected from Lending Club, P2P Lending based in the United States with some modifications in the data while the data is collected from a public data platform (www.kaggle.com). This research uses 887,379 loan applications from 1 June 2007 to 31 December 2015 with private information about the borrowers such as employment length, annual income, loan amount, dti ratio, total payment and borrowers grade. The summary of the loan condition is presented in Table 1. As shown in Table 1; 7,60% of all the borrowers have been marked as bad loans while 92,40% are good loan. Bad loan is a loan with status is charged off, default, in grace period, or late for payment (over 16 days).

Table 1. Loan Condition

Condition of Loan	Number	Percentage
Good Loan	819,950	92.40%
Bad Loan	67,429	7.60%

2.2. Data Set

Ferozi, Muhammad Nadeem (2018) by using random forest algorithm found that 7 attributes that have importance value is above 0.05 (figure 1) there is “employee length”, “annual income”, “loan amount”, “interest rate”, “dti ratio”, “total payment”, and “year”.

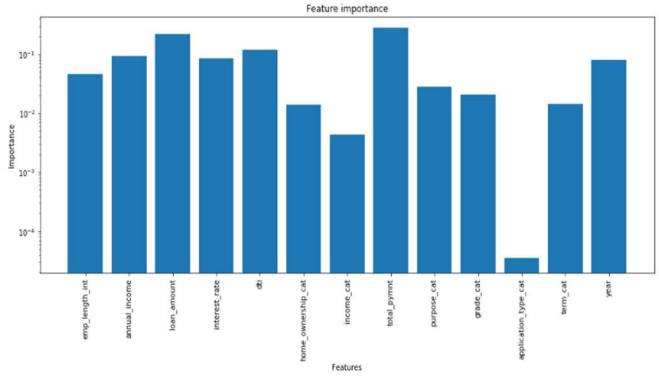

Figure 1. Important Feature Ordering Diagram

While using Pearson correlation there are only 2 attributes indicates a positive correlation with variable “loan condition category” there “interest rate” with correlation value 0.175, and “grade category” with correlation value 0.150.

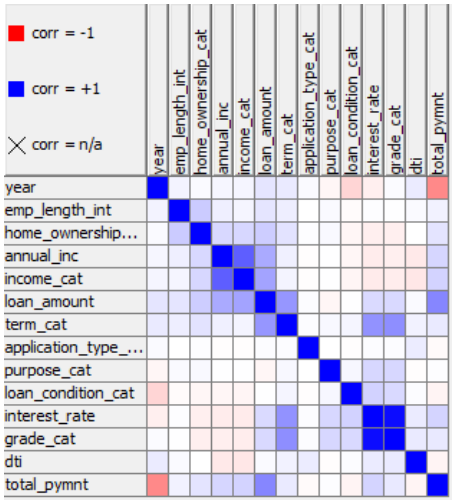

Figure 2. Pearson Correlation Coefficient Diagram

Therefore through the selection above, 7 attribute values were selected from 29 attributes in the final paper. The variable descriptions for the seven attributes are shown in Table 2.

Table 2. Description of Variable Properties

Variable	Variable Declaration	Variable Types
Employment length	The number of employment length in a year	Nominal variable
Annual income	The self-reported annual income provided by the borrower during registration	Continuous variable
Loan amount	The listed amount of the loan applied for by the borrower.	Continuous variable
Interest rate	Interest rate on the loan	Continuous variable
A debt to income (dti) ratio	The percentage of a consumer's monthly gross income that goes toward paying debts.	Continuous variable
Total payment	Payments received to date for total amount funded	Continuous variable
Grade	Borrowers' loan grade	Nominal variable

2.3. Data Preparation

Data preparation is the process of data manipulated and converted into forms that yield better results. Data preparation consists of two process data understanding and data preparation and data understanding itself consist of data collecting and data cleansing. Several steps in data preparation are:

a. Normalization outlier data: there is an outlier data in "dti" attribute, then we assigned an interquartile range (IQR) normalization.
b. Oversampling for imbalance data: The target (loan condition) is not equally distributed across the dataset. There is an imbalance between the two classes of 92,40% (good loan) and 7,60% (bad loan). Since one class is less numerous than the other, then the risk is going to be overlooked by the training algorithm. If the imbalance is not that strong, the data should be resampling before feeding the training algorithm. We decided to go for an oversampling of the minority class by using the SMOTE algorithm (Synthetic Minority Over-sampling Technique).

2.4. Classification

We randomly select 70% of the dataset as a training set, and 30% of the data as the test set. For the classification experiments, five classification algorithms are chosen, which are Gradient Boosted Trees, Random Forest, Naive Bayes, Decision Trees, and Logistic Regression and the proposed system we use is KNIME (Konstanz Information Miner).

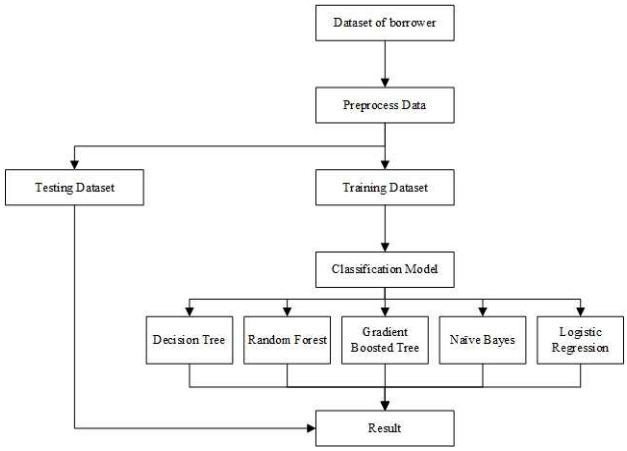

Figure 3. Proposed Framework

2.5. Gradient Boosted Trees

Gradient Boosted Trees is a method that builds the model in stages like other boosting methods but also generalizes these by optimizing an arbitrary differentiable loss function. Gradient Boosted Trees model can be used in both regression and classification problems. The final model is a function that takes as input a vector of attributes x ∈ Rn to get a score F(x) ∈ R so that Fi(x) = F(i-1) (x) + yihi (x), where each hi is a function that models a single tree and yi ∈ R is the weight associated with the i-th tree, so that these two terms are learned during the training phase.

2.6. Random Forest

Random Forest creates the forest with several decision trees. It makes the decision tree more robust and more precise, which has a supervised performance on classification and regression, and is widely used to deal with economic problems. Unlike linear models, Random Forest can deal with non-linear data as well. The formally models can be described as the following equation:

$$GG(x)=f0(x)+f1(x)+\cdots+fn(x)$$

Because of the using of multiple trees, compared to the decision tree, this algorithm reduces the probability of stumbling, which makes the prediction more credible. Besides, by creating multiple estimators, the influence of overfitting is reduced.

2.7. Naive Bayes

Naive Bayes classifier is one of the most famous classification algorithms for classification tasks. The core equation is based on the Bayes theorem.

$$P(ci|x) = p(x|ci)p(ci)/p(x)$$

Naive Bayes will calculate the probability of the specific category based on the prior information from data. Each category has its own posterior probability, and the largest one shows the possible candidate of its class. $p(ci)$ is easy to understand. It shows the ratio of each category. As for discrete information, likelihood probability $p(x|cc)$ could be calculated as the frequency of the appearance among data. For example, if feature A has appearance frequency n% for class $c1$, $p(A|\ cc1)$ = n%.

When it comes to continuous data, we need to use the normal distribution to model the data. For continuous data, it is hard to get such appearance frequency like discrete data. So, we need to calculate the mean and variance of one specific feature dimension, then use the pd.f of Gaussian distribution to get the probability of $p(x|c)$:

$$N(\mu,\sigma)=1/\sqrt{2\pi\sigma}\ \exp\ (-(x-\mu)22\sigma2)\ (11)$$

After understanding the possibility of each feature, just multiply them together, and we can know the posterior probability based on the multiplication law of probability. By comparing the posterior probability of each category, Naive Bayes can realize the classification task by selecting the class of the largest posterior probability.

2.8. Decision Trees

Decision Trees are decision making support tools that are used in data mining and artificial intelligence research and visualize the decision making process in the form of tree-shaped structures. The construction of a decision tree model requires a dataset of objects and a vector of attributes providing information about this dataset's object.

2.9. Logistic Regression

Logistic Regression is the preferred binary classification method, which is used for analysing a dataset that one or more independent variables may influence the outcome. The output is a discrete binary result between 0 and 1, and the purpose is to model the conditional probability p(Y=1|X=xx) and p(Y=0|X=xx). It based on the linear regression model, using the sigmoid function to compress the result of the linear model wTx to [0, 1].As we all know, linear regression can be described as the following equation:

$$w\cdot x=w0+w1x1+w2x2+\cdots wnxn$$

Formally, the logistic regression model is as follows:

$$p(Y=1|X=x)=(\exp(w\cdot x+b))/(1+\exp(w\cdot x+b))$$

$$p(Y=0|X=x)=1\ /\ (1+\exp(w\cdot x+b)\ )$$

And the sigmoid function is:

$$y=\ 1/(1+\exp\ (-x)\ )$$

From these above, the output of Logistic regression is a discrete binary result between 0 and 1, which indicates the probability that this sample belongs to a certain category.

3. Result and Discussion

The results were analyzed by comparing the performance of each algorithm using several measurements as shown in Table 3 where it appeared the highest accuracy compared to the other algorithms.

Table 3. Compare the Performance of The Algorithms

Algorithm	Accuracy	Error	Cohen's Kappa
Gradient Boosted Trees	80.30%	19.70%	0.234
Random Forest	**93.38%**	**6.62%**	**0.457**
Naive Bayes	59.60%	40.40%	0.086
Logistic Regression	69.58%	30.42%	0.097

The performance of gradient boosted trees, decision trees and random forest is equally good. Meanwhile other algorithms Naive Bayes, and Logistic Regression give a relatively low performance. One of the innovative points is that we develop a risk warning system based on predicting the risk of every borrower using logistic regression. It means that if a borrow is judged as negative by our method, then the probability of default is 69.58%

Besides that, to measure the performance of binary classification, we used Receiver Operating Characteristic (ROC) which showed each algorithm curve depends on (true positive rate) and (false positive rate) as shown in Figure 4. The ROC indicates the classification performance of a model. The performance of a classification model is better with larger ROC. The ROC can hence be seen as the ability to distinguish positive from negative classification. According to the result in Figure 4, Random Forest has higher ROC Value.

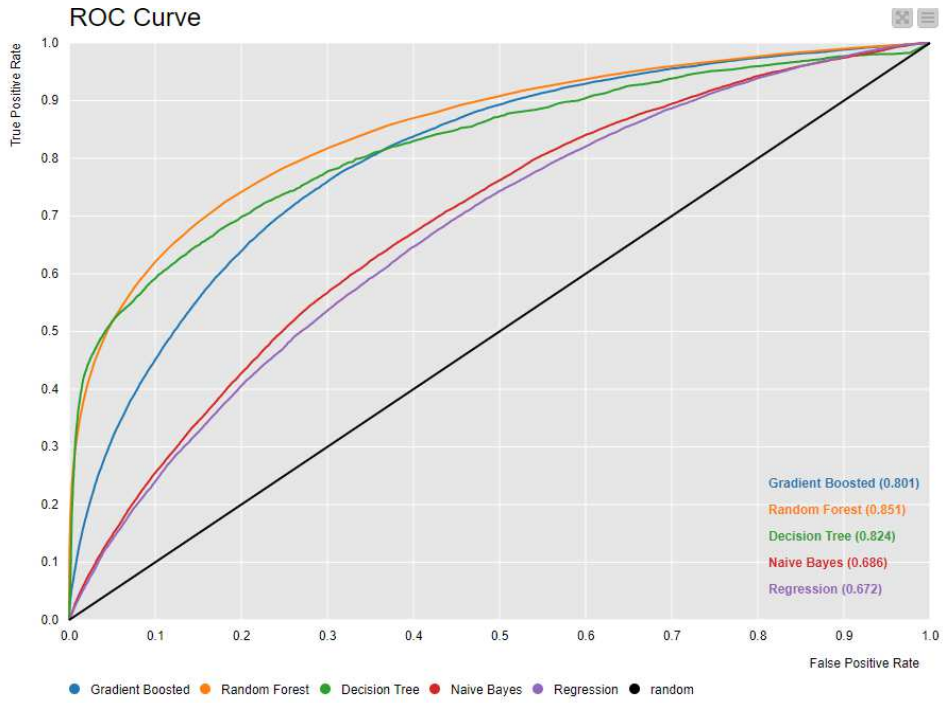

Figure 4. ROC Curve

4. Conclusion

The goal of this paper is to predict borrower loan conditions. For this study, our research dataset is collected from Lending Club, a P2P Lending based in the United States with 887,379 loan applications from 1 June 2007 until 31 December 2015 with 7 attribute values.

To conclude, based on five classification algorithms that we used, the Random Forest Model is capable of predicting borrower loan conditions because random forest can facilitate decision makers and predict the classification of borrowers with accuracy 93.38%.

References

[1] Pyle, D., *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, 1999.

[2] Provost, F. & Fawcett, T., *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*, O'Reilly (FV), 2013

[3] Ferozi, Muhammad Nadeem, *Loan Predictive Analysis*, www.kaggle.com, 2018

[4] Martinez Bachmann, Janio, *Lending Club Risk Analysis and Metrics*, www.kaggle.com, 2018

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. *SMOTE: Synthetic Minority Oversampling TEchnique. Journal of Artificial Intelligence Research*, 16:321–357, 2002

[6] Maria Tsami, Giannis Adamos, Eftihia Nathanail, Evelina Budilovich Budiloviča, Irina Yatskiv Jackiva and Vissarion Magginas, *A Decision Tree Approach for Achieving High Customer Satisfaction at Urban Interchanges*, p 194-202, 2018

[7] Xianyan Hou, *P2P Borrower Default Identification and Prediction Based on RFE-Multiple Classification Models*, 2020

[8] KNIME; *KNIME.com* AG, Germany; http://www.knime.org/

[9] Ryan Randy Suryono, and Indra Budi, *P2P Lending Sentiment Analysis in Indonesian Online News*, 2019

[10] Maan Y Alsalem, Safwan O Hasoon, *Predicting Bank Loan Risks Using Machine Learning Algorithm,* J.of comp & math's Vol 14 No.1, 2020

[11] Omer L. Gebizilioglu A. Belma Ozturkkal, *Predictive Modelling and Expectable Loss Analysis for Borrower Defaults of Mortgage Loans*, Journal of Modern Accounting and Auditing May 2018

[12] Zoran Ereiz, *Predicting Default Loan Using Machine learning*, 2019.