Received 8 April 2021
Revised 31 October 2021
27 May 2022
15 September 2023
14 April 2024
8 May 2024
Accepted 16 May 2024

A comparative analysis of consumer credit risk models in Peer-to-Peer Lending

Lua Thi Trinh
Department of Applied Mathematics,
Vietnam National University Ho Chi Minh City – University of Science,
Ho Chi Minh City, Vietnam

Abstract

Purpose – The purpose of this paper is to compare nine different models to evaluate consumer credit risk, which are the following: Logistic Regression (LR), Naive Bayes (NB), Linear Discriminant Analysis (LDA), k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), Classification and Regression Tree (CART), Artificial Neural Network (ANN), Random Forest (RF) and Gradient Boosting Decision Tree (GBDT) in Peer-to-Peer (P2P) Lending.

Design/methodology/approach – The author uses data from P2P Lending Club (LC) to assess the efficiency of a variety of classification models across different economic scenarios and to compare the ranking results of credit risk models in P2P lending through three families of evaluation metrics.

Findings – The results from this research indicate that the risk classification models in the 2013–2019 economic period show greater measurement efficiency than for the difficult 2007–2012 period. Besides, the results of ranking models for predicting default risk show that GBDT is the best model for most of the metrics or metric families included in the study. The findings of this study also support the results of Tsai *et al.* (2014) and Teplý and Polena (2019) that LR, ANN and LDA models classify loan applications quite stably and accurately, while CART, k-NN and NB show the worst performance when predicting borrower default risk on P2P loan data.

Originality/value – The main contributions of the research to the empirical literature review include: comparing nine prediction models of consumer loan application risk through statistical and machine learning algorithms evaluated by the performance measures according to three separate families of metrics (threshold, ranking and probabilistic metrics) that are consistent with the existing data characteristics of the LC lending platform through two periods of reviewing the current economic situation and platform development.

Keywords P2P lending, Lending club, Default risk, Credit risk models, GBDT

Paper type Research paper

1. Introduction

P2P lending is designed and built on a digital application platform to connect directly between borrowers and lenders (investors) without the need for financial intermediaries. The rapid global development of the P2P lending model over the past 15 years has created a new capital supply channel and has contributed to promoting comprehensive financial development. Similar to the traditional lending sector, credit risk prediction is one of the top concerns when assessing whether a borrower can pay off a loan in P2P lending, where the subject of asymmetric information is common. Today, these predictions are associated with prediction models, typically credit scoring and/or credit classification (good versus bad or no default risk versus default risk) models. Currently, there are many approaches to building

Journal of Economics, Finance and Administrative Science
Vol. 29 No. 58, 2024
pp. 346-365
Emerald Publishing Limited
2077-1886
DOI 10.1108/JEFAS-04-2021-0026

predictive models in traditional lending. Although the online P2P lending model is a relatively new field of research, the number of scientific contributions to this research is increasing and has been published in recent years (Reddy, 2016), thanks to the potential development of this platform and especially to the empirical contribution to solving many of the risky problems, including credit risk, associated with machine learning algorithms. An overview of the empirical studies on credit risk models on this loan platform shows the following main research flows: (i) important factors for forecasting (Lin *et al.*, 2016); (ii) proposing models or incorporating new factors to improve loan classification efficiency and increase predictability (Namvar *et al.*, 2018) and (iii) comparative performance of models (Tsai *et al.*, 2014). In particular, the performance of prediction models needs to be researched and compared clearly, providing a measure of choice to evaluate or classify a loan following the criteria of the other P2P lending platforms together. Furthermore, the literature review shows that studies on this lending platform often either use lengthy data without considering the impact of material real-life events (such as economic crises) and the developmental stages of the lending platform or they only consider data from the period of economic recovery and development, and the lending platform goes into a stable period to avoid bias values (Teplý and Polena, 2019). Meanwhile, Giannopoulos (2018) argues that the efficiency of all models decreased significantly during the recession period, suggesting that lending efficiency not only depends on the quality of the borrower but also on the economic scenario. Therefore, the main contributions of this paper to the empirical literature review include: comparing nine prediction models of consumer credit risk through statistical and machine learning algorithms evaluated by three separate performance measure families of metrics (threshold, ranking and probabilistic metrics), which is consistent with the existing data characteristics of LC lending platform through two periods of reviewing the current economic situation and the platform development.

2. Literature review

2.1 Credit risk models

Three groups of models are mentioned in Basel II for financial intermediaries to assess credit risks, including (i) Probability of default (PD); (ii) a credit position model, such as Loss Given Default (LGD) and (iii) expected loss and unexpected loss estimation models of portfolios (Odeh *et al.*, 2006). However, empirical studies focusing on PD model are well-established and account for a large proportion of the literature review. Common PD-based approaches include [1]: (i) application of credit scoring to a new candidate to rank in grade or subgrade with an appropriate loan interest rate or collection plan or even rejecting a loan if the credit score is not guaranteed according to the borrower's credit policy; (ii) classifying a new borrower as good or bad (with default risk [2]) when comparing the estimated probability of default risk to an appropriate threshold using some classification techniques that have been trained on data and customer behavior similar to the past or the present to find relationships and rules to screen and classify potential loan applications in order to decide whether to accept or reject a loan.

According to Hand and Henley (1997), the company can make significant savings in the future if one percent improvement is possible with the accuracy of the credit scoring technique. Consequently, a variety of classification techniques for predicting the probability of default have been introduced and used to develop accurate credit scoring models. The following common models for classifying methods used in loan classification can be mentioned: (i) parametric and nonparametric models; (ii) linear, nonlinear and rule-based algorithms; (iii) individual classification, homogeneous ensemble classification and heterogeneous ensemble classification and (iv) statistical technique and machine learning.

2.2 P2P lending review

As defined by the P2P lending association in the UK, P2P lending is a platform that can support financial services through a direct, one-to-one contract between a receiver and one or more investors (Reddy, 2016). As a result, the credit decisions on P2P platforms, which are quick and low-cost, also allow borrowers with short credit histories and small businesses to directly access financing loan channels. At the same time, yields for lenders and interest rates could be better than traditional credit channels. The first officially recognized online lending model was Zopa, a UK company, in March 2005, followed by the USA, Italy, Japan, China, etc. Nowadays, the number of companies with this form or a variation of this form is widespread, not only in developed countries but also in developing ones. In P2P lending, a serious problem can lead to future risks such as moral hazard, choice disadvantage and asymmetric information, which directly affect the lender through borrower fraud (Lin *et al.*, 2016). Generally, lending operations are carried out on an online platform, meaning that the parties involved have no physical contact, which results in the fact that some borrower information has not been verified correctly. Moreover, in some countries, this system grows too fast, typically in China. While lacking legal scrutiny, in some cases, to maintain the desired rates of return, these P2P intermediaries may not comply with borrower protection laws, which lead to industry chaos (Tao and Chang, 2019).

LC is the largest P2P lending platform in the USA, with lending information dating back to 2007. It was also the first platform to launch an IPO on the New York Stock Exchange, in December 2014. By the end of 2019, LC had more than 2.8m consumer loans, for a total loan amount of about $44bn. In the USA, loans on a P2P platform are considered a legitimate asset, and loans are of the same nature as unsecured loans issued by traditional banks. Consumer loan information will be continuously updated by the platform but it will still ensure the borrowers' right to private information.

P2P LC plays an intermediary role in collecting, filtering and classifying borrowers with corresponding interest rates (Figure 1) and then calculating a certain percentage of the loan for fees accompanying the service (Tao and Chang, 2019). P2P LC is not responsible for the borrower's default risks. However, they conduct their risk assessments using consumer reporting and the borrower's past performance.

2.3 Comparative analysis of credit risk models

The qualitative analysis discusses market emergence, overviews, investment strategies (Namvar, 2013) and factors influencing the P2P platform. Quantitative analysis, on the other hand, accounts for a larger share of literature review in this market, including major research

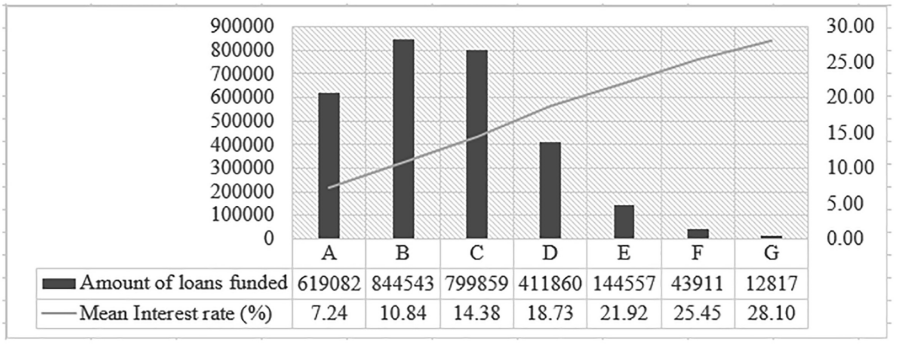


Figure 1.
Interest rate corresponds to grade according to loans funded at LC (2007–2019)

Source(s): Authors' own calculations based on LC data

flows: (i) determinants of funding success and default and (ii) credit risk analysis models for credit classification or scoring.

The approaches for comparing the existing risk classifications include: (i) comparing the models using multiple datasets (Zhang *et al.*, 2007); (ii) a comparison based on performance measures (Ferri *et al.*, 2009); (iii) comparing parametric and nonparametric models and individual models with ensemble models. However, it is not the purpose of all of the studies to compare the classifications as "comparative"; sometimes the main purpose is to highlight methods or techniques proposed in the study itself or to assess the role of other proposals, such as the addition of features (social media information [3], soft information, data transformation, etc.). In some cases, this can raise issues or controversy, such as a preference for the proposed models in optimizing parameters or the selection of appropriate performance indicators in line with such a proposal. Therefore, this paper attempts to review studies that neutralize methods of predicting default risk. Comparative studies based on LC data also contribute quite well to the existing literature review of empirical studies on evaluating and comparing the performance of models to predict default risk. Tsai *et al.* (2014) compare four algorithms: NB, RF, SVM and Modified LR (MLR) in the 2007–2013 period to predict the probability of default risk and compare the calculated return rate by MLR with the LC return rate. The study indicated that MLR outperformed the rest of the methods and classified grade A1 loans better than LC. Chang *et al.* (2015) compared LR, NB and SVM models after parameter adjustment to predict borrower status. The main results show that NB and Gaussian perform best with default predictions and can push LC return on investment up to 50%. After comparing different loan classification models (Classification Tree (CT), LR, Generalized Regression Models (GRM), Extreme Gradient Boosting (Xgboost) and RF), Reddy (2016) shows that the best performing model in both training data and testing data is Xgboost. Bae *et al.* (2018) deploy LR, Decision Tree (DT) and Multilayer Perceptron (MLP) models to predict borrower status. The results prove that MLP is superior to LR and DT in predicting the risk of default P2P. Teplý and Polena (2019) rank ten default risk prediction techniques by the average ranking of each method by six performance metrics. The classification techniques are: ANN, LR, LDA, Radial Basis Function (RBF) kernel, SVM (SVM-Rbf), Linear SVM (L-SVM), Bayesian Network, NB, k-NN, CART and RF. The performance measures include Accuracy Ratio (AR), Kolmogorov–Smirnov Statistic (KS), Brier Score (BS), AUC, Partial Gini Index (PG) and H-measure. The ranking results of the study indicate that LR, ANN and LDA are the three best credit classification algorithms based on LC data, while k-NN and CART are the two worst classification methods.

The most recent economic crisis began in mid-2007 in the USA with the onset of a subprime home mortgage crisis with nonperforming loans and bankruptcy. Around mid-September 2008, banks faced a liquidity crisis, losing billions of dollars due to the bankruptcy of customers and massive withdrawals by depositors. Meanwhile, financial intermediaries were not able to refuse all of their customers to avoid risks. Hence, classifying borrowers becomes more and more important and urgent in the lending process. However, some studies of traditional loan risk assessments express a fear that prediction models may not be available to predict borrowers' performance in difficult economic circumstances, especially if the historical period observed does not include deterioration conditions (Madzova and Ramadini, 2013). According to Dinh *et al.* (2013), not considering the general economic scenarios continuously can reduce the performance of classification models; whether or not the performance of the loan classification or credit scoring models is affected when considering the economic contexts of P2P lending platform data.

3. Method

3.1 Research design

Research data are divided into two phases: a difficult economic period (2007–2012) and a stage of economic recovery and development (2013–2019) [4]. The implementation process is briefly summarized in Figure 2. The data are cleaned and screened for suitable variables, and some data are transformed if necessary. These data are then split into training data and testing data. However, according to Namvar *et al.* (2018), imbalanced data sets are a common problem in credit risk assessment, which results in misclassification and causes the accuracy metrics of the models to become unreliable. Therefore, training data are balanced before being divided into subsets to optimize parameters, train the model and predict the borrower's credit risk on testing data.

3.2 Data and variables

3.2.1 Data collection. Consumer loan customer data is downloaded directly from P2P LC platform website, provided that the loan or investment account is registered. Data collected from 2007 to 2019 included 2,876,629 loans and 150 features, where the number of loans corresponding to two periods, 2007–2012 and 2013–2019, is 95,902 and 2,780,727 loans funded, respectively (Table 1). Compared with previous studies on P2P LC data, the whole data in this paper has the longest duration and the largest number of observations.

3.2.2 Variables selection. The principles of study variable rejection from 150 characteristics in LC data include: (i) more than 30% of attributes missed [5]; (ii) attributes (numeric or categorical) have only one value or do not contain the required information or categorical attributes have a categorical number that is too high (Chang *et al.*, 2015); (iii) attributes with similar or identical information (leading to multicollinearity); (iv) attributes are excessively informative, i.e. they contain information about the borrower after the loan

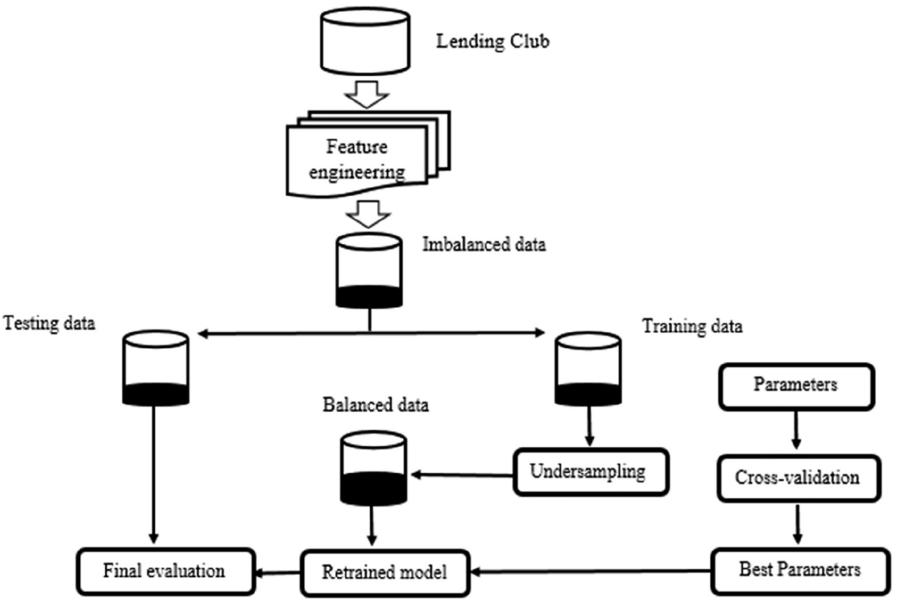


Figure 2.
Research implementation process

Source(s): Authors' own illustration

2007–2012		2013–2019	
Loan status	# Of loans	Loan status	# Of loans
Fully paid	78,839	*Fully paid*	1,495,504
Charged off	14,314	Current	890,873
Does not meet credit policy status: *fully paid*	1988	*Charged off*	366,331
		Late (31–120 days)	15,630
Does not meet credit policy status: *charged off*	761	In grace period	9,392
		Late (16–30 days)	2,571
		Default	426

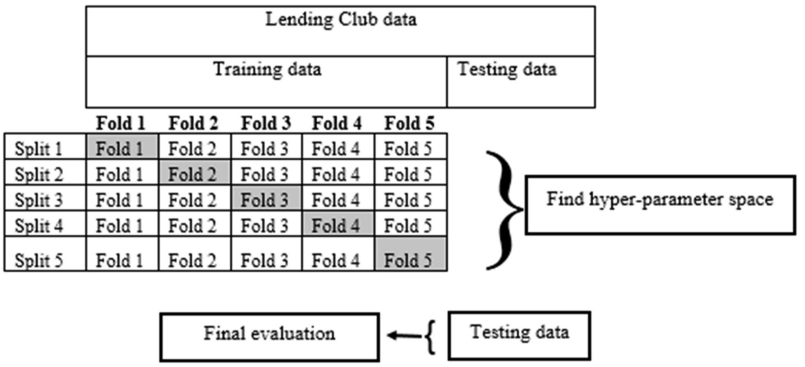
Source(s): Author's own calculations based on LC data

Table 1. Number of loans funded by loan status

has been accepted (except *loan status*). Namvar *et al.* (2018) argue that training models with such variables will produce suspiciously accurate prediction results, thus criticizing studies using such "leakage" variables and (v) default risk assessment attributes such as *grade, sub_grade* and *int_rate* because these variables have the power of risk perception from LC itself.

3.2.3 Data transformation. A categorical variable such as *addr_state*, representing the living state of the applicant, is often excluded from previous studies by a large number of categories (50 and 51 states for 2 study periods, respectively). This study converts *state* information to *region* where the borrowers live to reduce classification information [6]. Based on the majority of empirical studies performed on LC data, this paper creates a new variable named *fico_score*, which is calculated by taking the average of *fico_range_low* and *fico_range_high*. *emp_length, number_cr_line* and *term* variables are converted into numerical or timing values. The target variable of the study is *loan_status*, which describes the current state of a loan at the time of the data download. This study labels: (i) *0* (good loan application) for a loan application with the status *"Fully paid"* and *"Does not meet credit policy. Status: Fully paid"* and (ii) *1* (bad loan application) for a loan application with the status *"Charged off"*, *"Does not meet credit policy. Status: Charged off"* and *"Default"*.

3.2.4 Splitting the dataset. The research divides the data (both periods) into two parts, i.e. training data (70%) and testing data (30%). Figure 3 describes the K-fold cross-validation (cv) procedure in the training data section with K = 5, which means that the training data are



Note(s): See more: https://scikit-learn.org/stable/modules/cross_validation.html #cross-validation

Source(s): Modified figure by author according to scikit-learn (cross-validation)

Figure 3. Describe the K-fold cross-validation process

divided into five random subsets [7], with four subsets for training and one subset for validation, respectively. The hyperparameter space is selected based on the best cross-validation score.

3.2.5 Imbalanced data problem. According to Fernández *et al.* (2018), most machine learning algorithms for classification models are designed with the assumption of an even distribution of classes. Meanwhile, an imbalanced dataset is typical for credit datasets in general and for P2P in particular (Namvar *et al.*, 2018). Brown and Mues (2012) and Chang *et al.* (2015) show in their study that classification models may fail to predict borrower default risk because default risk observations, which are primarily interesting belong to a minority class, and while the accuracy of the models may be high, they could not predict or would predict only very few default risk records. This study uses a random undersampling sampling method to balance data for the following reasons: (i) the number of observations in LC data of both periods is quite large. Hence, an obvious undersampling technique should be considered and (ii) the research results of Namvar *et al.* (2018) on LC data show a clear effect of the random undersampling strategy when combined with classification models.

3.3 Classification models

The study overviews each approach, briefly describing the idea of some important computational algorithms or functions implemented in each approach, which readers can see in more detail in the studies of Hastie *et al.* (2009) and Shmueli *et al.* (2018).

The goal of the LR function for the binary target variable is to determine the probability of determining a class y for a data point x:

$$\pi_i = P(y_i = 1|x_i) = g(h(x_i)) \tag{3.1}$$

Where x_i is a characteristic variable vector and $h(x_i) = \sum_{j=0}^{k} \beta_j x_i^j$ is a linear prediction. With LR, model log-odd as $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$ is a linear function of the explanatory variable, where function *g(t)* is a logistic function:

$$g(t) = \frac{\exp(t)}{1 + \exp(t)} \tag{3.2}$$

NB is a probability classifier based on Bayesian theory and assumes that the input values of feature vector x are independent of each other.

Bayes rule is defined as follows:

$$p(y = c_k|x) = \frac{p(x|y = c_k).p(y = C_k)}{p(x)} \tag{3.3}$$

Thanks to its simplicity, NB model is quite fast in terms of training and testing. Under the above assumption, the probability that the data point falls in class y is:

$$p(x|y) = p(x_1, x_2, , \ldots, x_d|c) = \prod_{i=1}^{d} p(x_i|y) \tag{3.4}$$

where $p(x_i|y)$ is the conditional probability of the input variable.

LDA is a simple parametric statistical model, specially developed and used to distinguish two or more groups. LDA model derives from a simple probability model, modeling the class conditional distribution of data $p(x|y = c_k)$ for each class k and applying Bayes rule in (3.3) to

obtain predictive estimates for each training sample $x \in R^d$ through the posterior probability. This probability is calculated as follows: $\mathrm{p}(c_k|\mathrm{x}) = \mathrm{p}\ (\mathrm{X} \in c_k|\ \mathrm{X} = \mathrm{x}) = \frac{f_k(x)\pi_k}{f_1(x)\pi_1 + f_2(x)\pi_2}$ (3.4)

Where $f_k(\mathrm{x})$ is the conditional multivariate probability density of x for class k and π_k is the probability of class k. Assuming that the classes share the same covariance matrix, that is, $\sum_k = \sum \forall k$, the calculation function of LDA is:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log\, \pi_k \tag{3.5}$$

where the function's parameters are estimated as follows: $\widehat{\pi}_k = N_k/N$, N_k the record of class k; $\widehat{\mu_k} = \sum_k x_i/N_k$; $\widehat{\Sigma} = \sum_{k=1}^{K} \sum_k (x_i - \widehat{\mu_k})(x_i - \widehat{\mu_k})^T/(N-K)$.

The rule for building k-NN is to find the number of k predefined training samples at the closest distance to the new point and predict the label from there (Yeh and Lien, 2009). According to Shmueli *et al.* (2018), Euclidean distance is considered to be the most popular choice for measuring the distance between two observations because k-NN algorithm is based on many distance calculations. Euclidean distance between two observations $x = (x_1, x_2, \ldots x_d)$ and $u = (u_1, u_2, \ldots u_d)$ is calculated as follows:

$$d(x,u) = \sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \ldots + (x_d - u_d)^2} \tag{3.6}$$

The idea of SVM is to find a transformation so that the original data (non-discrimination linear) maps to a high-dimensional space where the data are linearly distinguishable (Tao and Chang, 2019). The mapping used in SVM algorithm requires that the scalar product of data vectors in the new space can be easily computed from coordinates in the old space. This scalar product is determined by kernel function *k(x,x′)* and is defined as follows:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \tag{3.7}$$

where Φ: $X \to H$ is the projection from a feature space to a high-dimensional feature space.

DT belongs to a rule-based classification and has a hierarchical organization structure, shaped like a tree, with each node dividing the data space into sections based on the values of attributes. Several algorithms have been developed to build DT, such as ID3, C4.5, C5.0 and CART. CART model is widely and successfully used for classification or credit-scoring loan applications (Giannopoulos, 2018; Ince and Aktan, 2009). Using the feature and the feature's threshold values, CART constructs a binary tree to obtain maximum information at each node. The splitting of each node is selected with Gini coefficient calculated at each node as follows:

$$\mathrm{Gini}(\sigma) = 1 - \sum_j \left(\frac{N(\sigma, \mathrm{j})}{N(\sigma)} \right)^2 \tag{3.8}$$

MLP model, a case of single-layer ANN, is arguably the most popular and widely used model in credit risk measurement studies (Teplý and Polena, 2019; Wendler and Grottrup, 2016). MLP function has one hidden layer; one neuron has the following form:

$$f(x) = W_2 g\left(W_1^T x + b_1\right) + b_2 \tag{3.9}$$

where W_1, W_2 is the weight of the input layer, the hidden layer relatively; b_1, b_2 is the bias added to the hidden layer and the output layer relatively; W_2, b_1, b_2 are model parameters;

activation function $g(.) : R \rightarrow$ R is defaulted by the hyperbolic tan, and is calculated as follows:

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{3.10}$$

RF is also a tree-based classification, using an algorithm that combines the predicting results of DT. The estimate function of RF classification takes the following form:

$$\widehat{C}_{rf}^{B}(x) = majority\, vote \left\{ \widehat{C}_b(x) \right\}_1^B \tag{3.11}$$

where B is the number of random forest trees and $\widehat{C}_b(x)$ is the class prediction of b.

Similar to RF, GBDT is also a tree-based classification, but the trees used for the ensembles are constructed sequentially to reduce association bias, with the idea of combining multiple weak learners (trees) to build strong models. The predictive function of GBDT classification takes the following form: $\widehat{y}_i = \sum_{k=1}^{t} w_k f_k(x_i)$ (3.12), where f_k is a function of kth DT, w_k is the weight corresponding to false classification, that is, the more a sample is misclassified, the more important a data sample with corresponding weight becomes. The basic GBDT process consists of three steps: (i) optimize loss function $l(\widehat{y}_i, y_i)$; (ii) focus on weak learning methods (trees) to predict observations (filtering out the ones they can process), thereby developing new methods to process new observations and repeating this process several times and (iii) use gradient descent to add trees to the model to reduce loss.

3.4 Tuning the model hyperparameter

Each model has a set of parameters trained directly with the training data set to represent each relationship between the features and target variables in the data. However, these parameters are advised to apply fine-tuning methods to achieve optimization, thereby improving the predictive power of the models (Tsai *et al.*, 2014). The hyperparameter space is set up with the best cross-validation score using the metric Area under the Receiver Operating Characteristic Curve (AUC) on the training set through the grid search method [8] (Niu *et al.*, 2019).

3.5 Performance measurements

This study used some of the performance measures divided into three groups, according to the study by Ferri *et al.* (2009): (i) family of metrics based on a threshold or qualitative understanding of error (threshold metrics) used to minimize the number of errors, including accuracy ratio (AR) and F-measure; (ii) family of metrics based on understanding the probability of errors (probabilistic metrics), i.e. measuring deviation from true probabilities, to assess the reliability of classifications, which are not only misclassified but also accompanied by the probability of choosing wrong class, including Log loss and Brier score; (iii) a family of metrics is based on the degree of goodness the model classifies observations with (ranking metrics), in order to evaluate the effectiveness of the classification, including AUC. However, according to Brownlee (2020) and Fernández *et al.* (2018), the performance metrics for the classification of problems on imbalanced data should be carefully considered because most of the measures evaluate the performance of models widely used that equally assume a balanced distribution between classes, leading to serious issues when assessing imbalanced problems. Hence, some of the measures just introduced can be adjusted to accommodate the imbalanced data problem. AR is not suitable for imbalanced data because it

only predicts the majority class while ignoring the minority class. Therefore, another metric proposed to replace AR is Balanced Accuracy Ratio (BAR) and F-measure is chosen to be F_2 with $\beta = 2$ (He and Ma, 2013).

4. Results

After the data are cleaned, the study selects and converts some variables so that the remaining variables for each period, 2007–2012 and 2013–2019, respectively, are 21 and 36 variables (Table A1, Appendix). The rest of the data are processed as follows: (i) delete the outliers' observations based on the experience of the variable and the data distribution and (ii) missed numerical data are replaced with median values. There is usually no missed information in the remaining categorical data or very little information, and if any, these records are deleted. The remaining number of observations for each period is 95,682 with a default rate of 15.71% (2007–2012) and 1,469,063 with a default rate of 19.66% (2013–2019). After a preliminary assessment of the variables used, the study computes a correlation coefficient matrix to check multicollinearity as well as the potential correlation relationships between the target and the independent variable, where the highest absolute correlation coefficient of the two periods, 2007–2012 and 2013–2019, is 0.57 (*revol_util* vs *fico_socre*) and 0.74 (*revol_until* vs *percenr_bc_gt_75*), respectively. However, according to Teplý and Polena (2019), sometimes excluding these variables will cause the loss of important information about the borrower and the loan, because these variables are all in a group of variables with a top correlation-to-target variable. The two independent variables that have a leading correlation with the target variable in both phases are *term* and *fico_score*.

Finally, variables with categorical data are converted to dummy variables to match most of the classification models. In addition, the input data includes variables with different units and range values. Moreover, some models assume that the input data have a Gaussian distribution with a mean value of zero and a variance of 1. Therefore, data normalization is necessary, and the normalization formula used is as follows: $x' = \frac{x - \overline{x}}{\sigma}$ where, $\overline{x}$ and σ are expectations and the square root of the variance of attribute X.

The operations in this paper, from data processing, descriptive statistics, modeling and hyperparameters optimization space (Table A2) and evaluation on testing data, are programmed through Python version 3.6 combined with the scikit-learn library to solve machine learning algorithms.

4.1 Compare credit risk models over two periods (Table 2)

Considering threshold metrics, BAR increased significantly in all models when comparing the 2013–2019 period with the 2007–2012 period, especially in the case of SVM model using RBF kernel function. This model has a marked increase in performance in both threshold measures in particular and the rest of the measures in general. In addition, considering the efficiency evaluation metrics separating two target classes through AUC scores, the results of all models in the 2013–2019 period performed better. The metrics assessing probabilistic predictions, such as BS and LL, are aimed at measuring the difference between the model prediction and reality, so the smaller these two metrics are, the better the results of the classification model are. The results show no clear difference between the two periods because, in some models, the results show that there are better (smaller) differences in the 2013–2019 period, while the other models show the opposite. The visual results show that the classification or credit scoring models in the 2013–2019 period are generally better than those for the 2007–2012 period in terms of measuring accuracy.

Table 2.
Comparing the classification models in two periods

Classifications	Threshold metrics						Ranking metric			Probabilistic metrics					
	BAR (07–12)	BAR (13–19)	Better	F2 (07–12)	F2 (13–19)	Better	AUC (07–12)	AUC (13–19)	Better	LL (07–12)	LL (13–19)	Better	BS (07–12)	BS (13–19)	Better
LR	0.63728	0.63894	13–19	0.66667	0.66237	07–12	0.68720	0.69323	13–19	12.29847	12.38141	07–12	0.223589	0.221877	13–19
NB	0.57301	0.57950	13–19	0.74581	0.46127	07–12	0.63574	0.64289	13–19	9.005112	19.98789	07–12	0.209889	0.411747	07–12
LDA	0.63682	0.63890	13–19	0.67532	0.66237	07–12	0.68604	0.69306	13–19	12.19018	12.35541	07–12	0.223446	0.222006	13–19
k-NN	0.59415	0.59975	13–19	0.63780	0.63221	07–12	0.63419	0.64211	13–19	13.44517	13.53651	07–12	0.233509	0.233088	13–19
SVM	0.53985	0.63936	13–19	0.39286	0.65377	13–19	0.57115	0.69497	13–19	22.32887	12.69005	13–19	0.2626	0.224374	13–19
CART	0.61209	0.61222	13–19	0.68246	0.62500	07–12	0.65242	0.66102	13–19	11.76782	13.89906	07–12	0.231949	0.232425	07–12
ANN	0.63685	0.64234	13–19	0.65796	0.65377	07–12	0.68472	0.69771	13–19	12.70277	12.83865	07–12	0.222975	0.224094	07–12
RF	0.62742	0.63300	13–19	0.64921	0.65232	13–19	0.67709	0.68622	13–19	13.23581	12.93221	13–19	0.227142	0.224676	13–19
GBDT	0.64026	0.64616	13–19	0.65796	0.66237	13–19	0.69257	0.70261	13–19	12.82671	12.48644	13–19	0.223542	0.220778	13–19

Note(s): Where 07–12: 2007–2012 and 13–19: 2013–2019
Source(s): Author's own calculations based on LC data

4.2 Compare ranking results of classification models with other studies on P2P data (Table 3)

The two-period comparative results show that the performance of credit risk models in the 2013–2019 period is better than that in the 2007–2012 period. Therefore, the study only uses the comparative results of the 2013–2019 period to compare with previous studies that had a comparison of models on P2P data. For the convenience of comparison, this comparative result table only includes the models used in this paper [9]. First, this research evaluates the ranking results of credit risk models used in the study across three families of metrics. Instead of ranking based on the average score of each classification across all performance metrics, such as Teplý and Polena (2019), this study only ranks classification models based on the average score for each classification according to three families of measures in a relative manner.

In general, GBDT model outperforms other credit risk models in predicting the default risk of borrowers because its prediction results surpass those of most measures or families of measures. Although GBDT model has not been as prevalent in credit scoring or risk classification studies as other models, such as LR, LDA and RF, the ensemble strategy of GBDT or Xgboost models have been the main strategy for recent data mining contest participants who aim to optimize the predictive measures of credit risk models (Shmueli *et al.*, 2018). LR, LDA and ANN models also had quite good ranking results in all three families of metrics; especially LR proves why it is popular in empirical studies on credit scoring or classification loan applications. Not only because of its intelligibility and its ability to generalize relationships between features but also because LR achieves high classification accuracy as well as lower predictive error when compared to complex models that require hyperparameter optimization. The performance of CART, k-NN and NB models measured in all three families of metrics is lower than that of other models, proving that these models are not suitable for classifying or credit scoring loan applications on P2P LC platform (only considering the performance model factor as the main criterion for choosing a model).

Although there are differences in the number, type of classification model, data preprocessing (e.g. balanced data) as well as performance metrics, this study has some similar results (Table 4) in comparison with Tsai *et al.* (2014) and Teplý and Polena (2019) when considering the relative ranking positions of some models, i.e. LR, ANN and LDA models are the best loan classification models [10] (in this study, they were just below GBDT). Meanwhile, CART and k-NN are ranked as the worst credit risk models based on LC data.

	Threshold metrics			Ranking metric	Probabilistic metrics		
Classifications	Ranking (BAR)	Ranking (F2)	avg	Ranking (AUC)	Ranking (LL)	Ranking (BS)	avg
LR	4	2	*2*	*4*	2	2	*1-2-3*
NB	9	9	*9*	*8*	9	9	*9*
LDA	5	3	*5*	*5*	1	3	*1-2-3*
k-NN	8	7	*7–8*	*9*	7	8	*7–8*
SVM	3	4–5	*4*	*3*	4	5	*4–5*
CART	7	8	*7–8*	*7*	8	7	*7–8*
ANN	2	4–5	*3*	*2*	5	4	*4–5*
RF	6	6	*6*	*6*	6	6	*6*
GBDT	1	1	*1*	*1*	3	1	*1-2-3*

Note(s): Where Ranking (avg.) is ranking based on the average score of the ranking results of metrics within the same family

Source(s): Author's own calculations based on Table 2

Table 3. Results of ranking classifications on three families of metrics 2013–2019

Table 4.
The results of a comparison of classification models on P2P data

Work	Period	Data		Classifiers													Metrics
		# Of obs	# Of vars	RF	ANN	MLP	LR	LDA	SVM	SVM-Rbf	SVM-L	SVM-P	NB	k-NN	CART	GBDT/Xgboost	
Tsai *et al.* (2014)	2007–2013	91,520	n/a	3			1			2			4				PPV
Chang *et al.* (2015)	2007–2015	n/a	n/a				3			5	2	4	1				AR, G-mean
Jin and Zhu (2015)	2007–2011	34,406	11		4	2			1						3		AR, LiCha
Reddy (2016)	2007–2016	1,254,308	40	2			3									1	AR
Bae *et al.* (2018)	2016–2017	153,250	14			1	3								2		AR
Teplý and Polena (2019)	2009–2013	212,252	23	5	2		1	3		6	4		7	9	8		AR, BS, AUC, H, KS, PG

Note(s): Where LiCha: Lift chart, CM: Confusion matrix, RMSE: root-mean-square, RU: Relative usefulness; SVM-L: SVM-Linear, SVM-P: SVM-Polynomial. Column ANN: other versions ANN except for MLP (ex: BP and RBF). Column SVM: other versions SVM except for SVM-Rbf, SVM-L and SVM-P
Source(s): Author's own synthesis

5. Discussion

5.1 Theoretical implications

LR model is widely used because of its ease of execution, relatively high predictive power, low implementation costs as well as its ability to explain the roles of the input variables. However, this model also suffers from some limitations, such as assumptions about its robustness and the fact that it cannot solve nonlinear problems. Due to its simplicity, NB model is quite fast in training and testing, but its assumption is also its main weakness, as in reality, dependencies can exist between the variables (Yeh and Lien, 2009). Although widely used, LDA model is often criticized for its assumption that for each type of target variable, independent variables are distributed as a multivariate normal distribution with a common covariance matrix (Xiao *et al.*, 2006). k-NN has several advantages in becoming the preferred class for classification loan applications (Yeh and Lien, 2009). However, k-NN does not create a simple classification probability formula and its predictive accuracy is greatly affected by the distance measure and the number of k of the nearest neighborhood. Due to the advantages of high-dimensional small sample processing and the ability to model nonlinear and high-dimensional space descriptions, SVM is favored in the study of predicting the risk of default. Due to the lack of assumptions about normal distribution and the ability to visualize, CART is widely and successfully used for classification or credit scoring loan applications (Ince and Aktan, 2009). According to Xiao *et al.* (2006), ANN may become an alternative to traditional regression models such as LR (which is criticized for its assumptions) for several reasons: (i) suitable for the case of the target and independent variables with non-linear relationship; (ii) there is no rigid restriction in the use of input and output functions; (iii) repeated, incorrect or missed data values are easily met and (iv) most importantly, when using ANN, it is possible to skip the analysis of the detailed architecture of an issue or a structure of use. However, it has the same disadvantages as the inexplicability of the learning process because the decision-making process is still in the "black box", resulting in a lack of generalization and expression of the importance of potential variables. Although RF can work efficiently on a large database, overcome overfitting problems, handle missing values and be robust to outliers, it involves a prohibitive computational cost to process big data. GBDT is highly appreciated because of its combination of multiple models of these methods to create more accurate predictions by reducing predictive error variance.

In general, the findings of this research show that the performance of the classification models for the 2013–2019 period is generally better than that for the 2007–2012 period in terms of measuring accuracy. This may be attributed to the fact that the quality of the input data set for the 2007–2012 period was not good in comparison with the 2013–2019 period. The data includes a lot of missed values, especially loan applications that do not meet the credit policy, which leads to a decrease in the number of important features or observations. In addition, the quality of the credit information could decrease significantly because of the information lag during the crisis (Madzova and Ramadini, 2013) and the market penetration stage. This result is completely consistent with the results and opinions of some previous studies, such as Dinh *et al.* (2013), Giannopoulos (2018) and Malik and Thomas (2010).

Comparing the ranking of classification risk models in studies conducted on the P2P platform in general and P2P LC in particular yielded numerous results. This can be explained for the following reasons: (i) purpose of the study, since some studies compared classifications to introduce the proposed model or to combine it with other factors to improve performance of the models such as balanced data techniques, soft information variables, etc. (Niu *et al.*, 2019); (ii) some models were not fine-tuned to achieve hyperparameter optimization space (Bae *et al.*, 2018); (iii) the choice of variables is different when some studies use "leaking" data, i.e. the data contains information obtained after a loan was accepted or variables with predictive power such as *grade*, *sub_grade*, *int_rate* created by LC itself to classify borrowers (Reddy, 2016).

5.2 Managerial implications

P2P lending services have been emerging in recent years as a channel for rapidly distributing and supplying capital to the market, especially to the consumer loan market. In addition to a lending system, these platforms also have to build a borrower credit risk analysis system to reduce investors' losses and increase the platforms' competitiveness. This becomes an inspiration for empirical studies that evaluate credit risk models matching the characteristics of this lending platform. However, when making a decision to choose one or more credit risk models, it is necessary to also consider other conditions, such as data, time, trade-offs between costs (calculation, machinery, science technology, etc.), and profit. Each classification model has its own advantages and disadvantages. Models such as ANN and GBDT give good predictive results, but sometimes they are not as popular as LR and LDA because these "black box" models are incapable of explaining or describing problems such as the reason why borrowers are rejected or the complexity of the algorithms and the modeling process that require an investment of specialized knowledge and skills by an employee or company. In addition, the processing speed of classification models is also evaluated as one of the potential criteria for financial intermediaries when choosing the right model. Models such as SVM or GBDT take a long time to train, find hyperparameter optimization space and predict results in large datasets in comparison with the rest of the methods included in this study. Moreover, macroeconomic factors such as economic crisis, unemployment, etc. should be taken into consideration because it is possible that the input data used to train a model in these cases will not ensure quality requirements, easily leading to misleading performance of the models, erroneous results and causing widespread consequences. Therefore, it is necessary to monitor market warning signs and choose an appropriate threshold decision to refuse a loan or a reasonable interest rate to compensate for the minimum default risk.

5.3 Limitations and future research agenda

The models used in this paper are those that are considered common when predicting borrower credit risk. Many of the proposed models achieve good results in data mining, such as Xgboost model or models that combine different machine learning models to give optimal results, which this paper has not yet mentioned. In addition, this study only assesses the differences in the performance of risk prediction models across economic scenarios but has not yet provided quantitative solutions to improve the models in those cases. Hence, the limitations of this topic can also become an inspiration for further research streams such as finding classification models combining many learners in order to better predict or classify loan applications or quantitative solutions to improve the quality of credit risk models in terms of economic scenarios. In addition, research directions to evaluate the default risk of the borrower after a loan has been accepted, such as assessing the probability of delinquent payment, can also be considered.

6. Conclusion

This study compares nine credit risk prediction models on consumer loan data of P2P LC platform, divided into two phases: a difficult economic period (2007–2012) and a stage of economic recovery and development (2013–2019). To avoid imbalanced data, the study uses a random undersampling technique and performance metrics for imbalanced data. The main results of the study show that the performance of credit classification or scoring models is influenced by economic factors. The performance of credit risk models in the 2013–2019 period is better evaluated through three families of measures compared to the difficult economic period (2007–2012). Therefore, threshold decisions to finance loans or lending interest rates should also be updated according to the economic situation to minimize credit

risk. In addition, the results of ranking-default risk classification models show that GBDT model is the best model in most of the study's metrics or families of metrics. This study also supports the research results of Tsai *et al.* (2014) and Teplý and Polena (2019) that LR, ANN and LDA models predict or classify default risk loan applications quite well (just below GBDT) and discourage using CART, k-NN and NB models to predict borrowers on P2P loan data.

Notes

1. The term *classification of loan application or credit scoring* is used to distinguish it from behavioral and performance scoring, referring to monitoring the repayment behavior of credit-granted customers.
2. According to Malekipirbazari and Aksakall (2015), default risk can be considered as missing three scheduled payments or missing three consecutive payments in the period.
3. Niu *et al.* (2019).
4. In 2007–2012, in addition to crisis and post-crisis, P2P LC has just participated in building and creating a market. In 2013–2019, P2P LC has stabilized and dominated the P2P lending market.
5. Credit data are characterized by a lack of information, but excluding all missed features can reduce the sample size and lead to the loss of valuable information. Therefore, in addition to observing the distribution of data to select the threshold of missed variables, the study also combines expert knowledge and the special perceptions of the borrower to remove these variables (Dinh *et al.*, 2013).
6. See a list of states by region from https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States.
7. The K-fold cross-validation used here is K-fold stratified, meaning that the data in each fold still guarantees a percentage of each subclass as original data.
8. Grid search lists all combinations of hyperparameters, performs a batch test and gives optimal hyperparameter space based on the given criteria.
9. This means that the number of models and ranking of models in the comparison table may not be consistent with those in the original papers, but basically, the relative ranking results between models in those studies will be unaffected.
10. Teplý and Polena (2019) does not include GBDT model, which is the best performing model in this research.

References

Bae, J.K., Lee, S.I. and Seo, H.J. (2018), "Predicting online peer-to-peer (P2P) lending default using data mining techniques", *Proceedings of 20th Asia-Pacific Conference on Global Business, Economics, Finance and Social Sciences*, Hong Kong, August 1-2, 2018, SAR – PRC, Paper ID:H808.

Brown, L. and Mues, C. (2012), "An experimental comparison of classification algorithms for imbalanced credit scoring data sets", *Expert Systems with Applications*, Vol. 39 No. 3, pp. 3446-3453, doi: 10.1016/j.eswa.2011.09.033.

Brownlee, J. (2020), *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*, Machine Learning Mastery, Vermont.

Chang, S., Kim, S.D. and Kondo, G. (2015), "Predicting default risk of lending club", Vol. CS229, Machine Learning.

Dinh, T.H.T., Kleimeier, S. and Straetmans, S.T.M. (2013), "Bank lending strategy, credit scoring and financial crises", in *Research Memoranda 053*, Maastricht University, Graduate School of Business and Economics (GSBE), doi: 10.26481/umagsb.2013053.

Fernández, A., García, S., Galar, M., Prati, R.C. and Krawczyk, B. (2018), *Learning from Imbalanced Data Set*, 1st ed. Edition, Springer, Berlin, doi: 10.1007/978-3-319-98074-4.

Ferri, C., Hernández-Orallo, J. and Modroiu, R. (2009), "An experimental comparison of performance measures for classification", *Pattern Recognition Letters*, Vol. 30 No. 1, pp. 27-38, doi: 10.1016/j.patrec.2008.08.010.

Giannopoulos, V. (2018), "The effectiveness of artificial credit scoring models in predicting NPLs using micro accounting data", *Journal of Accounting and Marketing*, Vol. 7 No. 4, doi: 10.4172/21689601.1000303.

Hand, D.J. and Henley, W.E. (1997), "Statistical classification methods in consumer credit scoring: a review", *Journal of the Royal Statistical Society*, Vol. 160 No. 3, pp. 523-541, doi: 10.1111/j.1467-985X.1997.00078.x.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Springer Series in Statistics, New York, doi: 10.1007/978-0-387-84858-7.

He, H. and Ma, Y. (Eds) (2013), *Imbalanced Learning: Foundations, Algorithms, and Applications*, John Wiley & Sons, doi: 10.1002/9781118646106.

Ince, H. and Aktan, B. (2009), "A comparison of data mining techniques for credit scoring in banking: a managerial perspective", *Journal of Business Economics and Management*, Vol. 10 No. 3, pp. 233-240, doi: 10.3846/1611-1699.2009.10.233-240.

Jin, Y. and Zhu, Y. (2015), "A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending", *Fifth International Conference on Communication Systems and Network Technologies*, pp. 609-613, doi: 10.1109/CSNT.2015.25.

Lin, X., Li, X. and Zheng, Z. (2016), "Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China", *Applied Economics*, Vol. 49 No. 35, pp. 3538-3545, doi: 10.1080/00036846.2016.1262526.

Madzova, V. and Ramadini, N. (2013), "Can credit scoring models prevent default payments in banking industry in the period of financial crisis?", *International Journal of Business and Technology*, Vol. 2 No. 1, pp. 32-38, doi: 10.33107/ijbte.2013.2.1.05.

Malekipirbazari, M. and Aksakall, V. (2015), "Risk assessment in social lending via random forests", *Expert Systems with Applications*, Vol. 42 No. 10, pp. 4621-4631, doi: 10.1016/j.eswa.2015.02.001.

Malik, M. and Thomas, L.C. (2010), "Modelling credit risk of portfolio of consumer loans", *Journal of the Operational Research Society*, Vol. 61 No. 3, pp. 411-420, doi: 10.1057/jors.2009.123.

Namvar, E. (2013), "An introduction to peer-to-peer loans as investments", *Journal of Investment Management*, First Quarter, doi: 10.2139/ssrn.2227181.

Namvar, A., Siami, M., Rabhi, F. and Naderpour, M. (2018), "Credit risk prediction in an imbalanced social lending environment", *International Journal of Computational Intelligence Systems*, Vol. 11 No. 1, pp. 925-935, doi: 10.48550/arXiv.1805.00801.

Niu, B., Ren, J. and Li, X. (2019), "Credit scoring using machine learning by combing social Network information: evidence from peer-to-peer lending", *Information*, Vol. 10 No. 12, p. 397, doi: 10.3390/info10120397.

Odeh, O.O., Featherstone, A.M. and Sanjoy, D. (2006), "Predicting credit default in an agricultural bank: methods and issues", *2006 Annual Meeting, Orlando, Florida 35359, Southern Agricultural Economics Association*, doi: 10.22004/ag.econ.35359.

Reddy, S. (2016), "Peer to peer lending, default prediction evidence from lending club", *Journal of Internet Banking and Commerce*, Vol. 21 No. 3, pp. 1-19.

Shmueli, G., Bruce, P.C., Yahav, I., Patel, N.R. and Lichtendahl, K.C. (2018), *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*, 1st ed., Wiley, Hoboken.

Tao, W. and Chang, D. (2019), "Credit risk assessment of P2P lending borrowers based on SVM", *Advances in Economics, Business and Management Research*, Vol. 80, pp. 182-190, doi: 10.2991/bems-19.2019.33.

Teplý, P. and Polena, M. (2019), "Best classification algorithms in peer-to-peer lending", *North American Journal of Economics and Finance*, Vol. 51, 100904, doi: 10.1016/j.najef.2019.01.001.

Tsai, K., Ramiah, S. and Singh, S. (2014), *Peer Lending Risk Predictor*, Stanford University, Stanford, California, Vol. CS229, doi: 10.13140/2.1.4810.6567.

Wendler, T. and Gröttrup, S. (2016), "Neuronal networks", in *Data Mining with SPSS Modeler: Theory, Exercises and Solutions*, Springer International Publishing, pp. 833-878, doi: 10.1007/978-3-319-28709-6.

Xiao, W., Zhao, Q. and Fei, Q. (2006), "A comparative study of data mining methods in consumer loans credit scoring management", *Journal of Systems Science and Systems Engineering*, Vol. 15 No. 4, pp. 419-435, doi: 10.1007/s11518-006-5023-5.

Yeh, I.C. and Lien, C.H. (2009), "The comparisons of data mining techniques for predictive accuracy of probability of default of credit card clients", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 2473-2480, doi: 10.1016/j.eswa.2007.12.020.

Zhang, D., Huang, H., Chen, Q. and Jiang, Y. (2007), "A comparison study of credit scoring models", Proceedings of the 3rd International Conference on Natural Computation, Haikou, China, pp 15-18, doi:10.1109/ICNC.2007.15.

(The Appendix follows overleaf)

Appendix

Abbreviated name	Description
addr_state	State provided by borrower
annual_inc	Self-reported annual income
application_type*	Individual or joint application with two co-borrowers
bc_open_to_buy*	Total open to buy on revolving bankcards
chargeoff_within_12_mths	Number of charge-offs within 12 months
delinq_2 yrs	Number of 30+ days past-due incidences of delinquency for the past 2 years
Dti	Ratio using borrower's total monthly debt payments on total debt obligations, excluding mortgage and requested LC loan, divided by borrower's self-reported monthly income
earliest_cr_line	Month borrower's earliest reported credit line was opened
emp_length	Employment length in years
fico_score	Average of fico_range_low and fico_range_high
home_ownership	Home ownership status provided by borrower during registration or obtained from credit report: RENT, OWN, MORTGAGE, OTHER
initial_list_status	Initial listing status of loan. Possible values are – W, F
inq_last_6 mths	Number of inquiries in the past 6 months (not including auto and mortgage inquiries)
loan_amnt	Listed amount of loan applied by borrower
loan_status	Current status of loan
mort_acc*	Number of mortgage accounts
num_accts_ever_120_pd*	Number of accounts 120 or more days past due
num_bc_tl*	Number of bankcard accounts
num_il_tl*	Number of installment accounts
num_rev_tl_bal_gt_0*	Number of revolving trades with balance >0
num_tl_90g_dpd_24m*	Number of accounts 90 or more days past due in the last 24 months
num_tl_op_past_12m*	Number of accounts opened in the past 12 months
open_acc	Number of open credit lines in borrower's credit file
pct_tl_nvr_dlq*	Percentage of trades that have never been delinquent
percent_bc_gt_75*	Percentage of all bankcard accounts >75% of limit
pub_rec	Number of derogatory public records
Purpose	Category provided by borrower for loan request
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate
tax_liens	Number of tax liens
Term	Number of payments on loan; values are in months and can be either 36 or 60
tot_coll_amt*	Total collection amounts ever owed
tot_hi_cred_lim*	Total high credit/credit limit
total_il_high_credit_limit*	Total installment high credit/credit limit
verification_status	Indicates if income was verified by LC, not verified or if the income source was verified
hardship_flag*	Flags whether the borrower is on a hardship plan

Note(s): Where * indicates variables which are only included in 2013–2019
Source(s): Own elaboration

Table A1. Attributes of selected features

Classifications	Best hyperparameter space (2007–2012)	Best hyperparameter space (2013–2019)
LR	n/a	n/a
NB	n/a	n/a
LDA	n/a	n/a
k-NN	{"n_neighbors": 16, "weights": "distance"}	{"n_neighbors": 16, "weights": "distance"}
SVM	{"C": 1, "gamma": 1, "kernel": "rbf"}	{"C": 10, "gamma": 0.001, "kernel": "rbf"}
CART	{"max_depth": 5, "min_samples_leaf": 100}	{"max_depth": 10, "min_samples_leaf": 150}
ANN	{"alpha": 0.1, "hidden_layer_sizes": (10), "max_iter": 50}	{"alpha": 0.1, "hidden_layer_sizes": (10), "max_iter": 50}
RF	{"n_estinamtors": 50, "max_features": 4, "min_samples_leaf": 4, "min_samples_split": 9}	{"n_estinamtors": 50, "max_features": 6, "min_samples_leaf": 2, "min_samples_split": 4}
GBDT	{"learning_rate": 0.05, "loss": "deviance", "max_depth": 8, "max_features": 0.1, "min_samples_leaf": 150, "n_estimators": 300}	{"learning_rate": 0.1, "loss": "deviance", "max_depth": 8, "max_features": 0.3, "min_samples_leaf": 100, "n_estimators": 300}

Source(s): Own elaboration

Table A2. Tuning the hyperparameters of an estimator

Corresponding author
Lua Thi Trinh can be contacted at: lua.trinh2017@qcf.jvn.edu.vn