Ensemble methods for credit scoring of Chinese peer-to-peer loans

Wei Cao, Yun He, Wenjun Wang, Weidong Zhu, Yves Demazeau

Ensemble methods for credit scoring of Chinese peer-to-peer loans

Wei Cao,[1] Yun He,[1] Wenjun Wang,[1] Weidong Zhu[1] and Yves Demazeau[2]

[1]School of Economics, Hefei University of Technology, 193 Tunxi Road, Hefei, Anhui 230009, People's Republic of China; emails: weicao@hfut.edu.cn, heyun779664@hfut.edu.cn, wjwang@hfut.edu.cn, zhuwd@hfut.edu.cn

[2]Laboratoire d'Informatique de Grenoble, Center National de la Recherche Scientifique, CS 40700, Grenoble 38058, France; email: yves.demazeau@imag.fr

(Received October 12, 2020; accepted February 23, 2021)

ABSTRACT

Risk control is a central issue for Chinese peer-to-peer (P2P) lending services. Although credit scoring has drawn much research interest and the superiority of ensemble models over single machine learning models has been proven, the question of which ensemble model is the best discrimination method for Chinese P2P lending services has received little attention. This study aims to conduct credit scoring by focusing on a Chinese P2P lending platform and selecting the optimal subset of features in order to find the best overall ensemble model. We propose a hybrid system to achieve these goals. Three feature selection algorithms are employed and combined to obtain the top 10 features. Six ensemble models with five base classifiers are then used to conduct comparisons after synthetic minority oversampling technique (SMOTE) treatment of the imbalanced data set. A real-world data set of 33 966 loans from the largest lending platform in China (ie, the Renren lending platform) is used to evaluate performance. The results show that the top 10 selected features can greatly improve performance compared with all features, particularly in terms of discriminating "bad" loans from "good" loans. Moreover, comparing the standard

evaluations, robustness tests and statistical tests suggests that the gradient boosting decision tree, random forest and rotation forest methods are the best. Our findings can help risk managers and investors by providing them with correct warning signals and the main factors influencing "bad" loans, so that they can take corrective actions and reduce risk.

Keywords: credit scoring; ensemble learning; feature selection; synthetic minority oversampling technique (SMOTE) treatment; Chinese peer-to-peer (P2P) lending.

1 INTRODUCTION

Peer-to-peer (P2P) lending, also known as person-to-person lending, has quickly emerged with the rise of the Internet and the development of private lending. In China, the first P2P lending platform was established in 2007, and the industry has since undergone rapid development as the core of inclusive finance and the sharing economy, receiving much attention. As of November 2018, the cumulative number of platforms was 6429 and the loan volume was Rmb8 trillion.

P2P lending plays a unique and important role in China and has certain particular characteristics. First, the default rate in China is much higher than that in other countries. By year-end 2018, the total number of failed platforms with risk issues (including transformed and closed platforms) had increased to 5242,[1] accounting for 81.54% of all platforms. Second, most P2P borrowers in China are small and medium-sized enterprises (SMEs) and self-employed individuals (Gao *et al* 2018). Thus, the high default rate has done enormous damage to the social economy, since SMEs contribute 60% of gross domestic product (GDP). Third, the main underlying cause of risk is the lack of efficient governance, in particular both a lack of corresponding P2P lending data and outdated credit scoring methods. Thus, in China, how to distinguish creditworthy borrowers from those who will probably default on repayment is a very important and challenging problem.

Many researchers have focused on credit scoring and have proposed useful methods. In general, with regard to credit scoring models, there are three main research streams. The first kind of approach is based on subjective judgments made by human experts using past experiences and basic principles (Abdou and Pointon 2011). However, this kind of method suffers from high training costs and, more importantly, from inconsistent decisions by different experts when facing the same application (Marqués *et al* 2012). The second kind of classical approach is based on statistical and mathematical models, such as linear logistic regression (LR) models (Hu and Ansell 2007; Hua *et al* 2007; Thomas 2009), linear discriminate analysis (LDA)

[1] Source URL: www.wdzj.com.

(Rosenberg and Gleit 1994) and multiple discriminant analysis (MDA) (Altman 1968). Since classical statistical models have some assumptions that cannot fit real-world conditions (eg, linear assumptions), various machine learning methods without any previous conditions have started to be employed. Among these artificial intelligence-based techniques, the most popular single classifiers applied to credit scoring are support vector machines (SVMs) (Harris 2015; Hens and Tiwari 2012; Huang *et al* 2007), artificial neural networks (ANNs) (Lee and Chen 2005; Malhotra and Malhotra 2002; West 2000; Zhao *et al* 2015) and decision trees (DTs) (Abellán and Castellano 2016; Bijak and Thomas 2012; Tsymbal *et al* 2005; Yap *et al* 2011).

Since a single classifier cannot capture the fine-grained nuances of various features and credit records, in recent years more attention has been paid to the use of classifier ensembles because of their ability to integrate multiple classifiers, which can lead to better performance (Abellán and Castellano 2016; Xiao *et al* 2016). Among them, there are six popular ensembles with promising results: the bagging (Hsieh *et al* 2010; Sun *et al* 2014), boosting (Bian and Wang 2007; Paleologo *et al* 2010; Sun *et al* 2016), random subspace (Marqués *et al* 2012; Nanni and Lumini 2009), rotation forest, random forest (Ala'raj and Abbod 2016; Twala 2010) and gradient boosting decision tree (GBDT) (Ma *et al* 2018) methods.

Although many studies have demonstrated the superiority of ensemble methods (Abellán and Castellano 2016; He *et al* 2018), few studies have focused on selecting the best ensemble in a specific context. To the best of our knowledge, no comparative study has been carried out to answer this question in the Chinese context, even though there is an urgent need to solve this problem. Moreover, it is important to find the most influential features that can discriminate "bad" loans from "good" loans, which will improve model interpretability and reduce the cost of risk control.

Taking these considerations into account, this paper aims to answer the following two questions.

(1) What are the most relevant features of P2P borrowers who will probably default in the future?

(2) What is the most accurate ensemble model for solving the Chinese credit scoring problem?

This study makes three contributions to the literature. First, it selects the most discriminative factors that can lead to better credit scoring by combining three kinds of typical feature selection methods. Second, it examines the performance of six effective ensemble methods with five well-known base classifiers. The results are analyzed with standard evaluation metrics (accuracy, Type I error and the F-measure) and statistical tests (Friedman and post hoc tests). Third, this study is among the small number of studies focusing on the Chinese credit scoring problem.

In this paper, we propose a hybrid system to select the most appropriate ensembles and find the major influencing factors that can discriminate between defaulting and nondefaulting borrowers. The remainder of this paper is organized as follows. The related literature is reviewed in Section 2. Section 3 provides a brief overview of the feature selection methods and ensembles used in this study. The proposed system, which involves four specific steps, is discussed in detail in Section 4. Section 5 presents the experimental results. Finally, Section 6 is devoted to our conclusions.

2 RELATED WORK

This study can be divided into two parts: feature selection methods, and ensemble-based credit scoring. In this literature review, the two issues are reviewed in detail.

2.1 Feature selection methods

A feature selection algorithm, as a preprocessing method of model creation, is used in the process of eliminating features from a data set that are irrelevant to the task being performed (Chandrashekar and Sahin 2014; Guyon *et al* 2003; Kumar 2014). The major benefits of such an algorithm are as follows.

(1) It facilitates one's understanding of data since it finds useful features to represent the data and removes the irrelevant features.

(2) It reduces computational time by reducing the dimensionality of data.

(3) It improves model quality and the interpretability of the outcome.

In general, there are three major types of feature selection approaches are applied to credit scoring (Jiménez *et al* 2016; Miao and Niu 2016). The first is the filter approach, which selects variables by ranking them with information generated from data and independent of the classification model. Filter methods that are often used in credit scoring include the information gain ratio (Wang *et al* 2017), mutual information (Zhang *et al* 2018), rough set (Ping and Lu 2011), principal component analysis (PCA) (Šušteršič *et al* 2009) and LDA (Chen and Li 2010). The second is the wrapper approach, which assesses a subset of features based on their performance with a given classifier model such as SVMs or LR models. Typical methods include the genetic algorithm (GA) (Koutanaei *et al* 2015), the wrapper method based on the genetic algorithm (Ala'raj and Abbod 2016) and recursive feature elimination (Bellotti and Crook 2009). The third is the embedded approach, in which feature selection is embedded in the classifier. The most typical models are DT-based approaches (Wang *et al* 2017), such as the classification and regression tree (CART) and C4.5 algorithms.

The studies cited above have demonstrated that using feature selection methods in credit scoring can improve prediction performance. However, as illustrated in Zhang *et al* (2018), different kinds of feature selection models have their own advantages and disadvantages, and there is no overall best feature selection method for the credit scoring problem. Currently, researchers are attempting to combine different kinds of feature selection methods to achieve better performance, and the studies by Chen and Li (2010) and Zhang *et al* (2018) have shown that the combination approach is superior to a single feature selection method.

2.2 Ensembles for credit scoring

Ensemble learning is a machine learning paradigm in which multiple learners are trained to solve the same problem. A classifier ensemble contains a number of individually trained base classifiers (eg, SVMs, ANNs) whose decisions are combined for use, and when classifying new samples, the most popular combination schemes are weighted or unweighted. There are many ensembles that have been applied to solving the credit scoring problem, and the bagging (Hsieh *et al* 2010; Sun *et al* 2014), boosting (Bian and Wang 2007; Paleologo *et al* 2010; Sun *et al* 2016), random subspace (Marqués *et al* 2012; Nanni and Lumini 2009), rotation forest, random forest (Ala'raj and Abbod 2016; Twala 2010) and GBDT (Ma *et al* 2018) methods are currently the most extensively used ensembles.

The works cited above have verified the superior stability and learning results of these models compared with single base classifiers. However, when dealing with the credit scoring problem, it remains an open question of which ensemble performs best. Recently, some researchers have begun to compare the effectiveness of ensembles. Lessmann *et al* (2015) compared individual base classifiers, homogeneous ensembles (eg, bagging, random forest) and heterogeneous ensembles (eg, dynamic ensembles) based on several credit data sets, including Australian, German, Benelux and UK credits, and they concluded that the random forest method achieves promising performance. García *et al* (2019) compared seven ensembles with different data sets, such as Australian, Finnish and German credits, and their results revealed that the performance of ensembles depends on the sample types in the data set; therefore, comparisons between various ensembles should be discussed with regard to a specific data set. The studies by Feng *et al* (2018) and Wang *et al* (2011) constitute similar research. However, most current studies focus on data sets in developed countries such as the United States, Australia and Germany, while minimal attention has been paid to the Chinese P2P market.[2] Moreover, to the best of our knowledge, comparisons between the six popular ensembles listed above are still lacking in the Chinese context.

[2] The reason may be that the Chinese P2P data are not open source.

In this study, we propose a hybrid system to find the most superior ensemble for Chinese P2P lending services. The system first selects the most relevant features by combining three kinds of typical feature selection methods. Subsequently, the selected features are fed into six popular ensembles with five base classifiers, and the results are then compared from standard and statistical perspectives.

3 METHOD

In this section, the three feature selection algorithms and six ensembles applied in this study are described in detail.

3.1 Overview of feature selection algorithms

Information gain ratio feature selection

This method is a classical filter approach that separates feature selection from classification methods, and features are ranked directly based on the data set (Koutanaei *et al* 2015). It is based on the information entropy concept. The standard form used in this paper is as follows: suppose there are n features for each borrower $F_1, F_2, \ldots, F_n$, and the value of an attribute i $(1 \leqslant i \leqslant n)$ is assessed by determining the information gain ratio (entropy difference) with respect to the class, where the information gain ratio is defined as the quotient of information gain and split information. The larger the information gain ratio is, the more likely that feature i is a discriminative factor.

GA feature selection

This method is a widely used wrapper approach that first uses an optimization algorithm to add or remove features and to generate different subsets and that then evaluates these subsets with classifiers (Wang *et al* 2017). It is based on the GA algorithm (Goldberg 2006). In this study, one chromosome is a set of features $(F_1, F_2, \ldots, F_n)$ for each borrower, and the Goldberg strategy is applied to discover an ideal set of features. The subset features are assessed based on the classification accuracy under the SVM classifier. The initial population, maximum number of generations, mutation probability, crossover probability, cross-validation and random seed number are 20, 0.033, 0.6, 10 and 1, respectively.

CART feature selection

This method is a typical embedded approach that uses all features to generate a model and then analyzes the model to infer the importance of variables (Questier *et al* 2005). This study selects the CART algorithm as a model and assumes that there are n features for each borrower $F_1, F_2, \ldots, F_n$. The CART algorithm is applied to identify

and construct binary DTs by using training samples; then, the subset features are assessed based on the predictive power of the trees.

3.2 Overview of classifier ensembles

Bagging

This ensemble method, also known as bootstrap aggregating, was proposed by Breiman (1996) and is rooted in bootstrap resampling and aggregating. The standard form used in this paper is shown in Algorithm 1: given a training set D (ie, "good" and "bad" loans) of size n, bagging generates m new training sets D_i, $i = 1, \ldots, m$, each of size n', by sampling from D uniformly and with replacement. Thus, some instances may be repeated in the training set, while others may be left out. Each sample D_i is then used to train a base classifier C_i. Finally, predictions of new samples are made by taking the majority vote of classifiers $C_1, C_2, \ldots, C_m$. Since each classifier is built with a different training set, the classifiers are different from each other. Additionally, bagging seeks to reduce the variance and to avoid overfitting due to the variance of classifiers C_i.

Algorithm 1 Bagging algorithm pseudocode.

Input: training set $D = (\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$; base classifiers Ω; number of iterations m

1: **for** $i = 1, 2, \ldots, m$ **do**
2: $\quad C_i = \Omega(D_i)$ $\quad$ % D_i is generated from D by bootstrap sampling
3: **end for**

Output: $H(\boldsymbol{x}) = \arg\max_y \sum_{i=1}^{m} \mathbb{I}(C_i(\boldsymbol{x}) = y)$

Boosting

This is a sequential ensemble method based on the work of Freund and Schapire (1996). It constructs a sequence of classifiers where each depends on its predecessors, and in particular, it focuses more on the error of the previous classifier. In this paper, we use the best known adaptive boosting (AdaBoost) algorithm in the boosting family. The procedure is shown in Algorithm 2: given a training set D (ie, "good" and "bad" loans) of size n, AdaBoost builds a sequence of base classifiers $C_1, C_2, \ldots, C_m$ with samples $D_1, D_2, \ldots, D_m$ that are tweaked in favor of instances misclassified by previous classifiers. The instances misclassified by C_{i-1} are more likely to appear in the next classifier C_i. The final predictions of new instances are obtained through a weighted vote of base classifiers, and the weights are computed based on the performance of the base classifiers.

Algorithm 2 AdaBoost algorithm pseudocode.

Input: training set $D = (\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$; base classifiers Ω; number of iterations m

Initialize distribution over the training set

$D_1(\boldsymbol{x}) = 1/n$

1: **for** $i = 1, 2, \ldots, m$ **do**
2: $C_i = \Omega(D_i)$
3: $\alpha_i = \frac{1}{2}\ln((1-\epsilon i)/\epsilon i)$ % ϵi is the weighted error
4: $D_{i+1}(\boldsymbol{x}) = \frac{D_i(\boldsymbol{x})\mathrm{e}^{-\alpha_i y_t C_i(\boldsymbol{x})}}{Z_i}$ % update distribution
5: **end for**

Output: $H(\boldsymbol{x}) = \mathrm{sign}(\sum_{i=1}^{m} \alpha_i C_i(\boldsymbol{x}))$

Random subspace

This method, also known as feature bagging, was proposed by Ho (1998). It is similar to bagging except that the features are randomly sampled, with replacement, for each classifier. The procedure used in this paper is depicted in Algorithm 3: given a training set D (ie, "good" and "bad" loans) of size n, random subspace generates m base classifiers $C_1, C_2, \ldots, C_m$ with samples $D_1, D_2, \ldots, D_m$. Each classifier uses only one subset of all the features (ie, ratios) in D. These features are randomly selected from the full feature set. Similar to bagging, predictions of new instances are generated via a simple majority vote algorithm. The method can improve the generalization error with the independent submodel (classifier).

Algorithm 3 Random subspace algorithm pseudocode.

Input: training set $D = (\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$; base classifiers Ω; number of iterations m; number of features F

1: **for** $i = 1, 2, \ldots, m$ **do**
2: $F_i = \mathrm{RS}(D, f1)$ % random select $f1$ features from D
3: $D_i = \mathrm{map}_{F_i}(D)$ % D_i is the data set for the features in F_i
4: $C_i = \Omega(D_i)$ % D_i is generated from D by bootstrap sampling
5: **end for**

Output: $H(\boldsymbol{x}) = \arg\max_y \sum_{i=1}^{m} \mathbb{I}(C_i(\mathrm{map}_{F_i}(\boldsymbol{x})) = y)$

Rotation forest

This method, proposed by Rodriguez *et al* (2006), creates an ensemble of models whose estimation is performed using a set of features extracted from original data. As shown in Algorithm 4, given a training set D (ie, "good" and "bad" loans) of size n, with feature set F, the method separates feature set F into K nonoverlapping subsets of equal size. Then a PCA on each K subset of features is applied, and new feature sets are constructed by integrating all principal components. Base classifier C_i is

finally trained with the new data set. This method enhances the individual accuracy of classifier C_i by contributing to diversity.

Algorithm 4 Rotation forest algorithm pseudocode.

Input: training set $D = (\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$; base classifiers Ω; number of iterations m; number of features F; number of subset features K

1: **for** $i = 1, 2, \ldots, m$ **do**
2: $F_{ij} = \mathrm{RS}(D, F/j)$, $j = 1, 2, \ldots, K$ % split F into K subsets
3: **for** $j = 1, 2, \ldots, K$ **do**
4: $X_{ij} = \mathrm{map}_{F_{ij}}(D)$ % X_{ij} is the data set for the features in F_{ij}
5: $R_i = \mathrm{PCA}(X_{ij})$ % construct rotation matrix R_i by applying PCA to X_{ij} and arranging the results
6: **end for**
7: $C_i = \Omega(X, R_i, Y)$
8: **end for**

Output: $H(\boldsymbol{x}) = \arg\max_y \sum_{i=1}^{m} \mathbb{I}(C_i(\boldsymbol{x}) = y)$

Random forest

This method is composed of several unrelated DTs, and each is constructed with samples and feature samplings (Breiman 2001). The procedure is shown in Algorithm 5: given a training set D (ie, "good" and "bad" loans) of size n with feature set F, in each iteration, a random forest selects a random subsample D_i of the included features by means of bootstrapping and then generates tree T_i from D_i using the CART algorithm. After constructing random trees, predictions on new instances are performed through a voting scheme. The method is robust to overfitting since each forest is presented with only a subset of all features.

Algorithm 5 Random forest algorithm pseudocode.

Input: training set $D = (\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$; base classifiers CART; number of iterations m

1: **for** $i = 1, 2, \ldots, m$ **do**
2: $C_i = \mathrm{CART}(D_i)$ % D_i is generated from D by bootstrap sampling, and the feature split is different between random forests and traditional decision trees
3: **end for**

Output: $H(\boldsymbol{x}) = \arg\max_y \sum_{i=1}^{m} \mathbb{I}(C_i(\boldsymbol{x}) = y)$

Gradient boosting decision tree (GBDT)

This advanced sequential ensemble method is another widely used tree-based boosting algorithm (Friedman 2001). The basic idea of boosting is to combine a series of

weak base learners to form a strong learner. Compared with the traditional boosting algorithm, the GBDT algorithm constructs a new model in a gradient direction of the residuals to minimize the loss function generated by each iteration, which in turn improves performance (Xia *et al* 2017). The specific process is shown in Algorithm 6.

Algorithm 6 GBDT algorithm pseudocode.

Input: training set $D = (\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$; base classifiers Ω; number of iterations m; loss function $L(y, F(x))$

Initialize the model with a constant value

$F_0(x) = \arg\min_\gamma \sum_{t=1}^{n} L(y_t, \gamma)$

for $i = 1, 2, \ldots, m$ **do**

  $r_{ti} = -[\partial L(y_t, F(X_t))/\partial F(X_t)], t = 1, 2, \ldots, n$ % compute pseudoresiduals

  $C_i = \Omega(x_t, r_{ti})$ % fit base learner

  $\gamma_i = \arg\min_\gamma \sum_{t=1}^{n} L(y_t, F_{i-1}(x_t) + \gamma C_i(x_t))$ % compute multiplier

  $F_i(x) = F_{i-1}(x) + \gamma_i C_i(x)$ % update the model

end for

Output: $H(\boldsymbol{x}) = \text{sign}(F_m(\boldsymbol{x}))$

4 THE PROPOSED MODEL

To conduct the research, we propose a hybrid approach that combines feature selection algorithms and ensemble methods. As shown in Figure 1, it consists of five main phases. In Phase I, data are collected from the Renren lending platform, which is the largest P2P platform in China as well as being one of the oldest. We collect five kinds of data: the characteristics of borrowers, the financial information of borrowers, the credit history of borrowers, the loan characteristics and the platform authentication information. An appropriate data preprocessing operation is involved. In Phase II, three feature selection algorithms are combined to obtain a subset of features that provides the most discriminative accuracy and interpretive power. In Phase III, the synthetic minority oversampling technique (SMOTE) algorithm is used to address the imbalanced data set. In Phase IV, the parameters of six ensembles and five base classifiers are tuned, and models are constructed. In Phase V, the models are compared and evaluated using three evaluation metrics and statistical tests.

4.1 Phase I: data collection and preprocessing

4.1.1 Data

The data used in this section are loan data from the Renren lending platform,[3] which cover the period from January 2013 to December 2016. Importantly, we only collect

[3] URL: www.renrendai.com.

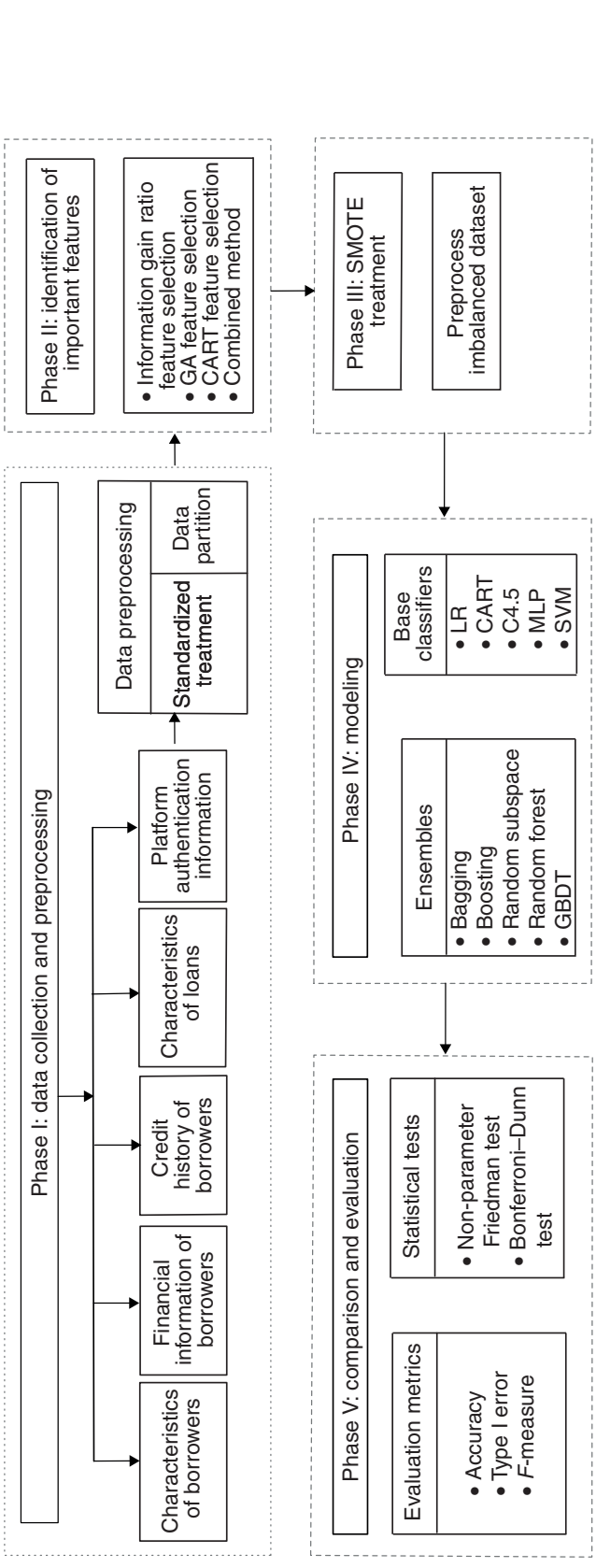


FIGURE 1 Overview of system design.

TABLE 1 Description of different loan statuses

Loan status	Description
Current	The loan is up to date on all outstanding payments
Op (1–29)	The loan has not been current for 1 to 29 days
Op (30–89)	The loan has not been current for 30 to 89 days
Op (90+)	The loan has been past due for more than 90 days
Charge off	The loan has been past due for more than 120 days, so there is no reasonable expectation of sufficient payment to prevent charge off
Fully paid	The loan has been fully repaid

information that is visible to all visitors, and no personal identifiable information exists in the data set. The data set is collected from all borrowers who applied for a personal loan over the observation period. The total number of loans is 73 395. According to the platform, six types of loan status exist (shown in Table 1). It is easy to see that only "fully paid" and "charge off" indicate that a loan has reached its final state; the other four statuses indicate that a loan is still in process. Therefore, in our experiment, we consider only "fully paid" ("good" loan) and "charge off" ("bad" loan) events. The final number of samples is 33 966, including 474 "bad" loans.

As described in Table 2 (details on the values and meanings of the features are listed in Table A1 in the online appendix) there are 22 features in the data set, which consists of five subsets. The first subset is the characteristics of borrowers, and it consists of seven features: age, educational level, marital status, length of employment, place of employment, industry and company size. The second subset is the financial information of borrowers, including their income, house, car, mortgage and auto loan. The third subset is the credit history of borrowers, which consists of the number of successful loans and the number of applications. The fourth subset is the characteristics of loans and includes the loan amount, loan purpose, loan term and loan type. The last subset is platform authentication information, and it consists of credit, identity, job and income authentication.

Table 3 presents the summary statistics of the numeric and binary features in our data set. The average marital status is 0.72, which is greater than 0.5 so indicates that the majority of borrowers are married. We also see that 94% of borrowers have a car, but only 61% of them own a house; more interestingly, 60% of borrowers have a mortgage, which is nearly equal to the number of borrowers who own a house. In addition, it is easy to see that the mean of applications is larger than the number of successful loans. In Table 3, we also find that most borrowers provide credit authen-

TABLE 2 Features listed by categories.

Category	Variable	Feature	Type
Characteristics of borrowers	F_1	Age	Ordinal
	F_2	Educational level	Ordinal
	F_3	Marital status	Binary
	F_4	Length of employment	Ordinal
	F_5	Place of employment	Nominal
	F_6	Industry	Nominal
	F_7	Size	Ordinal
Financial information of borrowers	F_8	Monthly income	Ordinal
	F_9	House	Binary
	F_{10}	Car	Binary
	F_{11}	Mortgage	Binary
	F_{12}	Auto loan	Binary
Credit history of borrowers	F_{13}	Successful loans	Numeric
	F_{14}	Applications	Numeric
Characteristics of loans	F_{15}	Amount	Numeric
	F_{16}	Purpose	Nominal
	F_{17}	Term	Nominal
	F_{18}	Type	Nominal
Platform authentication information	F_{19}	Credit report authentication	Binary
	F_{20}	Identity authentication	Binary
	F_{21}	Job authentication	Binary
	F_{22}	Income authentication	Binary

tication and identity authentication, but fewer of them provide job authentication and income authentication.

4.1.2 Preliminary data screening

To better understand the data set, we describe the default rates with features from the different feature categories.

Default rate with age. As shown in Figure 2(a), the default rate significantly decreases as age increases, which means that on this P2P lending platform, younger borrowers, especially those between 20 and 25 years old, are more likely to default than older borrowers.

Default rate with monthly income. As shown in Figure 2(b), the default rate pattern is similar to that in Figure 2(a), and the default rate decreases as monthly income

TABLE 3 Summary statistics of the numerical and binary variables.

Feature	Mean	Std	Max	Min
Marital status(F_3)	0.72			
House (F_9)	0.61			
Car(F_{10})	0.94			
Mortgage(F_{11})	0.60			
Auto loan (F_{12})	0.24			
Successful loans (F_{13})	1.04	0.277	20	0
Applications (F_{14})	1.10	0.588	22	0
Amount (F_{15})	35 336.60	17 133.320	425 971	1000
Credit authentication (F_{19})	0.99			
Identity authentication (F_{20})	0.99			
Job authentication (F_{21})	0.95			
Income authentication (F_{22})	0.95			

increases, which means that borrowers who earn a lower monthly income are more likely to default.

Default rate with term. As shown in Figure 2(c), the default rate first shows an upward trend as the term increases and then levels off. That indicates that the longer the borrowing period, the easier it is to default.

Default rate with income authentication. As shown in Table 4, the default rate of "yes" is significantly lower than that of "no", which means that borrowers who did not provide income authentication are more likely to default.

4.1.3 Data preprocessing

Data preprocessing is a crucial process to prepare data for model training, and in this paper, data preprocessing involves the following two steps.

Standardized treatment. Nominal features are assigned numeric values, and numeric features are normalized into [0,1].

Data partition. The original data set is divided into three parts in the proportion of 60%:20%:20%, ie, 60% of the observations are for the training set, 20% of the observations are for the validation set, and 20% of the observations are for the testing set. Here, we keep the testing set apart and perform 10-fold cross-validation to obtain the training set and the validation set.

FIGURE 2 Default rates calculated based on (a) age, (b) monthly income and (c) term.

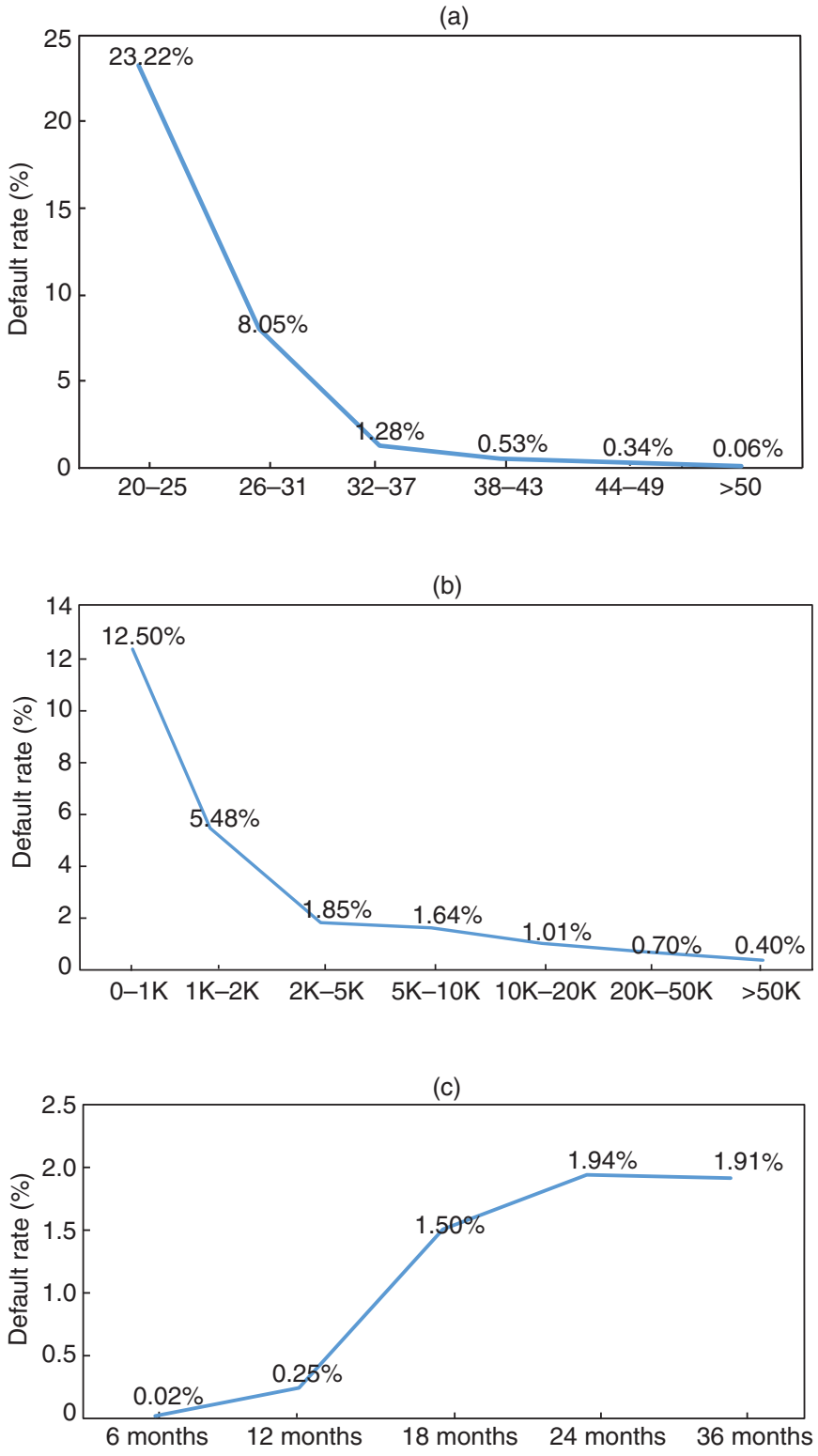


TABLE 4 Default rates calculated based on income authentication.

Income authentication	Default rate (%)
Yes	0.72
No	14.31

4.2 Phase II: identification of important features

The end goal of this step is to find a subset of features that offers the highest discriminative power for the next step. As demonstrated in Section 2.1, there is a trend of developing a hybrid feature selection method to realize an advantageous combination. Here, we combine the three feature selection algorithms listed in Section 3.1. The combination scheme is as follows: a feature F_i receives three scores $\{S(F_i{}^{\text{fs1}}), S(F_i{}^{\text{fs2}}), S(F_i{}^{\text{fs3}})\}$ based on the three algorithms fs1, fs2 and fs3; then, the final score of each feature $SF\{F_i\}$ is computed by an average voting mechanism as follows:

$$SF\{F_i\} = \frac{S(F_i{}^{fs1}) + S(F_i{}^{\text{fs2}}) + S(F_i{}^{\text{fs3}})}{3}. \tag{4.1}$$

Subsequently, the 10 features with the highest final scores are selected as the final results; the features are listed in Table 5. The reasons why we choose 10 features are as follows. First, when we calculate the 10-fold cross-validation F-measure[4] of the top k features ranked by the information gain ratio algorithm with the SVM classifier, some features will likely be noise, and the optimal feature number is 10 since it gives the best performance (see Figure 3). Second, the number of features selected by the GA approach is 10. Third, when comparing the top 10 selected features of the three methods (see Table 6), there are no significant differences. Moreover, when comparing the performance of the three single feature selection approaches with our combination approach using the SVM classifier, the results shown in Table 6 clearly verify the superiority of our combined approach. Thus, the top 10 features are used in the following sections.

4.3 Phase III: SMOTE treatment

A data set is said to be imbalanced if the number of samples from one class is higher than that from another. The class with more instances is called the major class, while the class with fewer instances is called the minor class (Chawla *et al* 2004; Longadge and Dongre 2013). As illustrated in Phase I, the number of "good" loans (major class)

[4] The reason we choose the F-measure as the evaluation metric rather than accuracy is that the data set is highly imbalanced, as illustrated in Section 4.3.

FIGURE 3 Classification performance of the information gain ratio feature selection method.

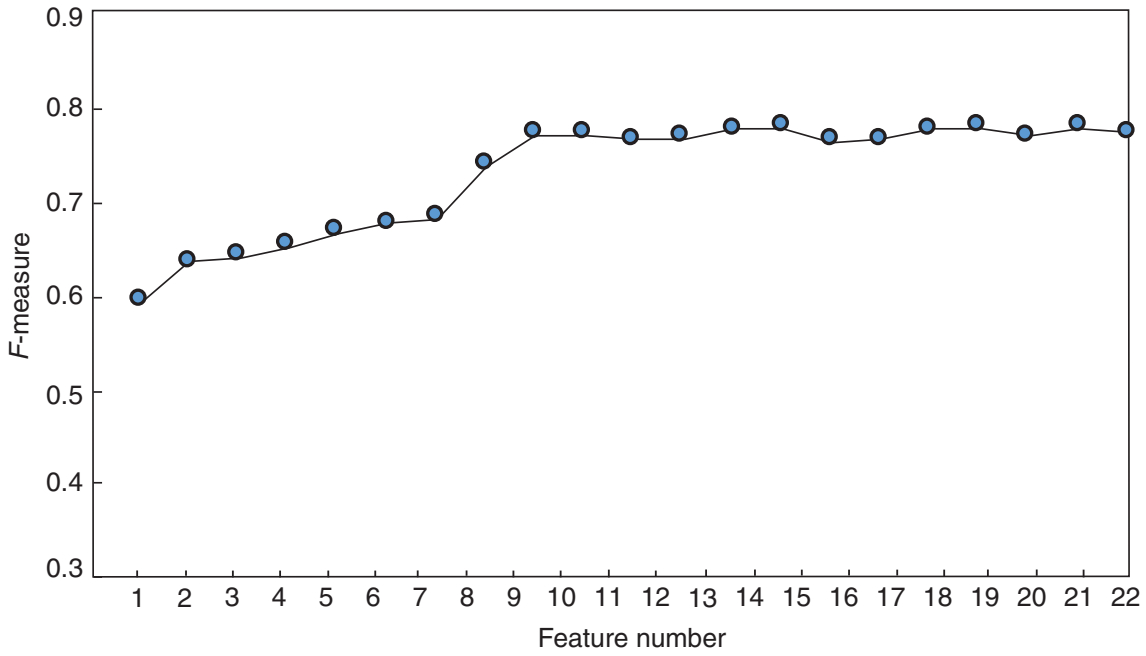


TABLE 5 Selected features.

Category	Feature	Meaning
Characteristics of borrowers	F_1	Age
	F_3	Marital status
Financial information of borrowers	F_8	Monthly income
Credit history of borrowers	F_{13}	Successful loans
	F_{14}	Applications
Characteristics of loans	F_{15}	Amount
	F_{17}	Term
	F_{18}	Type
Platform authentication information	F_{21}	Job authentication
	F_{22}	Income authentication

in our data set is 33 492, which is much higher than the number of "bad" loans (minor class), which is 474. Thus, it is easy to obtain that the imbalance rate is 70.66, which means that the data set is highly skewed.

TABLE 6 Performance of various feature selection algorithms.

Feature selection algorithm	F-measure	Selected features
Information gain ratio	0.769	F_1, F_{15}, F_{22}, F_3, F_8, F_{21}, F_{17}, F_{10}, F_{13}, F_{12}
GA	0.775	F_{22}, F_1, F_{14}, F_{17}, F_{21}, F_7, F_{15}, F_{18}, F_{13}
CART	0.772	F_3, F_{14}, F_1, F_{15}, F_{22}, F_{16}, F_{21}, F_9, F_{17}, F_{18}
Top 10 features	0.796	See Table 5

As illustrated by many researchers (see, for example, Shen *et al* 2019), an uneven distribution of samples will lead to skewed results due to data bias toward the major class, and this issue challenges existing models. Moreover, a highly skewed data set can lead to misleading results for some typical evaluation metrics. When facing a skewed data set, it is easy to understand that the accuracy metric will obtain biased results. Even receiver operator characteristic (ROC) curves, which are often employed to evaluate imbalanced data sets, can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution (Davis and Goadrich 2006).

Therefore, when facing the highly skewed data set in this study, preprocessing the imbalanced data set before modeling is highly important. As a simple and effective oversampling method, the SMOTE algorithm proposed by Chawla *et al* (2002) is employed in this study to handle the imbalance issue. Based on this algorithm, there exists a virtual sample between two real samples that are near one another. Therefore, in this study, the SMOTE algorithm artificially invents a new "bad" loan between the two real "bad" loans that are near each other based on the following two steps.

STEP 1 Given the unbalanced training data set D, $C_{\text{bad}} \subset T$ represents the minority class ("bad" loans), and for each sample of $x_i \in C_{\text{bad}}$, the k nearest neighbors ($\{x_{i1}, \ldots, x_{ib}, \ldots, x_{ik}\}$) are identified. In this paper, k is set to a default value 5 based on previous studies (Maldonado *et al* 2019).

STEP 2 A new "bad" loan x_{ib}^{new} around the original sample $x_i \in C_{\text{bad}}$ can be artificially invented with the randomly selected neighbor x_{ib} based on the following formula:

$$x_{ib}^{\text{new}} = x_i + (x_{ib} - x_i) \times \delta, \tag{4.2}$$

where δ denotes the random value in (0, 1).

In this manner, the procedure for synthesizing the instance of the minority class is repeated until the training data set is balanced. After the SMOTE treatment, there are 46 888 loans in the final training data set, where there are 23 444 instances of both "good" and "bad" loans.

4.4 Phase IV: parameter tuning and modeling

As described in Section 3.2 and Figure 1, six ensembles are tested in this paper – the bagging, boosting, random subspace, rotation forest, random forest and GBDT methods – with the following five popular base classifiers: LR, CART, C4.5, multilayer perceptron (MLP) neural networks with a back-propagation algorithm and SVMs with a linear kernel. The bagging, boosting, random subspace and rotation forest methods are implemented with the Waikato environment for knowledge analysis (WEKA) open-source data mining toolkit (Witten and Frank 2002; Witten *et al* 2011), while the random forest and GBDT methods are performed with the Sklearn package in Python 3.6.

In practice, to improve the forecasting performance of these models, a few parameters need to be optimized before classifier construction. In this study, the grid search method is used to determine the optimal parameters that generate minimum forecasting errors. The optimized results are listed in Table 7.

For base classifiers, the detailed parameter tuning processes are as follows.

CART and C4.5 algorithms. The test range of a minimal number of samples at the terminal leaf is from 2 to 10. In addition, the number of features is set to 10.

MLP method. A grid search is carried out to find the optimal number of hidden layers and the learning rate. The search ranges are [1,10] and [0.001,1], respectively.

SVMs. Various cost hyperparameters C and γ are evaluated. Moreover, different kernel types are tested, and the candidates are linear, polynomial, radial basis function and sigmoid.

LR and MLP. The cutoff values are set at a default value of 0.5.

For ensembles, the detailed parameter tuning processes are as follows.

Random forest method. Different "number of trees" and "number of attributes", used to grow each tree, are tested. The "number of trees" is adjusted from 10 to 100 with nine steps and the "number of attributes" is tested from 1 to 10.

GBDT method. Different "number of iterations", "learning rate" and "number of attributes" are tested. The "number of iterations" is tuned from 10 to 100 with nine steps, the "learning rate" is adjusted in the range [0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 1] and the "number of attributes" is tested from 1 to 10.

Bagging, boosting, random subspace and rotation forest methods. The default settings in the Weka package are adopted (Abellán and Castellano 2016; Marqués *et al* 2012).

TABLE 7 Main parameter settings of the base classifiers and ensembles.

Classification algorithms		Parameter settings
Base classifiers classifiers	LR	Ridge value in the log-likelihood: 1.0E–8
	CART	Minimal number of samples at the terminal leaf: 3; number of features: 10
	C4.5	Minimal number of samples at the terminal leaf: 2; number of features: 10
	MLP	Number of hidden layers: 3; learning rate: 0.01; learning algorithm: feed-forward
	SVM	Optimal pair of (C, γ) is (1, 0.1)
Ensembles	Bagging	Size of each bag: 100; batchsize: 100; number of iterations: 10; seed = 1
	Boosting	Batchsize: 100; number of iterations: 10; seed = 1
	Random subspace	Batchsize: 100; number of iterations: 10; seed = 1; subspacesize = 0.5
	Rotation forest	Batchsize: 100; number of iterations: 10; seed=1; projection filter: principal components analysis
	Random forest	Number of trees: 80; number of attributes: 10
	GBDT	Number of iterations: 70; learning rate: 0.1; number of attributes: 10

4.5 Phase V: comparison and evaluation

To evaluate the performance of various methods, two kinds of evaluation procedures are employed. The first involves standard evaluation metrics based on the confusion

TABLE 8 Confusion matrix.

	Actual "bad" loans	**Actual "good" loans**
Predicted as "bad" loans	True positive (TP)	False positive (FP)
Predicted as "good" loans	False negative (FN)	True negative (TN)

matrix described in Table 8; here, the "bad" samples are chosen as the positive class.[5] The corresponding metrics are as follows.

- Accuracy = $(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN})$, which represents the correctly discriminative samples, including "bad" and "good" loans.
- Type I error = $\mathrm{FN}/(\mathrm{TP}+\mathrm{FN})$, which is the proportion of actual "bad" loans predicted to be "good". Given its high probability of causing high costs, we must pay attention to this metric.
- F-measure = $2\mathrm{TP}/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$, which represents a type of harmonic mean for precision and recall, where precision is the fraction of true positive samples among the samples that the model classified as positive ($\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$), and where recall is the fraction of samples classified as positive among the total number of positive samples ($\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$). As illustrated in Soleymani *et al* (2020), widely used evaluation metrics, such as accuracy and the area under the ROC curve (AUC), tend to favor the correct classification of the most populated class, which will, together with imbalanced test data, cause biased high scores. To address this issue, in this paper we add the F-measure evaluation metric, which pays more attention to the minority class. For details, see Thomas *et al* (2017).

To compare the performance of various ensembles from a statistical perspective, the Friedman test (Demšar 2006; Friedman 1940), which is a rank-based nonparametric test, is employed to test the statistical significance of differences in performance. The Friedman test is distributed based on the chi-square distribution with $m-1$ degrees of freedom, where m is the number of methods (in this study, ensemble methods). The corresponding statistic is computed as follows:

$$\chi_F^2 = \frac{12n}{m(m+1)}\left[\Sigma_i(\Sigma_j(r_j^i))^2 - \frac{m(m+1)^2}{4}\right], \tag{4.3}$$

[5] As illustrated in Tan *et al* (2005), the minority class can be chosen as the positive class due to its high importance.

where r_j^i denotes the individual rank of each method i on each data set j, and h is the number of data sets (according to Section 5, there are three different data sets: one original data set and two robust testing data sets, which means that in this study $h = 3$). If χ_F^2 is larger than a critical value, the null hypothesis that all the methods are equivalent is rejected. Then, a post hoc test, specifically the Bonferroni–Dunn test (Dunn 1961), is employed to compare all the ensembles (eg, i) with a control ensemble k. The statistic for comparing methods i and k is computed as follows.

$$\mathrm{CD} = q_{p,\infty,m}\sqrt{\frac{m(m+1)}{12h}}, \tag{4.4}$$

where $q_{p,\infty,m}$ is computed based on the t-test statistic with a confidence level $p/(m-1)$. If the average ranks of methods i and k differ by at least the critical difference (CD), there is a significant difference in the performance of methods i and k.

5 EXPERIMENTAL RESULTS

The purpose of this study is to find useful features that can discriminate "bad" and "good" loans in a valid way and also to compare the performance of six ensembles in credit scoring for a Chinese P2P lending platform. To achieve this goal, we put forward four questions.

(1) Which ensemble performs best with Chinese P2P platform data?

(2) Can the results be trusted?

(3) Does the performance have statistical significance?

(4) Which feature impacts the results the most?

Hereafter, the questions are discussed in detail.

5.1 Performance of the six ensembles

The results of the discriminative performance of the six ensembles[6] on the testing set with full features and selected features as measured by accuracy, Type I error and F-measure are listed in Table 9 and Figure 4. Table 9 lists in detail the results of different ensembles with corresponding base classifiers, and Figure 4 shows a summary of the average performance[7] of the three measures. Importantly, we also compare the performance of ensembles and the grade feature provided by the lending

[6] Here we delete the bagging results with the LR and SVM base classifiers because the two strong base classifiers are not suitable for bagging, which focuses on reducing variance.

[7] Here the average refers to the mathematical average performance of different base classifiers by the bagging, boosting, random subspace and rotation forest methods.

platform. According to the platform, this feature is generated in the loan application process by evaluating the application documents, and the grade can be divided into seven levels (from low to high): HR, E, D, C, B, A and AA. Here, we select the lowest grade, HR, as the reference for "bad" loans and select the others as "good" loans; the results are listed in Figure 4. Moreover, to discover the effect of feature selection, we also compare the performance of all features (see Table 2) and the top 10 selected features (see Table 5).

As shown in Table 9 and Figure 4, we obtain the following findings. First, it is easy to see that the performances of the six ensembles are at different grades. The first grade includes the GBDT, random forest and rotation forest methods, which are powerful ensembles and hold the top three spots in terms of the three different evaluations. For example, as illustrated in Figure 4, with the top 10 features, the GBDT method achieves the highest F-measure (approximately 0.94), which is approximately 13% better than that achieved by the random subspace method (approximately 0.8). The second grade includes the bagging and boosting ensembles, which have similar performance. The third grade consists of the random subspace ensemble, which achieves the worst performance in terms of the three evaluation metrics. The detailed information listed in Table 9 shows that compared with other ensembles, the random subspace ensemble has the largest Type I error, which may be the key reason for the poor performance of random subspace.

Second, the grade feature has better performance than the random subspace ensemble in terms of different evaluation metrics and different features. Compared with the bagging and boosting ensembles, interestingly the grade feature achieves a better Type I error score but worse accuracy and F-measure scores. The reason behind this result may be that to avoid risk, the platform tends to give low grades.

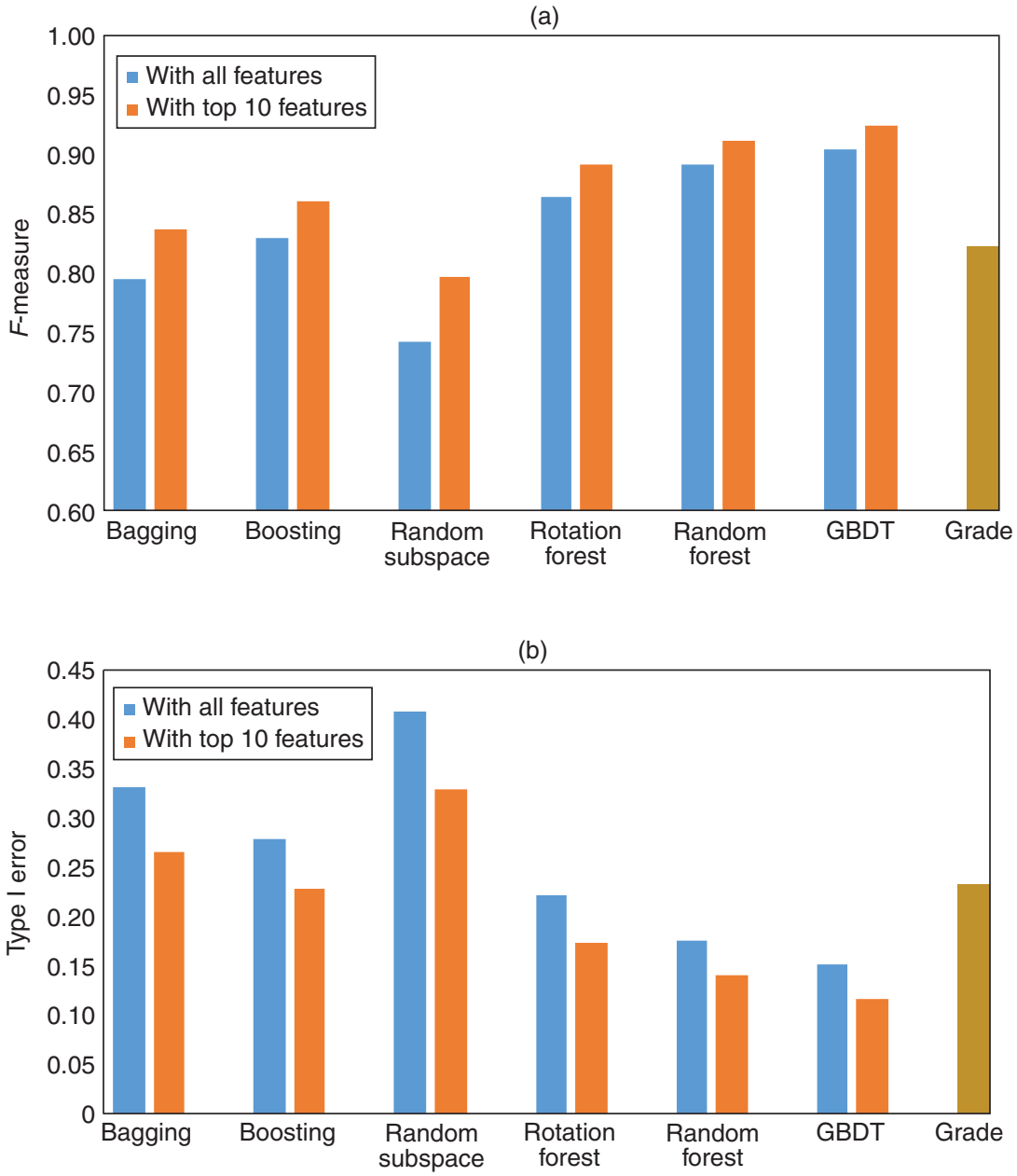
Third, each ensemble with the 10 selected features yields better performance than the ensemble with all 22 features, whether measured by detailed evaluations with base classifiers or average comparisons. This finding clearly indicates the effectiveness of our combined feature selection approach and the usefulness of the top 10 selected features.

Finally, regarding the Type I error metric, interestingly, the ensembles with the selected features can greatly reduce this error. For instance, Table 9 shows that even random subspace, the worst performing ensemble, combined with the C4.5 base classifier, improves performance by 9% on the F-measure with the selected features compared with performance that includes all features. In conclusion, the selected features show that they are capable of discriminating successfully between "bad" loans and "good" loans.

TABLE 9 Performance of six ensembles under three evaluation metrics.

Ensemble	Base classifiers	Accuracy (%)		Type I error (%)		AUC	
		With all features	With top 10 features	With all features	With top 10 features	With all features	With top 10 features
Bagging	CART	0.8333	0.8678	0.3333	0.2644	0.7832	0.8477
	C4.5	0.8563	0.8966	0.2874	0.2069	0.8322	0.8846
	MLP	0.8218	0.8413	0.3563	0.3095	0.8	0.8131
	LR	0.8492	0.8506	0.2989	0.2759	0.8243	0.84
Boosting	CART	0.8736	0.9195	0.2529	0.1609	0.8553	0.9125
	C4.5	0.8563	0.8736	0.2874	0.2529	0.8322	0.8553
	MLP	0.873	0.8889	0.254	0.2222	0.8545	0.875
	SVM	0.8621	0.8966	0.2759	0.2069	0.84	0.8846
	LR	0.7874	0.8175	0.4253	0.3651	0.7299	0.7767
Random subspace	CART	0.7976	0.8448	0.4048	0.3103	0.7463	0.8163
	C4.5	0.7897	0.8492	0.4206	0.3016	0.7337	0.8224
	MLP	0.8046	0.8294	0.3908	0.3333	0.7571	0.7962
	SVM	0.8161	0.8448	0.3678	0.3103	0.7746	0.8163
	LR	0.8793	0.8908	0.2299	0.2184	0.8645	0.8774
Rotation forest	CART	0.8966	0.9138	0.2069	0.1724	0.8846	0.9057
	C4.5	0.8851	0.9195	0.2299	0.1609	0.8701	0.9125
	MLP	0.8966	0.9195	0.2069	0.1494	0.8846	0.9136
	SVM	0.8908	0.9253	0.2184	0.1494	0.8774	0.9193
Random forest		0.9138	0.931	0.1724	0.138	0.9057	0.9259
GBDT		0.9253	0.9425	0.1494	0.1149	0.9193	0.939

FIGURE 4 Average performance of ensembles considering different measures.



(a) F-measure. (b) Type I error.

5.2 Robustness tests

To evaluate the effectiveness and robustness of the performance results listed above, we further compare the six ensembles with out-of-time instances. Moreover, we list the performance results with a different data partition ratio: 50% for training, 20%

for validation and 30% for testing. The results are listed in the online supplementary materials.

Here, the out-of-time loan data are collected from January 2017 to June 2019, and the loans that appear in the former data set (2013–16) are deleted from this data set. The final number of instances (considering only "fully paid" and "charge off") is 39 201, including 607 "bad" loans. To maintain consistency with the former comparison, the data are divided into three parts: 60% for training, 20% for validation and 20% for testing.

Table 10 and Figure 5 show the detailed and average performance of the six ensembles and the grade feature on the testing set. It is easy to see that the results are consistent with those of the previous experiment. For instance, Table 10 clearly shows that the GBDT method performs best, followed by the random forest and rotation forest methods, while the random subspace method still has the worst performance. Moreover, the grade feature has better performance than the random subspace ensemble. Interestingly, although it still performs worse than the other ensembles, it has better scores than those it obtained with the former data set listed in Section 5.1, which may imply that the P2P platform has achieved progress in identifying "bad" loans. In addition, Figure 5 clearly shows that the ensembles with the selected features outperform those with all features. Accordingly, we conclude that the comparison results are robust to different time periods.

5.3 Statistical tests

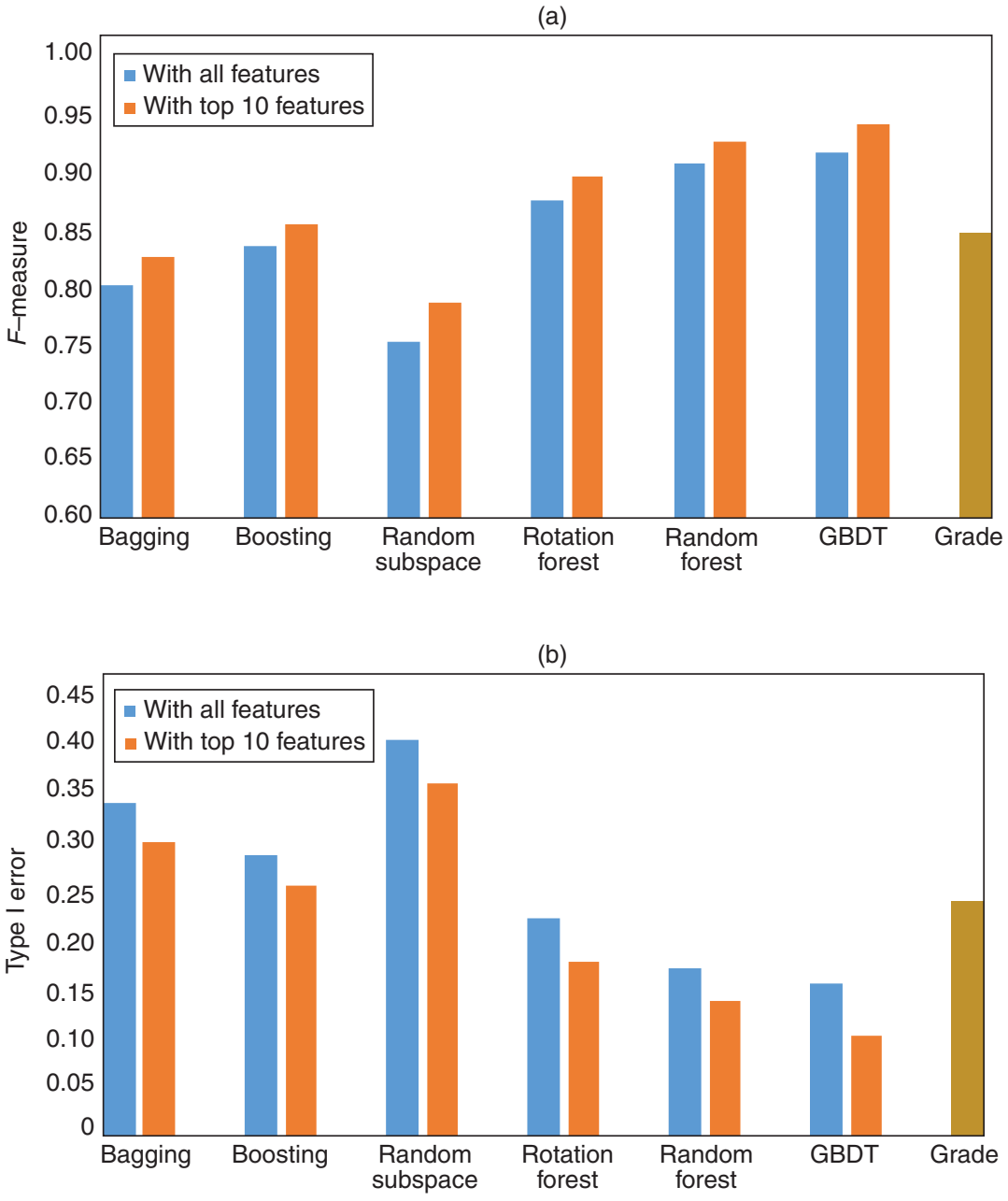
The nonparametric Friedman test and the post hoc (Bonferroni–Dunn) test were employed to test the statistical significance of the differences in performance. In this section, we also consider the values of Friedman's ranks for the accuracy and AUC measures. Performance with two different training, validation and testing ratios (ie, 60:20:20, 50:20:30) and out-of-time data sets are used to conduct the tests. The rank values represent an important reference for performance when several ensembles are compared and can be considered in order of their performance.

First, the Friedman test statistic is computed to verify whether the ensembles exhibit different prediction performances. The results are 35.59 and 35.47 on the two measures, both rejecting the null hypothesis that the ensembles exhibit no different performances at the significance level of 99.5%. Then, the Bonferroni–Dunn test is applied to conduct comparisons. The CDs at significance levels of 0.05 and 0.1 are computed as 10.532 and 9.781, and the average ranks of various ensembles and the grade feature are shown in Figure 6. Thus, the following findings can be obtained. First, the GBDT, random forest and rotation forest ensembles are generally the best. More specifically, the GBDT method with the top 10 features has the lowest rank in terms of two different metrics. Second, the grade feature holds the middle position

TABLE 10 Performance of six ensembles under three evaluation metrics.

Ensemble	Base classifiers	Accuracy (%)		Type I error (%)		AUC	
		With all features	With top 10 features	With all features	With top 10 features	With all features	With top 10 features
Bagging	CART	0.8430	0.8636	0.3058	0.2645	0.8155	0.8436
	C4.5	0.8388	0.8512	0.3223	0.2975	0.8079	0.8252
	MLP	0.8223	0.8430	0.3554	0.3058	0.7839	0.8155
	LR	0.8430	0.8554	0.3141	0.2893	0.8137	0.8309
Boosting	CART	0.8636	0.8802	0.2727	0.2397	0.8421	0.8639
	C4.5	0.8595	0.8802	0.281	0.2314	0.8365	0.8651
	MLP	0.8678	0.8843	0.2479	0.2314	0.8505	0.8682
	SVM	0.8595	0.8719	0.2645	0.2397	0.8396	0.8558
	LR	0.7686	0.8141	0.438	0.3719	0.7083	0.7716
Random subspace	CART	0.8182	0.8306	0.3554	0.3306	0.7800	0.798
	C4.5	0.8017	0.8182	0.3967	0.3471	0.7526	0.7822
	MLP	0.8099	0.8264	0.3719	0.3471	0.7677	0.7900
	SVM	0.8017	0.8264	0.3802	0.3388	0.7576	0.7921
	LR	0.876	0.8884	0.2397	0.2066	0.8598	0.8767
Rotation forest	CART	0.8967	0.9174	0.1983	0.157	0.8858	0.9107
	C4.5	0.8884	0.8926	0.2231	0.1901	0.8744	0.8829
	MLP	0.8884	0.9091	0.2149	0.1736	0.8756	0.9009
	SVM	0.9008	0.9256	0.1901	0.1322	0.8909	0.9211
Random forest		0.9174	0.9339	0.1653	0.1322	0.9099	0.9292
GBDT		0.9256	0.9463	0.1488	0.0992	0.9196	0.9437

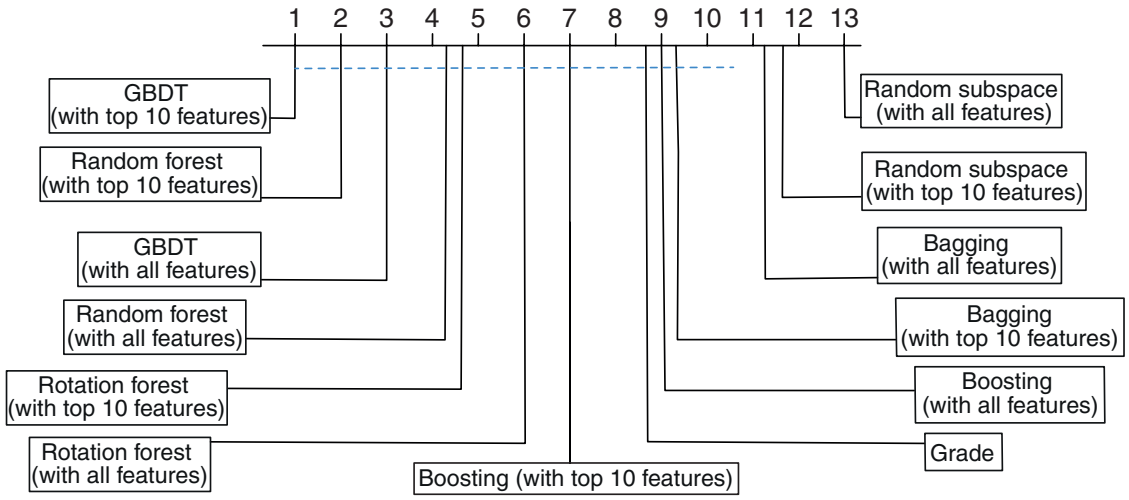
FIGURE 5 Average performance of ensembles considering different measures (out-of-time samples).



(a) F-measure. (b) Type I error.

among all of the ensembles. As shown in Figure 6, the average rank of the grade is 8, which means that it cannot beat the GBDT, random forest and rotation forest ensembles. Third, regarding the same ensemble with different numbers of features, it is easy to conclude that compared with the method with all features, the method

FIGURE 6 Friedman average ranking and post hoc test results for the F-measure.



Dashed blue line shows CD = 9.781, $\alpha = 0.1$.

with the top 10 selected features has the lower rank. Fourth, as shown in Figure 6, when the GBDT method with the top 10 features is used as the benchmark, the random subspace and bagging methods have average ranks that are higher than the CD, suggesting that in terms of accuracy, they are significantly worse than the GBDT approach at a significance level of 0.1. Additionally, this pattern is reproduced with the accuracy listed in the online supplementary materials, which further indicates that our results are robust.

In summary, the statistical tests also support the finding that the GBDT, random forest and rotation forest methods are generally the best for the Chinese P2P lending service since they outperform the other ensembles and the grade feature on various measures, different training-to-testing ratios and out-of-time samples. Moreover, the experimental results prove that the top 10 selected features can improve credit scoring performance, which also proves the effectiveness of the combined feature selection process in our proposed system.

5.4 Factor analysis

Since the evaluation test results and statistical tests demonstrate the effectiveness of the selected features, we conduct a detailed analysis of the factors. Table 11 lists the importance of the features selected by the GBDT ensemble model. By comparing the contributions of the features, the following insights can be obtained.

(1) As illustrated in Table 12, the contributions of the five categories of features in the original data set from high to low are the characteristics of loans, the

TABLE 11 Ranking results of the top 10 features (original samples).

Ranking	Features	Categories
1	F_{15} Amount	Characteristics of loans
2	F_{17} Term	Characteristics of loans
3	F_8 Monthly income	Financial information of borrowers
4	F_{18} Type	Characteristics of loans
5	F_1 Age	Characteristics of borrowers
6	F_{14} Applications	Credit history of borrowers
7	F_{13} Successful loans	Credit history of borrowers
8	F_{22} Income authentication	Platform authentication information
9	F_{21} Job authentication	Platform authentication information
10	F_3 Marital status	Characteristics of borrowers

TABLE 12 Ranking results of the top 10 features (out-of-time samples).

Ranking	Features	Categories
1	F_{15} Amount	Characteristics of loans
2	F_{17} Term	Characteristics of loans
3	F_{22} Income authentication	Platform authentication information
4	F_{21} Job authentication	Platform authentication information
5	F_8 Monthly income	Financial information of borrowers
6	F_1 Age	Characteristics of borrowers
7	F_{18} Type	Characteristics of loans
8	F_{14} Application	Credit history of borrowers
9	F_{13} Successful loans	Credit history of borrowers
10	F_3 Marital status	Characteristics of borrowers

financial information of borrowers, the credit history of borrowers, the characteristics of borrowers and platform authentication information. These results indicate that the characteristics of loans have the greatest impact. Interestingly, one of the characteristics of borrowers, F_1 (age), has a competitive ranking (rank 5), while another feature, F_3 (marital status), obtains the lowest rank, which indicates that age contributes more than the marital status. The reason may be that in China, age represents a borrower's economic strength, and a younger age may increase the default risk, as illustrated in Section 4.1.2.

(2) As depicted in Table 12, the ranks of the five categories of features in the 2017–19 data set (from high to low) are the characteristics of loans, platform authentication information, the financial information of borrowers, the credit

history of borrowers and the characteristics of borrowers. It is easy to see that the rank of platform authentication information increases with the rank in original data set, possibly because it was easier to obtain loans at an early time (2013–16). Thus, some "good" borrowers do not provide effective authentication information. However, over time the platform authentication information, as a crucial indicator of repayment ability, becomes increasingly important. Thus, borrowers who do not provide effective authentication information are more likely to be "bad" borrowers.

(3) Comparing the findings listed above with the similar impact factor analysis in Ma *et al* (2018), which focuses on US P2P lending services, we can make the following interesting remarks. First, the high proportions of features holding high ranks in the characteristics of loans subset (similar to the borrowing details in Ma *et al* (2018)) are similar in the two platforms, which means that loan details are the most relevant information for the final loan result. Second, the two platforms differ in that there fewer financial information features (similarly to the financial situation in Ma *et al* (2018)) but more authentication information features in the Chinese P2P platform than the US one. The reason behind this may be that authentication information features are the benchmark for credit authentication; also, the credit authentication system in China is not as mature as that in the United States. Overall, the platform needs more corresponding features to capture the borrower's credit situation.

6 CONCLUSION

As pointed out in the previous sections, there is an urgent need to improve the credit scoring method in Chinese P2P lending services to reduce risk. The future of credit risk analysis has an increasing reliance on machine learning models as opposed to human decision-making methods. Although credit scoring has attracted much attention, two questions have not been answered. First, what is the optimal set of features for the Chinese P2P credit scoring problem? Second, which ensemble method is the most appropriate for Chinese P2P services?

In this paper, a hybrid system of feature selection algorithms and ensemble methods was built to answer these questions. In the proposed model, three feature selection algorithms were first combined to obtain suitable features with higher discriminative power. After choosing the top 10 features, six popular ensemble methods with five base classifiers were used to conduct the credit scoring task. For the experiments, 33 906 samples from the Chinese Renren lending platform were collected, of which 474 samples were "bad" loans. Different robustness tests were conducted to evaluate the effectiveness and robustness of our system's performance. In addition, three

standard evaluation metrics (accuracy, Type I error and the F-measure) and nonparametric statistical tests were employed to compare the performance of the ensembles. The experimental results show that the selected features can greatly improve prediction performance. Moreover, the comparisons show the superiority of the GBDT, random forest and rotation forest methods in standard evaluations, robustness tests and statistical tests. We believe that our findings have practical value for creditors, investors and risk managers. Future work can build upon this study by examining the specific contributions of features. Additionally, the use of data sets from other P2P lending platforms would be interesting to consider.

DECLARATION OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their constructive comments. This work was supported by the Natural Science Foundation of China under Projects 71801072 and 71774047.

REFERENCES

Abdou, H. A., and Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management* **18**(2–3), 59–88 (https://doi.org/10.1002/isaf.325).

Abellán, J., and Castellano, J. G. (2016). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications* **73**, 1–10 (https://doi.org/10.1016/j.eswa.2016.12.020).

Ala'raj, M., and Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems with Applications* **64**, 1–20 (https://doi.org/10.1016/j.eswa.2016.07.017).

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* **23**(4), 589–609 (https://doi.org/10.2307/2978933).

Bellotti, T., and Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications* **36**(2), 3302–3308 (https://doi.org/10.1016/j.eswa.2008.01.005).

Bian, S., and Wang, W. (2007). On diversity and accuracy of homogeneous and heterogeneous ensembles. *International Journal of Hybrid Intelligent Systems* **4**(2), 103–128 (https://doi.org/10.3233/his-2007-4204).

Bijak, K., and Thomas, L. C. (2012). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications* **39**(3), 2433–2442 (https://doi.org/10.1016/j.eswa.2011.08.093).

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123-–140 (https://doi.org/10.1007/bf00058655).

Breiman, L. (2001). Random forests. *Machine Learning* **45**(1), 5–32 (https://doi.org/10.1007/0-387-21529-8-16).

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (https://doi.org/10.1613/jair.953).

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter* **6**(1), 1–6 (https://doi.org/10.1145/1007730.1007733).

Chen, F. L., and Li, F. C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications* **37**(7), 4902–4909 (https://doi.org/10.1016/j.eswa.2009.12.025).

Davis, J., and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240. Association for Computing Machinery, New York (https://doi.org/10.1145/1143844.1143874).

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* **7**, 1–30 (https://doi.org/10.2172/881587).

Dunn, O. J. (1961). Multiple comparisons among means. *Publications of the American Statistical Association* **56**(293), 52–64 (https://doi.org/10.2307/2282330).

Feng, X., Xiao, Z., Zhong, B., Qiu, J., and Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing* **65**, 139–151 (https://doi.org/10.1016/j.asoc.2018.01.021).

Freund, Y., and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, 148–156 (https://doi.org/10.1006/jcss.1997.1504).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**(5), 1189–1232 (https://doi.org/10.1016/s0167-9473(01)00065-2).

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics* **11**(1), 86–92 (https://doi.org/10.1214/aoms/1177731944).

Gao, Y., Yu, S.-H., and Shiue, Y.-C. (2018). The performance of the P2P finance industry in China. *Electronic Commerce Research and Applications* **30**, 138–148 (https://doi.org/10.1016/j.elerap.2018.06.002).

García, V., Marqués, A. I., and Sánchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion* **47**, 88–101 (https://doi.org/10.1016/j.inffus.2018.07.004).

Goldberg, D. E. (2006). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Boston, MA (https://doi.org/10.5860/choice.27-0936).

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**(6), 1157–1182 (https://doi.org/10.1162/P153244303322753616).

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications* **42**(2), 741–750 (https://doi.org/10.1016/j.eswa.2014.08.029).

He, H., Zhang, W., and Zhang, S. (2018). A novel ensemble method for credit scoring: adaption of different imbalance ratios. *Expert Systems with Applications* **98**, 105–117 (https://doi.org/10.1016/j.eswa.2018.01.012).

Hens, A. B., and Tiwari, M. K. (2012). Computational time reduction for credit scoring: an integrated approach based on support vector machine and stratified sampling method. *Expert Systems with Applications* **39**(8), 6774–6781 (https://doi.org/10.1016/j.eswa.2011.12.057).

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8), 832–844 (https://doi.org/10.1109/34.709601).

Hsieh, N. C., Hung, L. P., and Ho, C. L. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications* **37**(1), 534–545 (https://doi.org/10.1007/978-3-642-01307-2-33).

Hu, Y. C., and Ansell, J. (2007). Measuring retail company performance using credit scoring techniques. *European Journal of Operational Research* **183**(3), 1595–1606 (https://doi.org/10.1016/j.ejor.2006.09.101).

Hua, Z., Wang, Y., Xu, X., Zhang, B., and Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications* **33**(2), 434–440 (https://doi.org/10.1016/j.eswa.2006.05.006).

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), 847–856 (https://doi.org/10.1016/j.eswa.2006.07.007).

Jiménez, F., Jódar, R., Martín, M. d. P., Sánchez, G., and Sciavicco, G. (2016). Unsupervised feature selection for interpretable classification in behavioral assessment of children. *Expert Systems* **34**(4), e12173 (https://doi.org/10.1111/exsy.12173).

Koutanaei, F. N., Sajedi, H., and Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services* **27**, 11–23 (https://doi.org/10.1016/j.jretconser.2015.07.003).

Kumar, V. (2014). Feature selection: a literature review. *Smart Computing Review* **4**(3), 1–19 (https://doi.org/10.6029/smartcr.2014.03.007).

Lee, T. S., and Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications* **28**(4), 743–752 (https://doi.org/10.1016/j.eswa.2004.12.031).

Lessmann, S., Baesens, B., Seow, H. V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *European Journal of Operational Research* **247**(1), 124–136 (https://doi.org/10.1016/j.ejor.2015.05.030).

Longadge, R., and Dongre, S. (2013). Class imbalance problem in data mining review. *International Journal of Computer Science and Network* **2**(1), 83–87.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., and Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications* **31**, 24–39 (https://doi.org/10.1016/j.elerap.2018.08.002).

Maldonado, S., López, J., and Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing* **76**, 380–389 (https://doi.org/10.1016/j.asoc.2018.12.024).

Malhotra, R., and Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research* **136**(1), 190–211 (https://doi.org/10.1016/s0377-2217(01)00052-2).

Marqués, A. I., García, V., and Sáchez, J. S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications* **39**(11), 10 244–10 250 (https://doi.org/10.1016/j.eswa.2012.02.092).

Miao, J., and Niu, L. (2016). A survey on feature selection. *Procedia Computer Science* **91**, 919–926 (https://doi.org/10.1016/j.procs.2016.07.111).

Nanni, L., and Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications* **36**(2), 3028–3033 (https://doi.org/10.1016/j.eswa.2008.01.018).

Paleologo, G., Elisseeff, A., and Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research* **201**(2), 490–499 (https://doi.org/10.1016/j.ejor.2009.03.008).

Ping, Y., and Lu, Y. (2011). Neighborhood rough set and SVM based hybrid credit scoring classifier. *Expert Systems with Applications* **38**(9), 11 300–11 304 (https://doi.org/10.1109/bife.2009.41).

Questier, F., Put, R., Coomans, D., Walczak, B., and Heyden, Y. V. (2005). The use of cart and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems* **76**(1), 45–54 (https://doi.org/10.1016/j.chemolab.2004.09.003).

Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: a new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(10), 1619–1630 (https://doi.org/10.1109/tpami.2006.211).

Rosenberg, E., and Gleit, A. (1994). Quantitative methods in credit management: a survey. *Operations Research* **42**(4), 589–613 (https://doi.org/10.1287/opre.42.4.589).

Shen, F., Zhao, X., Li, Z., Li, K., and Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica* A **526**, 121073 (https://doi.org/10.1016/j.physa.2019.121073).

Soleymani, R., Granger, E., and Fumera, G. (2020). F-measure curves: a tool to visualize classifier performance under imbalance. *Pattern Recognition* **100**, article 107146 (https://doi.org/10.1016/j.patcog.2019.107146).

Sun, J., Li, H., Huang, Q. H., and He, K. Y. (2014). Predicting financial distress and corporate failure: a review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems* **57**(2), 41–56 (https://doi.org/10.1016/j.knosys.2013.12.006).

Sun, J., Fujita, H., Chen, P., and Li, H. (2016). Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vec-

tor machine ensemble. *Knowledge-Based Systems* **120**, 4–14 (https://doi.org/10.1016/j.knosys.2016.12.019).

Šušteršič, M., Mramor, D., and Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications* **36**(3), 4736–4744 (https://doi.org/10.2139/ssrn.967384).

Tan, P. N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*, 1st edn. Addison-Wesley, Boston, MA.

Thomas, L. C. (2009). *Consumer Credit Models*. Oxford University Press (https://doi.org/10.1093/acprof:oso/9780199232130.003.0001).

Thomas, L. C., Crook, J., and Edelman, D. (2017). *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA (https://doi.org/10.1137/1.9781611974560).

Tsymbal, A., Pechenizkiy, M., and Cunningham, P. (2005). Diversity in search strategies for ensemble feature selection. *Information Fusion* **6**(1), 83–98 (https://doi.org/10.1016/j.inffus.2004.04.003).

Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications* **37**(4), 3326–3336 (https://doi.org/10.1016/j.eswa.2009.10.018).

Wang, D., Zhang, Z., Bai, R., and Mao, Y. (2017). A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring. *Journal of Computational and Applied Mathematics* **329**, 307–321 (https://doi.org/10.1016/j.cam.2017.04.036).

Wang, G., Hao, J., Ma, J., and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications* **38**(1), 223–230 (https://doi.org/10.1016/j.eswa.2010.06.048).

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research* **27**(11–12), 1131–1152 (https://doi.org/10.1016/s0305-0548(99)00149-5).

Witten, I. H., and Frank, E. (2002). Data mining: practical machine learning tools and techniques. *ACM SIGMOD Record* **31**(1), 76–77 (https://doi.org/10.1145/507338.507355).

Witten, I. H., Frank, E. and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn. The Morgan Kaufmann Series in Data Management Systems. (https://doi.org/10.1016/c2009-0-19715-5).

Xia, Y., Liu, C., Li, Y., and Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications* **78**, 225–241 (https://doi.org/10.1016/j.eswa.2017.02.017).

Xiao, H., Xiao, Z., and Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing* **43**, 73–86 (https://doi.org/10.1016/j.asoc.2016.02.022).

Yap, B. W., Seng, H. O., and Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications* **38**(10), 13 274–13 283 (https://doi.org/10.1016/j.eswa.2011.04.147).

Zhang, W., He, H., and Zhang, S. (2018). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: an application in credit scoring. *Expert Systems with Applications* **121**, 221–232 (https://doi.org/10.1016/j.eswa.2018.12.020).

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., and Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scor-

ing. *Expert Systems with Applications* **42**(7), 3508–3516 (https://doi.org/10.1016/j.eswa.2014.12.006).