ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20221220085239
Original scientific paper

An Ensemble Learning Based Strategy for Customer subdivision and Credit Risk Characterization

Shuaiqi LIU, Guiying WEI, Sen WU*, Yiyuan SUN

Abstract: Credit customer subdivisions and borrower characteristics are essential tools for banks and lending companies to evaluate credit risk and make profits. This study proposes a Multi-Level Default Risk Rating (MLDRR) strategy based on a heterogeneous ensemble learning method. Further, a novel Eight Subdivisions Model (ESM) of credit customers is constructed. Through the model, the credit customers are subdivided into eight important categories, such as the defaulting customers that are easily missed, the customers with the highest risk, customers with the potential risk, and target customers, etc. Moreover, we describe the risk characteristics of the customer subdivisions and find deficiencies in the agency's existing risk ratings. Finally, the explored strategies are validated on one hundred thousand real credit data, demonstrating the effectiveness of ESM. Compared with the traditional customer segmentation and characteristics research, this paper develops a new credit customer segmentation method based on the perspective of default risk and describes the risk characteristics more comprehensively through eight customer subdivisions.

Keywords: credit customer subdivision; credit risk rating; customer characterization; ensemble learning

1 INTRODUCTION

In the financial market, credit risk evaluation plays an essential role in the credit risk management of financial institutions such as banks and insurance companies [1]. Effectively identifying risks, segmenting customers, and characterizing borrowers' characteristics are beneficial to improving credit risk assessment ability, which is an important means to reduce potential risks and improve profitability. With the rapid development and application of financial technology, driven by credit data, credit risk management innovations continue to emerge.

The core value of credit data lies in credit risk modeling and customer profiling. By segmenting customers and deeply exploring customer characteristics, it can provide effective decision support for risk management and precision marketing. In the related research of customer segmentation and customer characteristics based on machine learning technology, scholars summarize customer characteristics or form subdivide customers from different perspectives. Different segmentation objectives are mainly reflected in different dimensions, corresponding to different methods to apply to specific markets. Rachman et al. [2] segment credit card customers by K-means cluster analysis based on recency, frequency, and monetary (RFM) dimensions. He, C. Z. and Kong, L. [3] constructed a customer segmentation model considering the current value of credit card customers and customer churn prediction to improve the success rate of cross-selling. Zhang Jing et al. [4] achieved credit risk rating through improved KDLOR (kernel discriminant learning for ordinal regression) for social networking users. For the study of borrower characteristics, scholars have found that debt status, turnover rate, financial status, and property status are vital indicators to judge default risks [5]. Vasileios Giannopoulos et al. [6] identified that homeownership, the bank deposit relationship, the existence of ownacilities, etc. were crucial factors for non-performing loans avoidance in recession years. Shiyi Chen et al. [7] found that borrowers' demographic characteristics such as education level were the most efficient factor in forecasting P2P lending in China. Liu, S. Q. and Wu, S. [8] studied the "explicit knowledge" in borrower characteristics by attributes partition considering business process. From the perspective of the research methods adopted, scholars studied credit data through traditional statistical methods and machine learning methods for default risk assessment [9]. Ata Oğuz and Hazim Layth [10] found that the performance of random forest is better than NB, SVM, and KNN on credit card fraud detection.Malekipirbazari and Aksakalli [11] used the Random Forests to develop a lending decision model for credit assessment and demonstrated good accuracy. Ma et al. [12] used the Ada Boost algorithm to build a borrowers' default prediction model. Li, X. and Sun, Y. [13] establish a personal credit rating model by radial basis function neural network model combined with the optimal segmentation algorithm.

It can be seen from the traditional customer segmentation and customer characteristics research, scholars mostly take customer value, customer churn, or customer preference as the perspective and goal of customer segmentation and characteristics research to improve precision marketing. However, in the specific scenario of the personal credit market, the perspective of credit risk as customer segmentation also has practical significance. Many scholars regard default risk assessment as a binary classification problem, and borrowers are divided into defaulting and performing according to their actual labels. However, it is not nuanced enough to split the default risk of borrowers into only two categories. In real business, although the customers belong to the same category label, for example, in the defaulting customer, different samples also have different degrees of default risk, which is easy to understand. Among the performance customers, due to different degrees of default risk, there are also differences such as high-quality and potentially risky customers. These differences can be understood as the different significance of risk characteristics between different samples. In addition, many scholars have devoted themselves to constructing risk models and improving different methods to obtain higher prediction accuracy, but this does not explain their high application value. This is because the minority class of defaulting samples that are easily misclassified have not been fully learned. When machine learning classifies imbalanced data such as credit data, the minority class samples are often ignored or considered noise, so the prediction results are biased towards the majority class [14, 15]. However, in actual credit risk management, the minority class of defaulting samples is often the focus of risk control managers, the correct identification of such samples can not only improve the performance of the model but also contribute to the research of the characteristics of defaulting customers. Facing the above problems, that is necessary to identify some important customer subdivisions by multi-level default risk ratings and characterize these samples more elaborate from different angles, which is important to provide more valuable management advice.

The main contributions can be summarized as follows: (1) A novel strategy for credit customers' multi-level default risk ratings (MLDRR) based on heterogeneous ensemble learning (Section 3.1). (2) A new Eight Subdivisions Model (ESM) for credit customer subdivision based on the MLDRR (Section 3.2). (3) Customer characterization strategy of eight customer subdivisions (Section 3.3). (4) The practicability of the ESM model is verified through one hundred thousand real-world credit data and the discussions around important characteristics that lead to different default risk (Section 4).

2 RELATED WORK

In the field of credit research, borrower credit risk is also known as default risk. Borrower default risk assessment and borrower characterization have always been the core issue in financial risk management. Driven by credit data, borrowers' default risk assessment research methods can be summarized into traditional statistical and machine learning methods [16]. This section describes and comments on research related to credit risk assessment.

In the research based on statistical methods, the LDA and MDA methods have a long history and are still widely used by some institutions, and their prediction effect was close to the successfully applied in financial risk management due to its strong interpretability and high model robustness. It does not require any assumption of probability distribution and normal distribution. However, the disadvantage of the logistic regression model is that its accuracy is not as good as ensemble learning, neural network, and other methods in machine learning theory, especially when dealing with high-dimensional data. Since high-dimensional data is generally nonlinear, using LR models often requires derived variables, resulting in dimensional disaster, and the result often produces overfitting.

The application of big data technology can better help us make business decisions [19]. In the research based on machine learning methods, typical representatives of single classifiers are Support Vector Machines, Decision Trees, Bayesian Networks, etc. [20, 21]. Ghodselahi et al. [22] compared various methods on the German credit risk assessment dataset, and the SVM had the best prediction results, but it was complex and lacked interpretability. However, due to the issues of data sensitivity and overfitting of a single model, credit risk assessment encounters a bottleneck. Ensemble learning is integrating multiple single weak classifiers into a robust classifier, which can effectively improve the model's classification performance and generalization ability [23]. According to the combining strategy of the individual learner, ensemble learning can be roughly divided into Boosting-based and Bagging-based algorithm families. According to the type of individual learner, ensemble learning can also be divided into homogeneous individual learners and heterogeneous individual learners. The Boosting ensemble paradigm is a sequential ensemble method, the representative algorithms include Adaboost, GBDT, etc.,and the boosting trees is also an algorithm with high accuracy in machine learning [24]. The Bagging ensemble paradigm is a parallel ensemble method, and the representative algorithm is the Random Forest with high robustness [25]. In addition, some classical homogeneous ensemble learning algorithms can also be combined through some more complex strategies to achieve heterogeneous ensemble learning, such as the Stacked Generalization strategy [26]. The individual learners in the stacking algorithm are also called primary learners, and the learners used to combine are called meta-learners. Stacking usually uses meta-learner as a combination method to re-learn the output of primary learners, improving the model's generalization ability by increasing the individual classifier's diversity and complementarity. Tran Cao Son et al. [27] address inconsistency and bias in classification performance through the foundation of entropy-based stacking models. Xia Yufei et al. [28] developed a heterogeneous stacking ensemble (HSE) approach and improved the loss-given default (LGD) forecasting in the P2P lending domain. The HSE model outperforms the baseline model in most cases and achieves the best performance in all evaluation metrics. Liang K. et al. [29] combined lasso and stacking methods to study the prediction of default risk of P2P platforms in a high-dimensional imbalanced data environment. With the development of computer technology and deep learning, more and more deep learning algorithms are applied to the financial field, including credit risk assessment [30]. Sirignano et al. [31] used a deep neural network to predict the risk of overdue and early repayment of US real estate loans. The effect was better than logistic regression and artificial neural network models. However, some scholars have noticed that deep learning applications may not achieve good results on small data sets. For example, Shigeyuki et al. [32] compared several standard ensemble learning algorithms with deep neural networks on credit datasets. They verified that ensemble learning is better than deep learning models on such datasets. In summary, in terms of research methods, traditional statistical methods rely on data distribution assumptions, and the ensemble learning method in machine learning is more suitable for such problems.

Focusing on research issues, scholars pay more attention to the prediction of default risk. The forecast is mainly based on the two-category situation of default and non-default. Still, it is not enough to determine what kind of customer has what degree of default risk, which is vital for institutions to reduce losses and form a customer portrait. Facing the high-dimensional data and serious class imbalance characteristics, although the performance of different prediction models is not bad because of the correct classification of performing samples, the study is not sufficient for the minority defaulting samples. The significance of risk characteristics is also different in defaulting samples and performing samples, leading to varying degrees of default risk in samples with the same label. Still, there are few studies which have addressed these questions. Thus, the risk rating according to the identification of different significant degrees of default risk characteristics is the key. Then the characteristics of customer subdivisions based on the default risk ratings can be explored, which can provide more valuable and detailed management suggestions for relevant institutions and is also a director for the researchers that will continue to strive. Therefore, there is an urgent need for a new method to mine more profound "knowledge" in online loan data to meet the requirements for further decision-making.

3 ESM AND CHARACTERIZATION STRATEGY

Credit data has a natural class imbalance problem, and the proportion of positive and negative samples for most data is less than 100/1. That leads to a high-accuracy model in traditional loan data classification research but lacks practical significance and application value. It is because the negative class samples are not sufficiently learned due to their small number, and most of the negative class samples are mispredicted. In this case, the most noteworthy high-risk and potential-risk customers have not been correctly identified, but that is the most concerned issue of the risk models, so these models still have insufficient identification in the actual business. Therefore, in the problem of risk identification, facing the characteristic of data imbalance, we should pay more attention to the risk samples that are easily misclassified with a small proportion rather than the performance samples for a large proportion, and it is significant to discover and learn the characteristics of different important risk samples.

From the perspective of management significance and practical application value, we are the first to propose a credit customer eight subdivision model based on a novel method for credit customers' default risk rating and characterizing the customer subdivisions. The identification and subdivision of different default risks provide new strategies for risk assessment and helpful references for risk control management and customer characterization.

3.1 Multi-Level Default Risk Rating Strategy

This section will introduce the process of default risk identification and the formation of the multi-level default risk rating strategy. Traditionally, we divide credit customers into defaulting samples and performing samples by actual labels without considering that the samples' significance of the risk characteristics is different, which makes the identification of risks not sophisticated enough and does not rate the default risk. Based on amounts of credit data, we find that even the borrowers under the same label have different default risk characteristics, so the default risk is also discrepant. This study draws on the heterogeneous ensemble learning strategy, analyzes the prediction result of each individual learner, and realizes the customers' risk rating based on the different risk characteristics.

In this strategy, the number of individual learners through a heterogeneous ensemble is variable. The more individual learners there are, the more risk ratings are generated, and the more detailed risk levels are formed. In this section, the number of individual learners is four for the example analysis. It is worth mentioning that this strategy can also select an odd number of individual learners as an example, because this strategy has certain universality and this study focuses on risk rating and characterization of customer subdivisions, rather than the risk prediction, so there is no constraint on the number of individual learners. In future research, we will focus on the number and composition of individual learners for better prediction performance. The possible prediction results of four individual learners and the significance of risk characteristics for each sample are shown in Tab. 1.

Table 1 Significance of risk characteristics for different samples

Sample category	P1	P2	P3	P4	True label	Significance of risk characteristic
Sample a	1	1	1	1	1	●●●●●
Sample b	1	1	1	0	1	●●●●
Sample c	1	1	0	0	1	●●●
Sample d	1	0	0	0	1	●●
Sample e	0	0	0	0	1	●
Sample f	0	0	0	0	0	●
Sample g	1	0	0	0	0	●●
Sample h	1	1	0	0	0	●●●
Sample i	1	1	1	0	0	●●●●
Sample j	1	1	1	1	0	●●●●●

Samples "a-e" are default class samples, and their actual label is 1. In the learning process of the sample with the actual label of 1, the prediction result obtained by the individual learner can reflect the significant degree of default characteristics and different degrees of default risk. Specifically, the prediction results of the four individual learners of class "a" are all 1. It can be understood that class "a" samples have the most significant risk characteristics and are identified as having the highest level of default risk. In the same way as above, class "b" and class "c" samples are identified as the second-highest and the third-highest level of default risk. The individual learners' prediction results of class "d" samples and class "e" samples are composed of 1, 0, 0, 0, and 0, 0, 0, 0. Because these two classes of default samples have insufficient risk characteristics and data imbalance problems, most learners cannot correctly classify these two classes of samples. However, such samples are the most worthy of attention in credit risk prediction, and the correct classification of such samples is an important reflection of model performance and practical value. Therefore, identifying such samples and conducting further learning is of great practical significance.

Samples of classes "f-j" are performance samples, and their actual label is 0. Such samples are more likely to be fully learned and correctly classified due to their high proportion in the dataset. However, the performance samples also have different significant degrees of risk characteristics. Identifying these characteristics can obtain important subdivisions such as target and potential risk customers, which is of great significance. Specifically, the individual learners' prediction results of class "f" samples are composed of 0, 0, 0, and 0. It can be understood that these samples have not identified any default risks and have the most significant performance characteristics. Identifying such samples and learning their characteristics is significant to the portrait of target customers. Class "g", class "h", and class "i" samples were identified to have varying degrees of default risk. The individual learners' prediction results of class "j" samples are composed of 0, 0, 0, 0. Although it is a performance class sample, this kind of sample has significant risk characteristics, so this kind of sample has the highest potential risk. The identification and feature learning of such samples can enable institutions to effectively avoid losses caused by potential risks among performing customers.

3.2 Credit Customer Eight Subdivisions Model

This study further proposes an eight subdivisions model (ESM) for subdividing customers into eight types based on multi-level default risk rating. The algorithm is shown in Fig. 1. In the previous section, we took four individual learners for the example analysis, and the number of individual learners of this strategy is variable. Assume that T individual learners are integrated and correspond to T individual learning algorithms. In step 1, the original dataset D is trained separately by each individual learner. In step 2, a new data set is established based on the prediction results of step 1, which is composed of T features. Step 3 is to divide customers into eight subdivisions.

When the sample is in the default category, the actual label is 1, and the customers' subdivision according to the default risk is as follows:

(1) Customers with the highest default risk (customer A). If the sum of the prediction results of all individual learners is T (the number of individual learners), then we consider this sample to have the highest default risk, and the default characteristics of this sample are prominent.

(2) Customers with the second-highest default risk (customer B). If T-1 individual learners' prediction result is 1, it means that most individual learners have learned the default risk of the sample and only one misclassification. Therefore, we subdivide these samples as customers with the second-highest default risk.

(3) Customers with the average level of default risk (customer C). In the same way, if T-2 individual learners are predicting a result of 1 for the sample, we consider the default risk of the sample to be the third level. Thus, in the algorithm, if there are 1 to T-2 individual learners' prediction result is 1, we rate such samples as average default risk. To sum up, by setting the number of individual learners, this study can order different default risk ratings of customers.

(4) The positive samples are most likely to be missed (customer D). If the prediction results of the individual learners for a positive sample are all 0, it indicates that the risk feature of the sample is less significant, so the individual learner cannot correctly classify it. However, such samples are the key to affecting the model's performance. Identifying and strengthening the models' attention to such samples can effectively improve model performance.

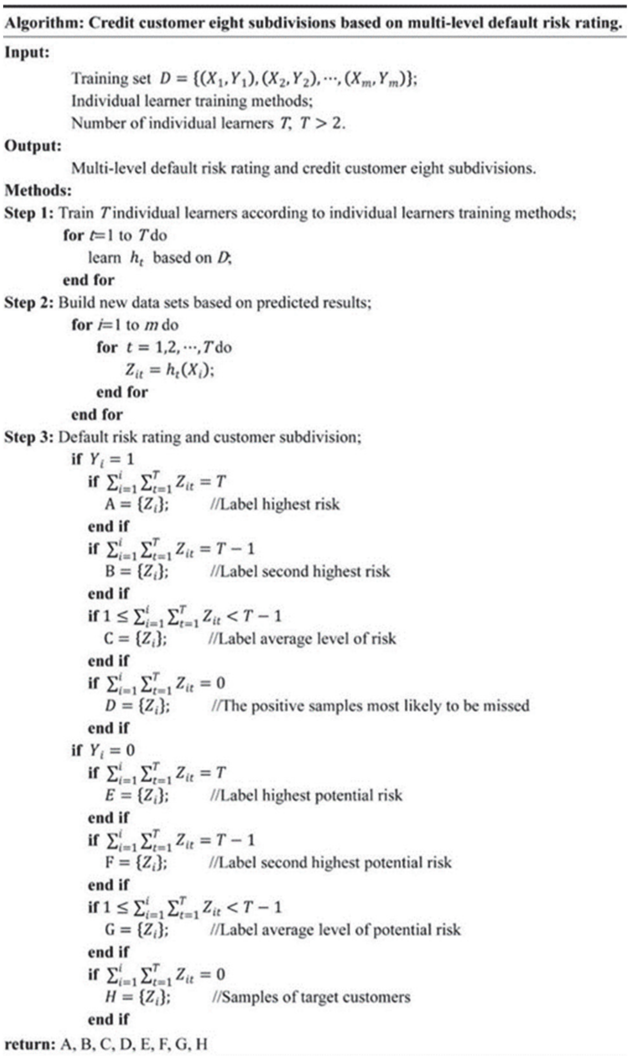
When a sample is a performance class sample, the actual label is 0, and the customer subdivisions according to the default risk are as follows:

(1) Customers with the highest potential default risk (customer E). Suppose the prediction results of the T individual learners for this sample are all 1. In that case, although this sample is a performance customer, it has relatively significant risk characteristics, so we view these samples as the customers with the highest potential risk.

(2) Customers with the second-highest potential risk (customer F). Similarly, if T-1 individual learners' prediction results are 1, the samples' significance of risk characteristics is weaker than the customers with the highest potential default risk. So we view this type of customer as having the second-highest potential risk.

(3) Customers with the average level of potential risk (customer G). In this algorithm, if there are 1 to T-2 individual learners' prediction results for a sample is 0, it indicates that such customers have different degrees of potential risk, so we classify such customers as customers with average potential risk. In addition, by setting the number of individual learners, this study can rate more different potential risks for customers.

(4) Target customer (customer H). If there are T individual learners' prediction results are all 0 for this sample, it means that this sample has not identified any default risk by the individual learners. So we view such customers as high-quality borrowers and subdivide them as target customers.



```
Algorithm: Credit customer eight subdivisions based on multi-level default risk rating.
Input:
        Training set $D = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_m, Y_m)\}$;
        Individual learner training methods;
        Number of individual learners $T$, $T > 2$.
Output:
        Multi-level default risk rating and credit customer eight subdivisions.
Methods:
Step 1: Train $T$ individual learners according to individual learners training methods;
       for $t$=1 to $T$ do
          learn $h_t$ based on $D$;
       end for
Step 2: Build new data sets based on predicted results;
        for $i$=1 to $m$ do
           for $t = 1,2,\cdots,T$ do
              $Z_{it} = h_t(X_i)$;
           end for
        end for
Step 3: Default risk rating and customer subdivision;
        if $Y_i = 1$
          if $\sum_{i=1}^{i}\sum_{t=1}^{T} Z_{it} = T$
            A = $\{Z_i\}$;        //Label highest risk
          end if
          if $\sum_{i=1}^{i}\sum_{t=1}^{T} Z_{it} = T - 1$
            B = $\{Z_i\}$;        //Label second highest risk
          end if
          if $1 \leq \sum_{i=1}^{i}\sum_{t=1}^{T} Z_{it} < T - 1$
            C = $\{Z_i\}$;        //Label average level of risk
          end if
          if $\sum_{i=1}^{i}\sum_{t=1}^{T} Z_{it} = 0$
            $D = \{Z_i\}$;        //The positive samples most likely to be missed
          end if
        if $Y_i = 0$
          if $\sum_{i=1}^{i}\sum_{t=1}^{T} Z_{it} = T$
            $E = \{Z_i\}$;        //Label highest potential risk
          end if
          if $\sum_{i=1}^{i}\sum_{t=1}^{T} Z_{it} = T - 1$
            F = $\{Z_i\}$;        //Label second highest potential risk
          end if
          if $1 \leq \sum_{i=1}^{i}\sum_{t=1}^{T} Z_{it} < T - 1$
            G = $\{Z_i\}$;        //Label average level of potential risk
          end if
          if $\sum_{i=1}^{i}\sum_{t=1}^{T} Z_{it} = 0$
            $H = \{Z_i\}$;        //Samples of target customers
          end if
return: A, B, C, D, E, F, G, H
```

Figure 1 Customer eight subdivisions model (ESM)

In this model, in addition to several customer subdivisions with obvious practical significance, such as customer A, customer D, customer H, etc., we subdivide customers with the average level of default risk into one category, because the customers with the highest risk and the second highest risk have more risk characteristics and identification value, and the same is true for the customers with potential risk. So far, through the identification of different default risks, this research has realized the rating of varying default risks of customers and obtained eight valuable customer subdivisions. Class A is the customer with the highest default risk; class B is the customer with the second-highest risk of default; class C is the customer with average default risk; class D is the most likely to be misclassified positive samples; class E is the customer with the highest potential risk; class F is the customer with the second-highest potential risk; class G is the customer with the average potential risk and class H is the target customer. This model is universal, and eight subcategories can be formed with different number of individual learners.

3.3 Customer Subdivisions Characterization

Based on the default risk rating and customer subdivision model proposed in the previous section, we have obtained eight customer subdivisions according to different default risks of customers, including four subdivisions of defaulting customers (A, B, C, and D) and four subdivisions of performing customers (E, F, G, and H). Based on the above division, this study proposes a study strategy of borrower characteristics under customer subdivision (Fig. 2).

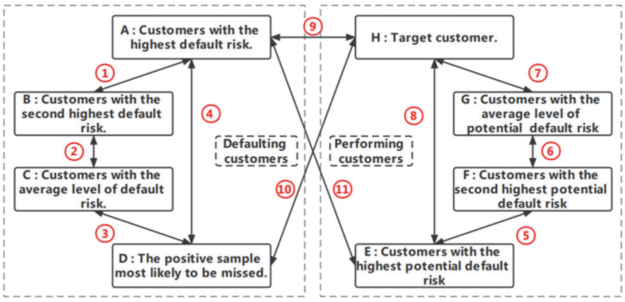

A : Customers with the highest default risk.
H : Target customer.
B : Customers with the second highest default risk.
C : Customers with the average level of default risk.
D : The positive sample most likely to be missed.
G : Customers with the average level of potential default risk
F : Customers with the second highest potential default risk
E : Customers with the highest potential default risk
Defaulting customers
Performing customers
①
②
③
④
⑤
⑥
⑦
⑧
⑨
⑩
⑪

Figure 2 Borrower characteristics under customer subdivision

By modelling the eight subdivided customers in pairs, important features affecting the differences between customer subdivisions were found, and the characteristics of borrowers are described in more detail, which provided a new idea for characterizing borrower characteristics. Each number in the figure represents a group of research objects. As shown in Fig. 2, we studied the characteristics among 11 groups of objects, while the characteristics among some subdivisions, such as class H and class B, were not studied. The characteristics of the labelled 11 groups of research objects are more representative of the actual business. The significance of modelling the 11 groups of research objects in the figure is shown in Tab. 2.

Based on the above analysis, this study firstly identifies varying degrees of default risks among customers through the prediction results of different individual learners and risk ratings for customers, and then subdivides customers through various default risks and obtains several vital subdivisions. Finally, every two sub-classes are modelled as research objects, and the important features that lead to the differences between sub-classes are explored. So far, the research on credit customer risk rating, customer subdivision, and customer characteristics has been completed, which significantly enriches the research on default risk identification and borrower characteristics and provides decision-makers with more helpful knowledge.

Table 2 Characterization of customer subdivisions

The Research Object	Characteristic Learning
Customers with the highest default risk Customers with the second-highest default risk Customers with the average default risk Defaulting customers that are easily misclassified	Characteristics that affect different levels of default risk among defaulting customers
Customers with the highest default riskDefaulting customers that are easily misclassified	Characteristics that lead to the defaulting sample not being correctly classified
Customers with the highest potential default riskCustomers with the second-highest potential default riskCustomers with average potential default risk	
Customers with average potential default risk Target customers Customers with the highest potential default risk	Characteristics that lead to different degrees of potential default risk among theperforming customers
Customers with the highest default risk Target customers	Characteristics that lead to customer performance and default behavior
Defaulting customers that are easily misclassified Target customers	Characteristics that lead to defaulting behavior in the case ofsignificant performance characteristics
Customers with the highest default risk Customers with the highest potential default risk	Characteristics that lead to performing the behavior in the case of significant defaulting characteristics

4 MODELING AND RESULTS

4.1 Data Collection and Preparation

This study is based on the data mining theory and process for the example analysis. To ensure the universality of the experimental conclusions, this study selects the credit data set published by the Lending Club platform as the research object, which is recognized and widely used in credit data research. The example data is selected from the public data in 2019. The original data contains 150 features and 115112 samples. According to the platform business process, 100000 samples and 30 features were selected for the case study after data preprocessing.

Data preprocessing is the most basic link for subsequent analysis, including data cleaning, descriptive statistics, feature selection, and data transformation. First, load the data sets for data cleaning. According to the lending club company's business process, the attributes generated after repaying and irrelevant attributes are removed, and the seriously missing attributes are removed. There are 30 valid attributes left after data cleaning. Some missing value has explanatory significance for the target variable. If the data missing value imputation method is used to fill in, it will change the actual situation of the sample; thus, the missing value of such variable is assigned as 0 in this study.

In the data transformation part, this study assigns numerical labels to the categorical variables in the data to facilitate the processing and construction of subsequent machine learning models. For example, replace the grade attribute's seven values of A, B, C, D, E, F, and G with 1, 2, 3, 4, 5, 6, and 7. For the target attribute loan_status, which the label is Late (16 - 30 days), Late (31 - 120 days), Charged Off, Default, or In Grace Period are the samples in the deferred or default state, denoted by 1; the label of Current, Fully Paid samples is in good loans condition, denoted by 0. Finally, for the numerical data, this study uses the z-score standardization method to standardize the data to improve the model learning efficiency.

4.2 Modeling and Results

In the multi-level default risk rating strategy, four individual learners are taken as an example for the case study. The four individual learners in this study are the RF (Random Forest), Adaboost, GBDT (Gradient Boosting Decision Tree), and ID3 algorithms. Among them, RF, Adaboost, and GBDT are three classic representative ensemble learning algorithms, and the ID3 is a traditional decision tree algorithm. First, modeling by the four individual classifiers, and the accuracy and the confusion matrix reflect the quality of the data and the performance of the individual classifier, shown in Tab. 3. It can be seen that the accuracy of the three classical ensemble learning algorithms is higher than the DT. The learning of positive and negative class samples is improved, but some positive samples are still not correctly classified.

Table 3 The modelling results

Algorithms		Actual "0"	Actual "1"	Acc
RF	Predicted as "0"	28124	930	0.956
	Predicted as "1"	395	551	
Adaboost	Predicted as "0"	28242	812	0.960
	Predicted as "1"	401	545	
GBDT	Predicted as "0"	28267	787	0.961
	Predicted as "1"	382	564	
DT (ID3)	Predicted as "0"	26601	2453	0.907
	Predicted as "1"	340	606	

Tab. 4 summarizes the prediction results of the testing set and eight subdivisions of the customer. It can be seen from the experimental results that this study realizes the identification of the customer's default risk and the risk rating in defaulting customers and even in performing customers. We identified several important customer subdivisions, such as customers with the highest risk, the most misclassified defaulting customers, the target customer, and the performing customers with the highest potential default risk, etc.

Under the realistic background of class imbalance of credit data, a large number of negative classes are correctly classified, resulting in high model accuracy but a lack of learning for risk samples. However, these samples are the most worthy of attention. By identifying such positive samples that are most likely to be misclassified, the model performance can be improved by increasing the attention of the classifier to such samples in the subsequent modelling. In the performing customer, identifying customers with different potential risks and the target customers helps financial institutions to lend more precisely.

Table 4 Customer subdivision results

Customer subdivision	Individual learners predict results	Actual label	Number of samples
Customers with the highest default risk	1, 1, 1, 1	1	544
Customers with the second-highest default risk	1, 1, 1, 0	1	20
Customers with the average level of default risk	1, 1, 0, 0; 1, 0, 0, 0	1	26
Defaulting customers most likely to be misclassified	0, 0, 0, 0	1	356
Target customer	0, 0, 0, 0	0	26592
Customers with the average level of default risk	0, 0, 0, 1; 0, 0, 1, 1	0	1702
Customers with the second-highest potential default risk	0, 1, 1, 1	0	31
Customers with the highest potential default risk	1, 1, 1, 1	0	729

Finally, we study the characteristics of the customers' subdivisions by the Random Forest algorithm. The RF is an ensemble learning method based on the decision tree, which has better anti-noise ability and generalization ability and can obtain important features that affect different target variables. The top 5 important features of each customer subcategory and the feature importance are shown in Tab. 5.

Table 5 Top 5 important features

The modeling object	Top 5 important features	Feature importance
customers with the highest risk and target customers	sub_grade	0.457
	int_rate	0.038
	emp_length	0.030
	dti	0.029
	avg_cur_bal	0.029
Defaulting customers most likely to be misclassified and target customers	addr_state	0.378
	zip_code	0.312
	dti	0.109
	annual_inc	0.102
	emp_length	0.029
customers with the highest risk and Defaulting customers most likely to be misclassified	sub_grade	0.321
	zip_code	0.314
	emp_length	0.146
	addr_state	0.122
	emp_title	0.057
customers with the highest potential risk and target customers	dti	0.309
	int_rate	0.256
	addr_state	0.184
	sub_grade	0.103
	zip_code	0.079

Through the analysis of the experimental results, it can be seen that the risk characteristics of the most easily missed default customer and the target customer are similar, and the credit rating of borrowers by the platform (sub_grade) can distinguish the "good" and "bad" loans in most cases, but it is not accurate enough. Because "sub_grade" is not included in the top five important features that affect the most easily missed defaulting customers and target customers, it reveals that "rating" does not distinguish between defaulting customers and performing customers with insignificant risk characteristics. However, this study finds that the natural customer information of "customer's state" and "postal code" is an important feature to distinguish the default customers from the performance customers. In addition, the important features affecting the highest potential risk customers and target customers are "loan interest rate", "monthly debt repayment ratio", "state", etc., while the importance of "rating" is significantly lower than the former, revealing that "rating" does not distinguish these two types of customers, and the loan interest rate is the key feature leading to the difference between them. This study makes the research on customer characteristics more detailed and specific, provides more abundant information for the relevant institutions and has practical application value.

5 CONCLUSION

The study of default risk assessment and borrower characteristics is of great significance in the field of credit. Few studies have segmented customers according to different default risks to describe customer characteristics. To fill the research gap, this study proposes a novel multi-level default risk rating (MLDRR) strategy and a new Eight Subdivisions Model (ESM) for credit customer segmentation based on heterogeneous ensemble learning. Furthermore, we characterize the important features between different customer subdivisions. Specifically, we divide customers into eight subdivisions according to the default risk, including the customers with the highest default risk, the defaulting customers most likely to be misclassified, the customers with the highest potential risk, and the target customers, etc. In actual business scenarios, facing the class-imbalance problem of credit data, identifying such customer subdivisions can provide decision-makers with richer information and make financial institution lending more targeted to increase profits remarkably. Finally, we explore the specific characteristics that lead to differences in customer subdivisions and find that the credit rating by the platform has some shortcomings, which provides a new idea for the study of credit risk assessment and customer characteristics.

In the future, we will research credit data from more perspectives, such as improving the classification accuracy of easily misclassified samples and digging deeper into borrower characteristics.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 71971025.

6 REFERENCES

[1] Ciurea, C., Chiriță, N., & Nica, I. (2022). A Practical Approach to Development and Validation of Credit Risk Models Based on Data Analysis. *Economic Computation And Economic Cybernetics Studies And Research*, *56*(3), 51-68. https://doi.org/10.24818/18423264/56.3.22.04

[2] Rachman, F., Santoso, H., & Djajadi, A. (2021). Machine Learning Mini Batch K-means and Business Intelligence Utilization for Credit Card Customer Segmentation. *International Journal of Advanced Computer Science and Applications(IJACSA)*, *12*(10). https://doi.org/10.14569/IJACSA.2021.0121024

[3] He, C, Z. & Kong, L. (2013). Credit card customer segmentation model based on CLV elements. *Statistics and decision*, (11), 183-185.

[4] Zhang, J., Guo, J., & Ren, Y. (2021). Robust ordinal regression: user credit grading with triplet loss-based sampling. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM).* https://doi.org/10.1145/3408303

[5] Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, *47*(1-3), 54-70. https://doi.org/10.1080/00036846.2014.962222

[6] Vasileios, G., Eleftherios, A., & Antonios, G. (2022). Borrower characteristics and loan performance: evidence from micro and small Greek firms. *International Journal of Banking, Accounting and Finance*, *13*(1), 1-31. https://doi.org/10.1504/IJBAAF.2022.121547

[7] Shiyi, C., Yan, G., Qingfu, L., & Yiuman, T. (2020). How do lenders evaluate borrowers in peer-to-peer lending in China? *International Review of Economics and Finance*, *69*, 651-662. https://doi.org/10.1016/j.iref.2020.06.038

[8] Liu, S. Q. & Wu, S. (2020). Multi-angle p2p borrower characterization analytics by attributes partition considering business process. *Intelligent Systems, IEEE*, *PP*(99), 1-1. https://doi.org/10.1109/MIS.2020.2986973

[9] Vishwakarma, A, C. & Solanki, R. (2018).Analysing Credit Risk using Statistical and Machine Learning Techniques. *International Journal of Engineering Science and Computing*, *8*(6), 18397-18404.

[10] Ata, O. & Hazim, L. (2020). Comparative Analysis of Different Distributions Dataset by Using Data Mining Techniques on Credit Card Fraud Detection. *Tehnicki vjesnik-Technical Gazette*, *27*(2), 618-626. https://doi.org/10.17559/TV-20180427091048

[11] Malekipirbazari, M. & Aksakalli, V. (2015). Risk Assessment in Social Lending Via Random Forests. *Expert Systems With Applications*, *42* (10), 4621-4631. https://doi.org/10.1016/j.eswa.2015.02.001

[12] Ma, L., Zhao, X., Zhou, Z., & Liu, Y. (2018). A new aspect on p2p online lending default prediction using meta-level phone usage data in china. *Decision Support Systems,* S0167923618300836. https://doi.org/10.1016/j.dss.2018.05.001

[13] Li, X. & Sun, Y. (2020). Application of RBF neural network optimal segmentation algorithm in credit rating. *Neural Computing and Applications*, (2). https://doi.org/10.1007/s00521-020-04958-9

[14] Li, J. M., Sun, J. T., Huang, W. H., Zhang, Q. Y., Tian, Z. Z., Lu, N. (2020). Unsupervised Text Topic-Related Gene Extraction for Large Unbalanced Datasets. *Tehnicki vjesnik-Technical Gazette*, *27*(3), 842-852. https://doi.org/10.17559/TV-20191111095139

[15] Li, Y., Guo, H., & Liu, X. (2016). A boosting based ensemble learning algorithm in imbalanced data classification. *Systems Engineering-Theory & Practice.*

[16] Dahiya, A., Gautam, N., & Gautam, P. K. (2021). Data Mining Methods and Techniques for Online Customer Review Analysis: A Literature Review. *Journal of System and Management Sciences*, *11*(3), 1-26. https://doi.org/10.33168/JSMS.2021.0401

[17] Serrano-Cinca, C. & Gutiérrez-Nieto, B. (2013). Partial least square discriminant analysis for bankruptcy prediction. *Decision Support Systems*, *54*(3), 1245-1255. https://doi.org/10.1016/j.dss.2012.11.015

[18] Chun, Y. H. & Cho, M. K. (2022). An empirical study of intelligent security analysis methods utilizing big data. *Journal of Logistics, Informatics and Service Science*, *9*(1), 26-35. https://doi.org/10.14704/WEB/V19I1/WEB19311

[19] Mekvabidze, R. (2020). From business modeling to business management: an exploratory study of the optimal decision making on the modern university level. *Journal of Logistics, Informatics and Service Science*, *7*(1), 67-86. https://doi.org/10.33168/LISS.2020.0106

[20] Tampubolon, P. & Girsang, A. S. (2021). Classification of Attacks through the Type of Protocol Using Data Mining. *Journal of System and Management Sciences*, *11*(2), 1-14. https://doi.org/10.33168/JSMS.2021.0201

[21] Meltem, K., Nevcihan, D., Mucella, C, M., & Tarik, D. H. (2016). Prediction of magnetic susceptibility class of soil using decision trees. *Tehnicki vjesnik-Technical Gazette*, *23*(1), 83-90. https://doi.org/10.17559/TV-20140807111130

[22] Ghodselahi, A. & Amirmadhi, A. (2011). Application of artificial intelligence technology for credit risk evaluation. *Int. J. of Modeling and Optimization*, *1*(3), 243-249. https://doi.org/10.7763/IJMO.2011.V1.43

[23] Zhou, Z. H. (2012). Ensemble Methods: Foundations and Algorithms. *Taylor & Francis.* https://doi.org/10.1111/insr.12042_10

[24] Augustyn, L., Małgorzata, K., Tone, L., & Maciej, S. (2020). Predicting the Probability of Cargo Theft for Individual Cases in Railway Transport. *Tehnicki vjesnik-Technical Gazette*, *27*(3), 773-780. https://doi.org/10.17559/TV-20190320194915

[25] Özgur, C. & Sarikovanlik, V. (2022). Forecasting BIST100 and NASDAQ Indices with Single and Hybrid Machine Learning Algorithms. *Economic Computation And Economic Cybernetics Studies And Research*, *56*(3), 235-250. https://doi.org/ 10.24818/18423264/56.3.22.15

[26] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2): 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

[27] Tran, C. S. (2021). Modeling Credit Risk: A Category Theory Perspective. J*ournal of Risk and Financial Management*, *14*(7), 298-298. https://doi.org/10.3390/jrfm14070298

[28] Xia, Y., Zhao, J., He, L., Li, Y., & Yang, X. (2021). Forecasting loss given default for peer-to-peer loans via heterogeneous stacking ensemble approach. *International Journal of Forecasting*, 37. https://doi.org/10.1016/j.ijforecast.2021.03.002

[29] Liang, K., Zhang, C., & Jiang, C. (2022). Analyzing default risk among P2P platforms based on the LAS-STACK method by considering multidimensional signals under specific economic contexts. *Electronic Commerce Research*, 22. https://doi.org/10.1007/s10660-021-09505-9

[30] Heaton, J. B., Polson, N. G., & Witte, J, H. (2017). Deep Learning in Finance. *Applied Stochastic Models in Business and Industry*, *33*(1), 561-580. https://doi.org/10.1002/asmb.2230

[31] Sirignano, J., Sadhwani, A., & Giesecke, K. (2017). Deep Learning for Mortgage Risk. *Social Science Electronic Publishing*, *22*(6), 134-216. https://doi.org/10.2139/ssrn.2799443

[32] Shigeyuki, H., Minami, K., & Takahiro, K. (2018). Ensemble Learning or Deep Learning? Application to Default Risk Analysis. *Risk and Financial Management, 11*(1), 12-25. https://doi.org/10.3390/jrfm11010012

Contact information:

Shuaiqi LIU, PhD candidate
School of Economics & Management,
University of Science and Technology Beijing,
No.30 Xueyuan Road, Beijing 100083, China
E-mail: b20190404@xs.ustb.edu.cn

Guiying WEI, PhD, Associate Professor
School of Economics and Management,
University of Science and Technology Beijing,
No.30 Xueyuan Road, Beijing 100083, China
E-mail: weigy@manage.ustb.edu.cn

Sen WU, PhD, Full Professor
(Corresponding author)
School of Economics & Management,
University of Science and Technology Beijing,
No.30 Xueyuan Road, Beijing 100083, China
E-mail: wusen@manage.ustb.edu.cn

Yiyuan SUN, PhD candidate
School of Economics & Management,
University of Science and Technology Beijing,
No.30 Xueyuan Road, Beijing 100083, China
E-mail: s20190990@xs.ustb.edu.cn