Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CUS-LightGBM-based financial distress prediction for small- and medium-sized enterprises with imbalanced data

Sumei Ruan
Anhui University of Finance and Economics
Jiayong Zhang
Anhui University of Finance and Economics
Wei Li (liweiaufe@163.com)
Anhui University of Finance and Economics https://orcid.org/0000-0002-4420-131X

Research Article

Keywords: Financial distress prediction, Imbalanced data, Small- and medium-sized enterprises, CUS-LightGBM model, Ensemble learning

Posted Date: May 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-415706/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

CUS-LightGBM-based financial distress prediction for small- and medium-sized enterprises with imbalanced data

Sumei Ruan, Jiayong Zhang, Wei Li*

School of Finance, Anhui University of Finance and Economics, Bengbu 233030, Anhui, China

Abstract: In this paper, by analyzing the financial data of small- and medium-sized enterprises, it is found that there is a general problem of imbalance data. Therefore, an effective financial distress prediction model based on clustering under-sampling and LightGBM model is constructed. Based on the idea of ensemble learning, this paper proposes the CUS-LightGBM (cluster-based under-sampling with LightGBM) model. First, the data are divided into minority and majority class samples. Then the K-means algorithm is used to cluster the majority class samples, and some data are selected from each cluster to form the balanced data. Finally, it is fused with the LightGBM algorithm based on decision tree to form an efficient prediction model. In addition, there are a large number of redundant features in the proposed model, which will reduce the prediction accuracy and efficiency of the model. Therefore, this paper adopts the feature selection based on ensemble strategy, and determines the main risk factors according to the principle of minority obeying majority. Finally, through the experimental analysis of real financial data, the results show that the CUS-LightGBM model can significantly improve the recognition ability of small- and medium-sized enterprises in financial distress, and the proposed model is more effective in processing financial ratio data than the benchmark model.

Keywords: Financial distress prediction; Imbalanced data; Small- and medium-sized enterprises; CUS-LightGBM model; Ensemble learning.

* Corresponding author.

E-mail addresses: ruansumei0116@163.com (S. Ruan); jiayongmac@163.com (J. Zhang); liweiaufe@163.com (W. Li) ORCID iD: 0000-0002-4420-131X

1. Introduction

Bankruptcy prediction(Hu 2020) has been a hot and challenging field of research for decades. During the Asian financial crisis of 1997 and the global financial crisis of 2008, many well-known enterprises ran into financial distress. Since the wave of corporate bankruptcy tends to threaten the security of the banking system, financial statement analysis and risk management become particularly important in financial market. Lenders (creditors), other investors, and regulators (auditors) all require a timely understanding of the risk of default in the portfolio of loans and derivatives. Specifically, banks require the development of effective internal rating systems and the establishment of default probability models suitable for different enterprises of different sizes to improve corporate risk management to cope with changes in the macro environment. Enterprises require the establishment of financial distress prediction model to help internal managers to realize the deterioration of financial situation earlier and take measures to prevent it as soon as possible. Investors need to be able to use the enterprises' predictive information to choose whether to buy the corporate bonds or stocks. Misjudgment will bring huge economic losses to investors and even lead to economic recession. And it is crucial for auditors to make effective use of the predicted information for a business-sustainability assessment. Therefore, in the past two decades, financial distress prediction of enterprises has been widely studied in the field of accounting and finance but establishing an effective financial distress prediction model is a complicated and arduous task for financial institutions. The financial distress prediction aims to predict whether a new applicant, including enterprises and individuals, will be in financial distress. If the prediction model cannot provide a model with a high prediction error rate well, it will lead to making wrong decisions to new applicants, and then make financial institutions or investors face huge losses.

The research for financial distress prediction dates back to the 1960s. Between 1966 and 1980, Altman, Beaver and Ohlson were pioneers in the study of financial distress. Although these three researchers used different methods and data respectively, they provided important reference value for the research of bankruptcy prediction. For example, Beaver in 1966 used the univariate analysis method to predict the failure probability of an enterprise. Later, as computing power improved, and financial data sets increased, more advanced statistical methods became popular. For example, Fisher in 1936 first proposed a discriminant analysis.

Altman in 1968 considered multivariate discriminant analysis (MDA) to quantify the critical value of Z-score and proposed Z-score model. Ohlson in 1980 improved the bankruptcy prediction model and used logistic regression (LR) to estimate the bankruptcy probability of enterprises, avoiding common problems related to MDA. After the 1990s, with the development of information technology, many financial distress prediction models were built on the basis of complex artificial intelligence approaches, so the popular trend of bankruptcy prediction method changed from statistical method to artificial intelligence method. Li, Y. et. al (2016)(Li et al. 2016) combined the initial artificial neural network (ANN) and LR models and proposed a hybrid ANN / Logistic model. Kruppa, J. et. al (2013)(Kruppa et al. 2013) used the probability-estimated random forest (RF-PET), k-nearest neighbor (k-NN), and Bagging-based k-NN machine learning methods to estimate the required probability of default and proved that RF was superior to the standard LR model and even the adjusted LR model. Bao, L. et. al (2016)(Bao et al. 2016) improved classifier performance by using improved Near-miss, Under-sampling and SVM ensemble model. Olson, D.L. et. al (2012)(Olson et al. 2012) compared the bankruptcy prediction effect of statistical methods and intelligent methods respectively and found that intelligent methods were superior to statistical methods because intelligent methods could consider a large number of attributes and evaluate the complex relationship between them.

The famous "no free lunch theorem"(Wolpert and Macready 1997) also applies to machine learning, that is, a single classifier is not a good solution to the problem of financial distress. Because different enterprises have their own features in data set size, data structure and prediction variables, a single algorithm cannot meet all the problems that need to be solved. Therefore, it is necessary to apply ensemble in financial distress prediction. Ensemble learning combines multiple algorithms to process different assumptions and then make good predictions(Papouskova and Hajek 2019). Its structure can be roughly divided into two categories, parallel and sequential(Xia et al. 2017). Parallel ensemble is a combination of different learning algorithms, each of which generates an independent model in parallel. In contrast, sequential ensemble is the first learning algorithm generates a model and then the second learning algorithm corrects the previous model, and so on. Although the application trend of ensemble learning in financial distress prediction is very active, researchers focus mainly on parallel ensemble, because in parallel ensemble, each basic model decides a specific

problem independently and then combines these decisions to make the final decision. However, the advanced sequential ensemble is rarely considered, and the parallel structure is easy to apply parallel and distributed learning, which speeds up the learning process. And through the enhancement technology, the disadvantages of high error rate can be overcome to some extent. Therefore, sequential ensemble, such as AdaBoost, were used as the benchmark model, and then parallel integration was used to make the model practical for large-scale data and achieve better overall performance.

Generally speaking, classifier ensemble is based on training a certain number of classifiers for the same field problem, combining the output of trained classifiers to obtain the final output on the given unknown data samples. In related literature studies, Bagging and Boosting are two widely used ensemble methods. Wang, B. and Pineau, J. (2016)(Boyu and Joelle 2016) describes two theoretically reasonable online cost-sensitive bagging algorithms and online cost-sensitive boosting algorithms, and proves that the convergence of these algorithms is guaranteed under certain conditions. Lusa, L. (2016)(Blagus and Lusa 2017) successfully eliminated a few event deviations by using the GBDT and found that the GBDT algorithm was superior to other considered ensemble classifiers through comparison, which also demonstrated the flexibility and interpretability of the proposed method. Although many relevant studies have proved the superiority of classifier ensemble over many individual classifiers, most studies have only constructed specific classifier ensemble types for bankruptcy prediction, such as decision tree ensemble, neural network ensemble(Shen et al. 2019). In addition, most of these classifier ensembles are only based on a specific combination method. In practice, when the feature dimension is high, and the data volume is large, the efficiency and scalability are still unsatisfactory. The main reason is that for each function, they need to scan all data instances to estimate the possible split points of the information gain for all instances, which is time-consuming. Moreover, real-world corporate data sets tend to be imbalanced (imbalanced data refers to situations where minority groups are far less than majority groups, resulting in uneven sample distribution). This means that the number of trustworthy good samples is much larger than the number of potentially defaulted bad samples(Shen et al. 2019). Therefore, imbalanced data becomes a common problem in credit rating or financial distress prediction, which can significantly affect the performance of traditional financial distress prediction models. Data are

biased towards most categories, the classification results may be biased, and because the wrong samples are wrongly classified as good samples, significant economic losses may be caused. Therefore, although data mining methods have been widely used in business and management decision-making, class imbalance is still an important challenge due to the particularity of classification model. This means that it is important to pre-process imbalanced data before modeling.

To solve the problem of imbalanced data, this paper is inspired by ensemble and ML data processing technology, using the cluster-based under-sampling algorithm to balance the imbalanced financial ratio data. Since the cluster-based under-sampling (CUS) has been found to be effective in solving the problem of data imbalance, it has attracted great attention from scholars and practitioners in the past few years. In addition, the existing learning algorithms improve the classification accuracy to the maximum extent by correctly classifying the majority of classes but misclassify the minority groups. In practice, a few class instances are more interesting than most class instances. Traditional ML algorithms, such as DT, NB and KNN, establish classification models to maximize the classification rate, but ignore a few classes. The CUS can solve this problem by using k-means clustering algorithm to cluster most class instances into several classes, and then forms a balanced data set that most class instances are almost equal to a few class instances. Previous studies have found that the learning efficiency and accuracy of weak learners can significantly affect the performance improvement of the final model, and there is a proportional coefficient between the ability of weak learners and the ability of the final ensemble predictor. In order to improve the performance of ensemble model, this paper proposes a new ensemble model, which is a CUSBoost method based on what Rayhan, F. et. al (2018)(Rayhan et al. 2017) proposed. It divides imbalanced datasets into two parts: majority class instances and majority class instances. Then, the k-means clustering algorithm is used to cluster the majority class instances into several classes, and the majority class instances are selected from each cluster to form a balanced data set. Inspired by the imbalanced learning technique combining sampling and boosting method, this paper adopts the CUS to effectively balance imbalanced data and LightGBM algorithm to optimize weak learners. The ensemble model of CUS-LightGBM is used to predict. By comparison with other methods (such as XGBoost, GBDT, etc.), we found that CUS-LightGBM algorithm has better

solution quality, higher success rate, fewer parameters, simpler program, and better prediction accuracy than CUSBoost. Thus, the classification performance of basic learners is significantly improved, and the model has obvious advantages in various performance indicators. Therefore, this model has a good potential to improve the effectiveness and efficiency of financial distress assessment.

The structure of the rest of this paper is as follows: Section 2, we briefly review the related papers and progress about financial distress prediction; Section 3, the construction of the model is described; Section 4, we conducted a preliminary statistical analysis of the data, use the feature selection technique to make a deep analysis of the importance of the target variables of different features, and analyze the experimental results. Section 5, we give the conclusion and put forward some opinions for future research.

2. Literature review

2.1. Research on feature selection

In classification, training sample reduction and attribute selection are usually combined to overcome the problems related to particularity, noise and contradictory data. Zoričák, M. et. al (2020)(Zoričák et al. 2020) used three supervised filtered feature selection techniques (tree-based feature selection, Fisher score, and ReliefF), unsupervised LAP-scored feature selection methods, and SVM-based recursive feature elimination for feature selection. Ke, G. et. al (2017)(Ke et al. 2017) proposed a LightGBM method using GOSS and EFB techniques to reduce the number of features. Papouskova, M., & Hajek, P. (2019)(Papouskova and Hajek 2019) uses MOEFS, which uses ENORA-based MOEFS algorithm to minimize the number of selected features for screening. Li, K. et. al (2016)(Li et al. 2016) used the Filterd to rank features in light of the probability density estimation model. Guo, H. et. al (2016)(Haixiang et al. 2016) used BPSO as the feature selection algorithm, which belongs to Wrapper mode, and proposed the ensemble algorithm of BPSO-AdaBoost-KNN, which integrated feature selection and boost into the system(Verikas et al. 2010). Ng, W. et. al (2016)(Ng et al. 2016) proposed a self-encoder-based feature learning method to learn a set of features with better classification capabilities for minority and majority classes to address the problem of classification imbalance.

2.2. Research on processing methods of imbalanced data

The category unbalanced data set means that the sample size of enterprises with good financial status is much higher than that of enterprises with poor financial status when making financial distress prediction(Li et al. 2016). Many researchers have made efforts to solve this problem and have proposed some effective imbalanced learning methods. One is to adopt cost-sensitive learning methods. This kind of learning method is to increase the weight of misjudgment cost of default samples by adding a cost matrix composed of penalty coefficient of class misjudgment, and then realize the cost sensitivity. For example, Cheng, F. et. al (2016)(Cheng et al. 2016) considered the cost of misclassification, which simplified the original imbalanced data classification problem to optimization problem and minimizes the misclassification of all data samples with better results. Maurya, C. et. al (2016)(Maurya et al. 2016) found that cost-sensitive learning methods were more computationally effective than data sampling techniques in classifying big data. Another is to use the sampling technique widely to solve the problem of imbalanced data. By adjusting the number of samples in different categories, it tries to balance the original data on the basis of a series of sampling algorithms, and then uses classification algorithms to train the new "balanced" data. For example, Shen, F. et. al (2019)(Shen et al. 2019) proposed a SMOTE method to rebalance the target training dataset, inserting new samples between minority individuals and neighbors in the same class instead of directly copying samples from minority classes and then using the PSO algorithm to optimize the weak classifier. He, H. et. al (2018)(He et al. 2018) used EBCA to reduce redundant samples in majority classes to make all kinds of data as balanced as possible to construct a set of adjustable data. Sun, J. et. al (2020)(Sun et al. 2020) proposed two kinds of imbalanced dynamic financial distress prediction models based on SMOTE and ADASVM-TW, i.e. S-SMOTE-ADASVM-TW model and E-SMOTE-ADASVM-TW model. An ensemble classification method is also used to solve this problem. For example, Rayhan, F. et. al (2018)(Rayhan et al. 2017) used ensemble learning to deal with highly imbalanced datasets, which is a new CUSBoost based method. Gong, J. and Kim, H. (2017)(Gong and Kim 2017) adopted the mixed sampling strategy, which combined under-sampling with ROSE sampling and adopted the AdaBoost algorithm as an ensemble technique to form an RHSBoost method to effectively deal with class imbalance. Sigrist, F. and Hirnschall, C. (2019)(Sigrist and Hirnschall 2019) introduced a new Grabit model of binary classification to solve the class

imbalance. Raghuwanshi, B. and Shukla, S. (2018)(Raghuwanshi and Shukla 2018) used the variant class nucleation of ELM to effectively address class imbalance.

2.3. Research on the financial distress prediction

Ma, X. et. al (2018)(Ma et al. 2018) applied LightGBM and XGBoost algorithms to the P2P network credit default prediction model. According to the output of the model, the multi-observation data cleaning method is better than the multi-dimensional method, and the classification prediction effect of the LightGBM algorithm is better on the same data set with an accuracy of 80.1%. Xia, Y. et. al (2017)(Xia et al. 2017) adaptively adjusted the hyperparameters of XGBoost by using the Bayesian hyperparameter optimization, used feature selection to remove the system redundant variables, and then train the model with the selected feature subset. The results showed that the XGBoost-TPE model achieved the best probability prediction ability. Karabadji, N. et. al (2019)(Karabadji et al. 2019) used an effective particle swarm optimization (PSO) algorithm to combine attribute selection and data adoption process to search for "optimal" solutions and optimize the automatic construction of the decision tree. Finally, an improved decision tree (PSO-DT) model is constructed. Sun, J. et. al (2018)(Sun et al. 2018) established an effective DT ensemble model, which combines SMOTE, Bagging and Differential Sampling Rate (DSR), is called the DTE-SBD model. This model solves the problem of imbalance between high-risk enterprises and low-risk enterprises and improves the performance of financial evaluation. Huang, X. et. al (2018)(Huang et al. 2018) established an enterprise credit risk assessment model based on NN algorithm by combining Bootstrap aggregation with a few sampling techniques based on the data set of private small- and medium-sized enterprises in China.

3. Methodology

3.1. Overview for financial distress prediction model

In supervised machine learning algorithms, our goal is to learn a stable and well-behaved model in all aspects, but the actual situation is often not so ideal. It may be that model "A" is better than model "B" in some respect, but model "B" is not always bad. In supervised machine learning algorithms, the goal is to learn a model that is stable and performs well in all aspects, but the reality is often less than ideal. Maybe model "A" performs better than model "B" in

some respect, but model "B" does not perform poorly everywhere. That is to say, each weak classifier has a certain "accuracy", and there are "differences" among classifiers. Many studies have shown that the model constructed by the ensemble idea is better than the single model. In this paper, a new accurate machine learning (ML) model is constructed by the ensemble idea based on the decision tree (DT). The constructed model method framework is shown in figure 1. Ensemble learning is to construct and combine several "good and different" learners to accomplish the learning task. Even if one weak classifier gets the wrong prediction, other weak classifiers can also correct the error. Finally, the results of ensemble learning are generated by the voting method "the minority is subordinate to the majority". However, ensemble learning also has the disadvantage of high computational complexity and the difficulty of explaining the model.

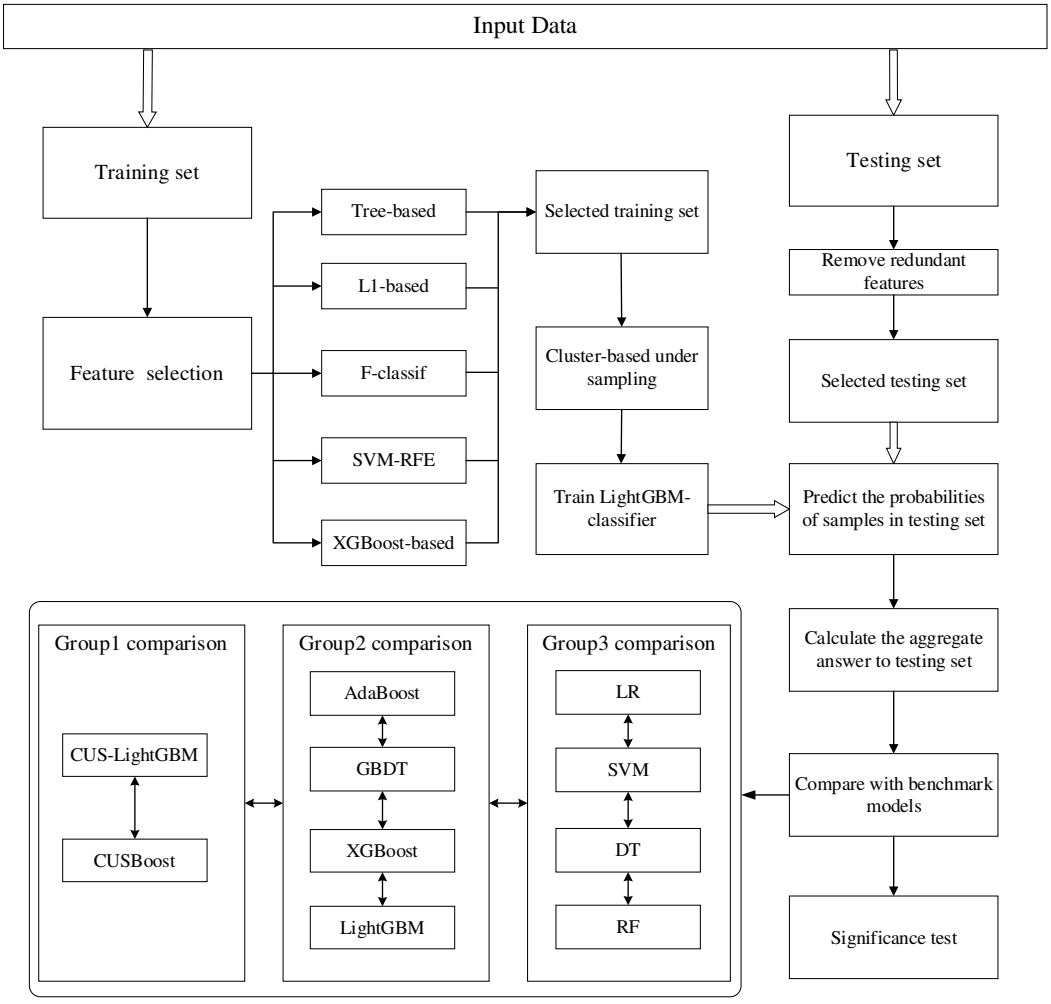

Fig 1. Flowchart of the proposed financial distress prediction model

3.2. CUS-LightGBM algorithm

The CUS-LightGBM method is an improvement on the basis of the CUSBoost method. CUSBoost uses a combination of cluster sampling from majority classes and AdaBoost algorithms. It separates the majority and minority class instances from the original dataset and uses k-means cluster to optimize the hyper-parameter of the majority class instance cluster, and then performs random under-sampling for each created cluster by randomly selecting 50% of the instances and deleting the remaining 50% of the instances. Since clustering is used before sampling, theoretically, the algorithm performs best when the dataset has a high clustering capability. These representative samples are then combined with minority class instances to obtain a balanced dataset. The strength of the algorithm lies in the fact that it considers all the examples of subspaces from most classes, because k-means clustering places each instance in a certain group. Other similar methods often fail to obtain the proper representation of majority classes. Specifically, CUSBoost uses the C4.5 algorithm to consider a series of decision trees and classifies new instances in light of the voting of each individual tree.

$$error(M_i) = \sum_{i=1}^{d} \omega_i \times err(x_i)$$

If in an instance, the x_i in the No. i iteration is correctly classified and its weight is $\frac{error(M_i)}{1-error(M_i)}$, the weight of all instances (including the misclassified instance) is normalized. To normalize a weight, we multiply it by the sum of the old weights, divided by the sum of the new weights. Therefore, the weight of the wrong classification instance is increased, and the weight of the correct classification instance is reduced. If the error rate of model M_i exceeds 0.5, we discard M_i and export the new M_i by generating a new sub-dataset D_i.

CUS-LightGBM is a new method proposed in this paper. It is an improvement on CUSBoost algorithm. The key point of CUS-LightGBM algorithm is the clustering-based under-sampling method (k-means), and then LightGBM algorithm is combined to get the final model. The central idea of CUS-LightGBM is to divide the data into minority and majority class instances first, then use K-means algorithm to cluster the majority class instances and select part of the data from each cluster to form the balanced data. Then it is combined with a fast, distributed and high-performance gradient lifting framework LightGBM algorithm based on decision tree algorithm to form an efficient predictive output model. Figure 2 shows the

frame diagram of cluster-based LightGBM. It contains four steps. In this paper, a two-class training dataset *D* is given. One is a dataset containing majority samples, denoted as $D_{majority}$, and the other is the minority dataset, denoted as $D_{minority}$(Tsai et al. 2019).

The first step is to use the clustering analysis algorithm (k-means) to divide the $D_{majority}$-similar data samples into a number of *K* groups (that is, the clustering process for majority category instances). Since there is a specific clustering correlation between the data samples of K groups and the samples outside K groups, each identified cluster can be regarded as a "pseudo" subclass data set of $D_{majority}$ based on the clustering results. Therefore, we note that majority class datasets of this "pseudo" subclass are $D'_{majority}$.

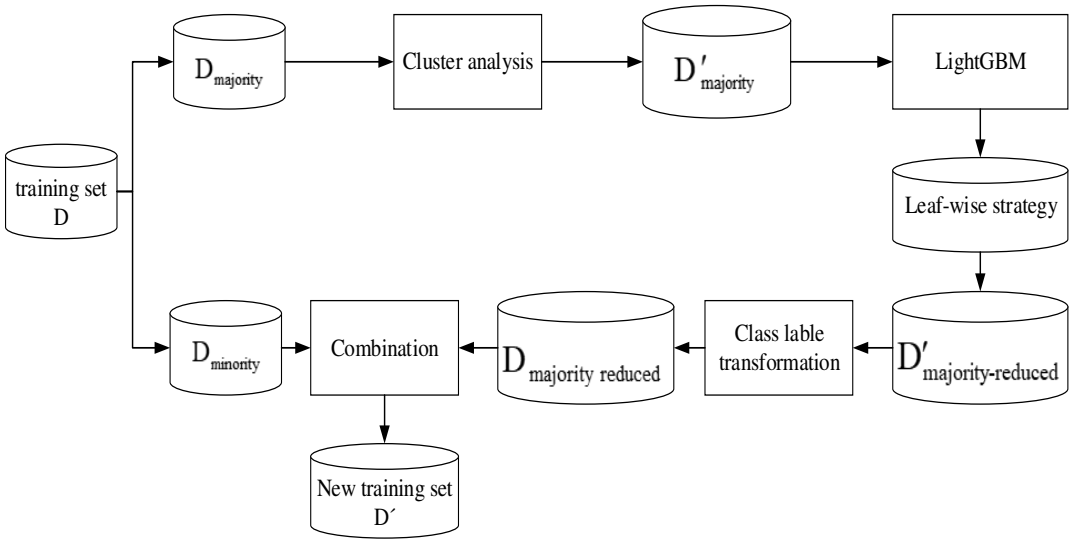

Fig 2. The steps of the proposed CUS-LightGBM approach

The second step is to use $D'_{majority}$ to execute the LightGBM algorithm to create a classifier set. In the classifier training phase and 10 iterations of CUS, the boost algorithm uses the optimal leaf-wise strategy to split leaf nodes from the main class $D'_{majority}$. When growth reaches the same leaf node, some data samples in $D'_{majority}$ are filtered out, which results in a subset of $D'_{majority}$ denoted by $D'_{majority-reduced}$. It contains less samples than $D'_{majority}$.

Next, the k-class label of $D'_{majority-reduced}$ is converted to the original class label of the majority-class dataset, expressed as the less majority-class dataset of $D_{majority-reduced}$.

Finally, $D_{majority-reduced}$ and $D_{minority}$ are combined to generate a new balanced training data set (D').

Since the clustering method helps us select more differentiated data among the data of majority categories (the data in the same cluster is relatively few). The data loss is smaller, and the accuracy is higher than that of the random sampling methods (the data of majority categories are discarded randomly). Also, LightGBM is faster, requires less memory, and is more accurate than XGBoost compared to other existing enhancements. This combines sampling and gradient enhancement approaches to form an efficient class imbalance learning algorithm.

3.3. The related comparison method

AdaBoost reflects the "two heads are better than one" thought, it is an iterative algorithm, and its core idea is to train different weak classifiers for the same training set, and then form a stronger final classifier by combining different training classifiers in order and giving relatively large voting weights to the best performing classifiers, which constructs a high-quality ensemble system. Its specific principle is: first, N sample points on the training data set are given the same weight $(1/N)$, which constitute vector D. These weights are then adjusted according to the adaptive performance of each base learner. If the current base learner does not correctly classify an instance in the training data set, the algorithm will increase the weight of the sample in the next training cycle. In the second training of the classifier, the weight of each sample will be readjusted, in which the weight of the sample accurately classified in the first time will be reduced, while the weight of the sample incorrectly classified in the first time will be increased. The sample set after updating the weights is used to train the next classifier, and the whole training process proceeds iteratively. At the end of each training phase, the total prediction error of the current classifier is calculated based on the sample weights of all the wrong classifications. Then the voting weights of the basic learners in the final set are calculated. Finally, the weighted voting method is used to obtain the final classification results from all the weak classifiers.

GBDT, also known as MART, is an iterative decision tree algorithm, which is composed of multiple decision trees. The results of all the trees add up to make the final output. GBDT = gradient enhancement + decision tree. That is, if each sub-model in gradient enhancement is a decision tree, then the model is GBDT, which has the common features of gradient enhancement and decision tree. The core idea of GBDT algorithm is that in the iterative process, the one-time iterative variables increase the weak classifier $f_1(X)$ one by one, so that the loss

function $L(F_m(X),Y)$ is constantly reduced. That means $L(F_m(X),Y) < L(F_{m-1}(X),Y)$, so that the composite model is finally obtained: $F_m(X) = \partial_0 f_0(X) + \partial_1 f_1(X) + \cdots + \partial_m f_m(X)$. It is a sequentially trained decision tree ensemble model. In each iteration, the decision tree is learned by fitting residuals, and each calculation is made to reduce the residuals of the previous iteration. GBDT builds a new model in the direction of residuals reduction (negative gradient).

XGBoost, recently proposed by Chen, T. and He, T. (2016)(Chen and Guestrin 2016) is an advanced gradient lifting system that USES learning rates, enhancement numbers, maximum tree depth, and subsampling to avoid overfitting. XGBoost algorithm uses regular function $\Omega(f)$ to control over-learning of the model. Its objective function is $min\{Obj\} = min\{L(F_m(X),Y)\} + \Omega(f) + C = min\{\sum_i l(y_i,\hat{y_i}) + \sum_i \Omega(f_t) + C\}$. That is to introduce the definite iterative target into the computer, and then selectively omit the constant term C according to the specific problem of the study. The regular term function is the model of "leaf number T" and "leaf node output ω_j" of a function in each sub-decision tree, which can be expressed as: $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$. With the progress of iteration, the loss function gradually decreases, and the regular term keeps expanding. Then, the objective function is expanded into a Taylor second order series. By using the second derivative of the extreme value, we can find the weak classifier with the minimum objective function. Since the optimal solution that minimizes the objective function is found, the asterisk is used: $f_t(x)^* = -\frac{\sum_{i\epsilon I_t} \partial_{\hat{y}^{(t-1)}} l(y_i,\hat{y}^{(t-1)})}{\sum_{i\epsilon I_t} \partial^2{}_{\hat{y}^{(t-1)}} l(y_i,\hat{y}^{(t-1)}) + \lambda}$, where l_t represents the counting range of the loop, which is set by the algorithm parameters in the above equation. Within the maximum number of variables involved in the iteration, the algorithm will iterate automatically. The loss function makes the target decision tree grow on the layer to fit the data, and the regular function removes the redundant branches with less information through computer iteration. In this way, we can obtain the compound decision tree model with the minimum value of the objective function under the condition of equilibrium, so as to classify and predict the relevant data effectively.

LightGBM is a kind of GBDT, which is proposed to solve the problems encountered by GBDT in massive data, so that GBDT can be applied better and faster in practice. LightGBM is a kind of GBDT, which is proposed to solve the problems encountered by GBDT in mass data, so that GBDT can be applied in practice better and faster. The decision tree sub-model in

LightGBM splits nodes by splitting leaves. As a result, its computing costs are relatively small compared to XGBoost. However, when using this method, the depth of the tree and the minimum data of each leaf node must be controlled to avoid the phenomenon of fitting. LightGBM selects the histogram-based decision tree algorithm to divide the continuous eigenvalues into many small "buckets", then searches for the partitions on these "buckets" and packs them into discrete bins, thus reducing the storage and computing costs(Chen and Guestrin 2016). It takes less memory to use discrete bins to store and replace consecutive values. In addition, it produces more complex trees through the leaf-wise splitting method than the level-wise splitting method, which makes it more accurate than other boosting algorithms. Therefore, the reliability and flexibility of LightGBM will greatly promote the development of credit rating system.

Logistic Regression (LR) is a very widely used classification machine learning algorithm. That is, the output value of the linear model is processed by the sigmoid function, and the output value is positioned between 0 and 1 for the task of binary classification. It can predict not only the category, but also the probability of the event.

Support Vector Machine (SVM) is another credit scoring technique based on artificial intelligence. It is also a dichotomous model, which can now deal with both multivariate linear and nonlinear problems and regression problems after evolution. The core of SVM is to map the original training set to the high-dimensional feature space, in which nonlinear separation features are replaced by linear discriminant functions(Antunes et al. 2017). Then, according to the principle of structural risk minimization, SVM searches for a separated hyperplane to segment the samples. The principle of segmentation is to make the outlier points close to the hyperplane have a larger interval, which is finally transformed into a convex quadratic programming problem to solve, ensuring good generalization ability.

Decision Tree (DT) is a basic classification and regression method, and the decision tree of classification is mainly discussed here. In the classification problem, it represents the process of classifying examples based on features, which can be considered as an if-then set, or as a conditional probability distribution defined on feature space and class space. When a decision tree is used to classify, a feature of an instance is generally tested from the root node. According to the test results, the instance is assigned to its child node and each child node corresponds to

a value of that feature so that the instance is tested and assigned recursively until it reaches the leaf node, and finally the instance is divided into the class of the leaf node. The goal of decision tree learning is to choose the optimal decision tree in the sense of loss function. Specifically, a decision tree model is constructed based on the given training dataset to enable it to correctly classify instances(Karabadji et al. 2019).

Random Forest (RF) is an algorithm that integrates multiple trees through the idea of ensemble learning. Its basic unit is decision tree, and its essence belongs to a big branch of machine learning - ensemble learning method. It not only can process the input samples with high dimensional features, but also does not need to reduce the dimension. In fact, from an intuitive point of view, every decision tree is a classifier (assuming that the classification problem is now targeted), so for an input sample, N trees will have N classification results. In addition, an unbiased estimation of the internal generation error can be obtained during the generation process. The random forest integrates all the classification voting results and specifies the category with the most votes as the final output.

4. Experimental analysis

In this section, we briefly introduce the financial distress prediction dataset for training and testing and analyze the experimental results of the ensemble financial distress prediction model in detail. Finally, some commonly used machine learning classification models are used for comparative analysis.

4.1. Data description

To evaluate the performance of the proposed ensemble classification model, we used the 2017 and 2018 imbalanced financial datasets of small- and medium-sized enterprises in china for training and testing. This dataset can be publicly available from the CSMAR and RESSET databases. We decided to use the small- and medium-sized financial ratio dataset for three main reasons. First of all, large enterprises or enterprises listed on the stock exchange are more likely to get more financial information disclosure. The accounting records of small- and medium-sized enterprises are incomplete, and the true financial status of the enterprise will not be fully reflected in the financial statements. Therefore, for the potential bankruptcy risk that may exist, we established a financial distress prediction model to evaluate the bankruptcy risk of

enterprises with real world loans. Secondly, we want to verify the effectiveness of the proposed method in dealing with imbalanced data. According to the uniqueness of financial distress prediction and the data properties analyzed, it is found that most of the researchers' data indicators extract financial parameters from the annual financial statements of small- and medium-sized enterprises. Because the frequency of financial ratio data used is annual, it may cover significant fluctuations during the two reporting periods, and these data reveal the true ratio between financially distressed enterprises and non-financial distressed enterprises, which facilitates bankruptcy prediction and credit risk assessment. Finally, in the financial distress prediction domain, these small- and medium-sized enterprises datasets are often used to test the performance of classification models, which makes it easy to use this dataset to test the classification performance of our proposed ensemble model and to compare the results with other benchmark models. Table 1 provides a brief description of the dataset.

Table 1
Degree of imbalance of dataset.

Year	Positive Sample	Negative Sample	Imbalanced Proportion
2017	893	47	0.0526
2018	897	43	0.0479

The 2017 financial ratio dataset of small- and medium-sized enterprises contains 940 examples, of which 893 samples were assessed as healthy small- and medium-sized enterprises and 47 samples as poorly run enterprises. A total of 55 financial variables are used to describe applicant's financial distress features. If the loan is approved, the final variable is 0; if the loan is rejected, the final variable is 1. The 2018 financial ratios test set of small- and medium-sized enterprises contains 940 examples, of which 897 samples are marked as creditworthy and 43 samples are marked as having a poor credit rating. For each instance, there are 24 customer credit information variables, and the final binary variables represent the financial forecast evaluation results for each instance. Value 1 indicates a good financial rating and value 2 indicates a bad financial rating. To protect the confidentiality of enterprise information, all identifiable attribute values are changed to meaningless symbols.

4.2. Feature selection

Related research shows that the accuracy and stability of the prediction can be significantly improved by feature selection. Generally speaking, there are two methods to form a small number of important features: feature extraction and feature selection. Both methods can achieve the effect of dimension reduction. They are essentially trying to reduce the attributes of feature dataset, but the two methods are different. The method of feature extraction is mainly through the relationship between attributes, such as combining different attributes to get new attributes, and then changing the original feature space. It is obvious disadvantage is that it is difficult to relate the new features with the original features, and the obtained new feature is not a direct representation of the known financial attribute. The method of feature selection is to select the most relevant features used to predict the target variable from the original feature set and delete the irrelevant features. Therefore, this paper removes irrelevant or redundant feature attributes by feature selection, and then reduces the dimension of space. Feature selection can be solved by three methods, namely filtration, wrapper and embedded(Xia et al. 2018).

Filtration evaluates the divergence or correlation of the attributes according to the data features, that is, giving weight to the features of each dimension. Such a weight represents the importance of the dimensional feature and then is ranked according to the weight. Since we collect discrete variables, which belong to the problem of dichotomy, we use the F test function f_classif in sklearn in this paper. The classification task of f_classif is to divide the sample into n subsets ($S_1, S_2, \cdots, S_n$), we want the mean value of each of these subsets, $\mu_1, \mu_2, \cdots, \mu_n$, to differ. When sample *S* belongs to the positive class, Si takes some specific values (as set *S*+), and when sample *S* belongs to the negative class, Si takes some other specific values (*S*−). We expect the sets *S*+ and *S*− to show great differences, so that feature Si can predict categories better. In the corresponding problem of the analysis of variance, it becomes that we need to test the original hypothesis: $\mu_{S+} = \mu_{S-}$. we want to construct f as large as possible. The larger the value of f, the greater the assurance of rejecting H0. we have more reason to believe that $\mu_{S+} \neq \mu_{S-}$, and more confident that set *S*+ and *S*− show great differences, which means S_i is more helpful to predict the category. In this way, we can judge the help of feature S_i to the prediction category according to the f value of the *i*-th feature Si of the sample. The larger the f value, the stronger the predictive power and the greater the correlation. Therefore, we can select features based on this.

Wrapper is to search the space of attribute subset by enumeration. Its main idea is to treat the selection of subset as a search optimization problem, generating different combinations, evaluating the combinations, and comparing it with other combinations. The most commonly used Wrapper method is the recursive feature elimination algorithm (RFE). RFE is to use a base model for multiple rounds of training. After each round of training, several features of weight coefficients are removed, and then the next round of features are carried out based on the new feature set. In this paper, the most classical SVM-RFE algorithm is used for feature selection. This algorithm uses support vector machine (SVM) to do the machine learning model selection feature of RFE. It in the first round of training, will choose all of the features to train. After obtaining the classified hyperplane $\omega\dot{x} + b = 0$, if there are n features, then RFE - SVM will selects and excludes the feature corresponding to the ordinal i with the least squared value of the component ω_j^2 from ω. In the second category, the number of features is left at (n-1), and we continue to train the SVM with these (n-1) features and output values, similarly, remove the feature corresponding to the minimal ordinal i of ω_j^2, and so on until the remaining number of features meet our need. This method is usually more accurate than the filtration, but it costs a lot of computation.

The main idea of the embedded method is to evaluate the best attributes to improve the accuracy of the model when training the prediction model to avoid overfitting the model. The most common is to use L1 regularization and L2 regularization to select features. In this paper, L1 regularization is used to select features. The larger the L1 regularization penalty, the smaller the coefficient of the model will be. When the regularization penalty term is large to a certain extent, the partial feature coefficient becomes 0, and this partial feature of coefficient 0 can be sifted out. In addition, tree-based feature selection is also used in this paper. This method is based on the gradient lifting algorithm of decision tree. In general, the importance of the feature indicates the role of the feature in constructing the ascending tree. The more times a feature is used as a partitioning attribute in all trees, the more important the feature is. The importance of a single decision tree is calculated by the amount of improved performance measures at each attribute split point and is weighted by the number of observations the node is responsible for. The performance measure can be used to select the purity of the split point or another more specific error function. Finally, the importance of factors is averaged across all decision trees

in the model. Finally, the importance of each feature is obtained, and then features can be sorted or compared.

The initial features of this paper include profitability, operational capacity, solvency, structural ratio, development capacity, risk level, per share index and relative price index. A total of 68 financial index attributes are collected from these eight aspects. It is convenient for small- and medium-sized enterprises to have a less set of features, and these features can be used as the initial index of potential bankruptcy. Therefore, we used feature selection of F function, RFEVC method, recursive feature elimination based on support vector machine estimation, feature selection based on tree and feature selection based on L1 norm to identify the most relevant features, so as to obtain a fewer and more compact group to help people more effectively analyze and evaluate the results. Because of the "no free lunch" theorem, there is no single method that performs better than other methods. To provide robust feature selection, each of these techniques is based on a different theoretical background. In addition, once we know those are the most important indicators of bankruptcy, we can focus on these effective financial attributes and determine the causes of potential bankruptcies. The selected results are shown in table 2.

Table 2
The result of feature selection.

Feature selection method	F-classif	SVM-RFE	L1-based	XGBoost_based	Tree_based	Ensemble choice
Optimal features	F16，F64，F35，F22，F18，F10，F23，F26，F44，F25，F11，F60，F12，F29，F65，F52，F61，F51，F56，F8，F13，F21，F27，F39	F1，F2，F3，F4，F5，F9，F12，F13，F14，F15，F18，F22，F28，F29，F30，F35，F65，F36，F39，F44，F50，F52，F60，F61	F1，F2，F3，F4，F5，F10，F14，F19，F33，F36，F39，F45，F50，F54，F57，F65，F51，F38，F49，F55，F43，F9，F64，F58	F4，F8，F42，F33，F50，F65，F37，F3，F32，F45，F11，F30，F24，F16，F36，F26，F52，F18，F2，F67，F62，F41，F28，F44	F2、F41、F16、F62、F5、F14、F48、F8、F29、F10、F60、F51、F13、F15、F26、F3、F22、F25、F29、F30、F34、F46、F56、F12	F2，F3，F4，F5，F8，F10，F12，F13，F14，F16，F18，F22，F26，F28，F29，F30，F36，F39，F44，F50，F55，F52，F60，F65

We applied the above five feature selection methods to the 2017 financial indicator data, and table 2 shows the results of these five-feature selection. Each method lists the 24 most

important feature attributes according to the importance of the features, with one exception being feature selection of the SVM-RFE method, which obtains only a set of selected attributes without providing the ranking of the features. In addition to the five methods mentioned above, we also selected features based on the votes of all the feature selection methods (ensemble feature selection) (the last column in table 2). Specific ensemble feature selection voting procedures are based on the results of voting selection, the minority is subject to the majority. In addition to the five methods mentioned above, we also selected features based on the votes of all the feature selection methods (integrated feature selection) (the last column in table 2). Specific integrated feature selection voting procedures are based on the results of voting selection, the minority is subject to the majority. Integrating the feature set selected by feature selection enables us to identify a set of features that at least three of the five approaches consider as important attributes for predicting financial distress. Among them, F2, F3, F4, F35 and F65 are selected several times by the five methods, and these indicators have a high degree of confidence in predicting bankruptcy distress. The final column contains the attributes that all the methods selected most frequently for the entire dataset in 2017. Table 3 lists the 24 selected financial ratio attributes and the specific meanings that each attribute represents.

Regarding the functional differences of the selected methods, we need to consider the problem of feature selection instability. It has been proved by many experiments that applying feature selection to slightly varying data can produce different results. Although the output of other methods may vary considerably, some provide relatively stable results. As can be seen from table 2, most of the important characteristics are profitability, solvency and operational capacity. On the other hand, development capacity, structural ratios and related price indicators are not important. Financial ratios in the solvency category imply external sources of capital, such as debt and financial ratios in the profitability category imply lower total returns, which may not be sufficient to repay the liability.

4.3. Evaluation metrics

When datasets are imbalanced, accuracy is affected because of a bias toward majority classes. Therefore, in order to evaluate the performance of machine learning models, it is crucial to evaluate the models effectively and guide classifier learning with appropriate evaluation criteria. There are usually many measurement indicators to choose from, but the evaluation

indicators we choose must reflect the overall predictive performance of the model. Therefore, we adopt the following three methods as the evaluation criteria(Shen et al. 2019).

Table 3
Descriptive statistics of financial indicators.

Type	Financial Ratios	Q25	Median	Q75	Mean
Earning Power	F2: Rate of Return on Total Assets	2.4682	4.4332	7.7642	300.5512
	F3: Net profit margin on current assets	106.506	176.6557	279.5488	242.8734
	F4: Net profit margin on fixed assets	3.4587	5.3609	9.3719	10.7828
	F5: Return on Equity	1.4249	3.0408	7.1280	7.1279
	F8: Operating Margin Ratio	0.3881	0.5791	0.8466	0.6871
	F10: Ratio of Profits to Cost	1.3064	1.8824	2.8686	2.5203
	F12: Earnings Before Interest and Tax	0.6872	1.0735	1.7934	1.5019
	F13: Rate of Return on Investment	0.2008	0.4218	0.8036	0.7497
Operating Capacity	F14: Receivables Turnover Ratio	3.8187	10.2198	58.4725	58.4725
	F16: Operating Cycle	0.2703	0.4332	0.5716	0.4362
	F18: Working Capital Turnover Rate	0.1359	0.2418	0.4369	0.3536
	F22: Equity Turnover	0.0213	0.0416	0.0664	0.0416
	F26: Currency Ratio	0.0315	0.0548	0.0866	0.0548
Debt-paying Ability	F28: Debt Asset ratio	0.1848	0.2766	0.3687	0.2973
	F29: Tangible Debt Asset ratio	0.0382	0.0940	0.1515	0.0961
	F30: Debt equity Ratio	0.0426	0.1062	0.1975	0.9045
Structural Ratio	F36: Fixed Assets Ratio	0.0929	0.1782	0.2860	0.7899
	F39: Current Debt Ratio	0.7864	0.9052	0.9731	3.3301
Development Capability	F44: Net Profit Growth Rate	0.4173	0.1146	0.8903	0.8903
Risk Leve	F50: Operating Leverage	1.1989	1.3864	1.6146	1.6146
	F51: Comprehensive Leverage	1.2834	1.6064	2.2647	2.2647
Index Per Share	F52: Earnings Per Share	0.1194	0.2910	0.5846	0.3797
	F60: UDPPS	0.6072	1.1431	1.9226	1.3699
CKD	F65: Price to Earnings ratio	30.4016	47.6991	98.9254	98.9254

One of the most popular evaluation methods in financial distress prediction is ACC, which is defined as the proportion of correctly classified samples to the test set. Specifically, ACC is defined as the confusion matrix in Table 4.

Table 4
Confusion matrix.

Observed	Predicted		
	Good	Bad	
Good	TP	FN	TP+FN
Bad	FP	TN	FP+TN

Here TP and TN represent the number of good borrowers and bad borrowers of the correct classification, respectively. FP and FN represent the number of bad borrowers and good borrowers misclassified, respectively. The calculation formula of ACC is as follows:

$$ACC = \frac{TP + FN}{TP + FP + FN + TN}$$

Another useful tool for evaluating the effectiveness of a model and visualizing its performance is the ROC curve, which is the optimal decision boundary family for the relative costs of true positives (TP) and false positives (FP).If an instance is marked positive and is also classified as positive, it is called true positive (TP).If an instance is marked negative but classified as a positive instance, it is called a false positive (FP).At each point on the ROC curve, there are a pair of TPR and FPR, corresponding to a certain threshold, which indicates the behavior of the classification model, but does not take into account the category distribution or misclassification costs. In order to compare the ROC curves of different classification models, Tom Fawcett, in 2006, calculated the Area Under the Curve (AUC) value to quantify the ROC performance of the model, which is another recognition capability measure based on the ROC curve. Because it is independent of the selected decision criterion and prior probability, AUC can compare the dominance relationships between the established classifiers. If the ROC curves intersect, the total AUC is the average comparison between models. In general, the greater the AUC value is, the better the performance of the classification model will be. For a perfect model, the AUC value is 1. That is, all positive instances are correctly classified. Instances without negatives are misclassified as positive instances.

There is another performance evaluation indicator that considers the distribution of classes, such as F-measure, which are widely used to solve the evaluation problem of machine learning

classification models. F-measure is an effective evaluation indicator whose value is between 0 and 1, representing a harmonic mean based on accuracy and recall(He et al. 2018). In the F-measure function, F1-measure synthesizes the results of P and R when the parameter $\alpha = 1$, and the test method is more effective when F1-measure is higher. The mathematical formula of f1-measure is shown as follows:

$$\mathrm{F1-M}easure = \frac{2 \times Precision \times TPR}{Precision + TPR}$$

$\mathrm{F1-M}easure = \frac{2 \times Precision \times TPR}{Precision + TPR}$ represents the proportion of TP samples in the total number of predicted positive samples. $TPR = \frac{TP}{TP+FN}$ is the "true case rate" representing the ratio of true positive instances to all positive instances(Raghuwanshi and Shukla 2018). Therefore, in order to fully understand the features of the algorithm, this study used three performance evaluation indicators: ACC, F1-measure and AUC.

4.4. Experimental results

Table 5 and figure 3 list three groups of experiments using traditional machine learning methods, ensemble learning methods, and machine learning methods based on clustering under-sampling and ensemble to measure the performance of the new CUS- LightGBM. From the first set of experiments, it can be seen that the performance of the traditional four model methods is obviously lower than that of the ensemble learning method, and the performance of the DT model is better than that of KNN, SVM and LR. From the second group of experiments, it can be seen that AdaBoost has low performance, with an AUC of 0.77, followed by GBDT, and LightGBM has a better prediction method and stable performance. From the experimental results of the third group, the performance of the ensemble learning method based on clustering under-sampling is obviously better. Namely CUS-LightGBM method based on CUSBoost is superior to CUSBoost and has the best performance. And the overall performance is higher than the eight comparison experiments(Du et al. 2020). The results show that the proposed CUS-LightGBM method is superior to the 9 methods listed in the comparison experiment, regardless of the technical combination. Although the performance of the CUSBoost is similar to that of the proposed best ensemble method CUS-LightGBM, the AUC and ACC of CUS-Light GBM were 0.9919 and 0.9130, respectively, which is better than CUS boost.

Table 5

Comparative experimental results.

Model	ACC	F1-Measure	AUC
LR	0.6722	0.5333	0.6434
DT	0.8964	0.8723	0.6963
SVM	0.7878	0.3618	0.7640
RF	0.7922	0.4084	0.7752
AdaBoost	0.7755	0.6281	0.7499
XGBoost	0.9436	0.6543	0.8033
GBDT	0.9403	0.7496	0.8016
LightGBM	0.9425	0.6506	0.8068
CUSBoost	0.9467	0.8009	0.9654
CUS-LightGBM	0.9130	0.8165	0.9919

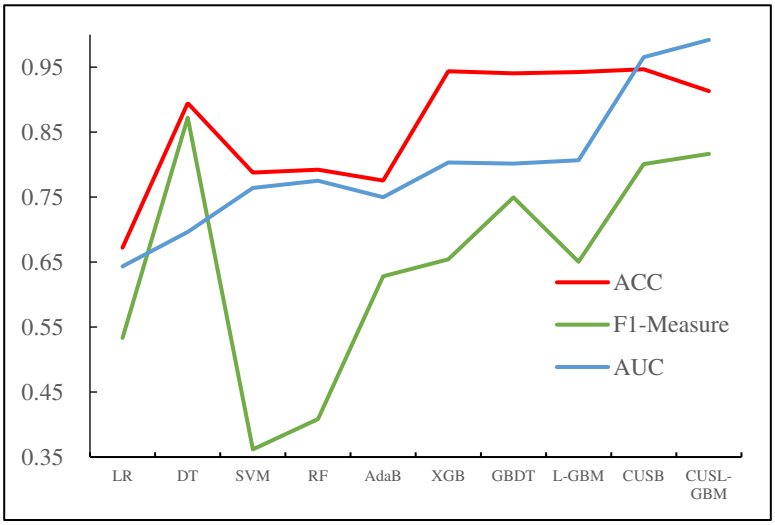

Fig. 3. Performance of proposed models is compared.

In summary, the performance of the nine models in this study, all of which were statistically significant. Therefore, they are useful for measuring the financial distress of small- and medium-sized enterprises in China. AUC is directly used to estimate the prediction accuracy of the model. In terms of measuring the usefulness of financial distress, the ROC value of the best model is 1. Other models are generally better, but the traditional artificial intelligence methods DT, KNN, SVM and LR show that the prediction ability of default probability is low. In addition, the results for AdaBoost, XGBoost, and XGBoost are similar to those for LightGBM. Three experiments show that the model CUS- LightGBM in the third experiment is the best model, which is a good choice for enterprise managers, bank administrators and

auditors. That is, when the effective financial ratio variable is added to the model, the value of the influence factor on default decreases.

4.5. Discussion

Based on small- and medium-sized enterprises listed enterprises as the research object, according to the principles of comprehensiveness, ease of availability and effectiveness, from the eight aspects in Table 3, this paper fully reflects the enterprise's financial condition and operation management of financial indicators to construct the index system of financial ratios, takes preliminary selection of 68 financial ratios as an indicator of financial distress prediction, and by 5 feature selections selects 24 financial distress indicators which may have important effects on small- and medium-sized enterprise. Then, ST standard is used to define financial distress. The feasibility and effectiveness of the CUS-LightGBM method in estimating and predicting financial distress model are studied.

From the perspective of social practice, our research has important implications. It can be seen from the results after feature selection that liquidity index is one of the key indicators of financial distress prediction of the small- and medium-sized enterprises, which tells us that the financial managers of the small- and medium-sized enterprises in the growing stage should keep and strictly track the liquidity of the enterprise, and shorten the time of payment delay as much as possible. It should also be recognized that allowing customers to delay payments for more than 30 days increases the risk of bankruptcy for the enterprise, Although the enterprise performs well and is still at risk of financial distress for a certain period of time, because once the enterprise's liquidity level reaches a critical value and begins to lead to late payments, it is too late to adjust the payment terms. Therefore, the financial manager should fully explain the terms of the payment as soon as possible and set a certain proportion of the pure advance payment terms and so on.

Compared with the internal control of enterprises, the bank policy has even higher influence on small- and medium-sized enterprises, and the bank credit limit significantly affects the small- and medium-sized enterprises' business bottleneck(Schwab et al. 2019). To put it simply, Banks tightened restrictions on credit services of small- and medium-sized enterprises after the financial crisis with the aim of reducing the risk of bank failure. Some research evidence shows that legal restrictions will harm small- and medium-sized enterprises ' access

to financing, thus restricting their development. In this paper, another method is found for bank managers to predict the financial status of target credit method enterprises, so that banks can adjust their loans according to the financial health of customers. This model can help enterprises calculate bankruptcy risk and help identify the financial position of a small- and medium-sized enterprises.

The main purpose of comparing the traditional AI method and ensemble method with our proposed method is to evaluate the effectiveness of the CUS-LightGBM approach and to achieve a good balance between accuracy and complexity. First of all, among the existing methods, CUS-LightGBM method is an advanced technology. Compared with the traditional artificial intelligence method and ensemble method, CUS-LightGBM has shown better performance in multiple experiments. Experimental results show that CUS-LightGBM can improve the performance of the model under the same environment. Our experiments also show that the CUS-LightGBM method performs better than the benchmark method without excessive computational costs. Second, CUS-LightGBM is more effective and accurate than other Boosting tools. CUS-LightGBM is faster, requires less memory, and is more accurate than RUSBoost. In addition, experiments show that CUS-LightGBM can achieve linear acceleration by using multiple machines for specific training. These features indicate that the proposed model is suitable for practical application. Third, from a cost-effective point of view, even if 1% of the answer is wrong, it will bring huge losses to financial institutions. therefore, the ultimate goal of the financial distress prediction research is to improve the predictive ability of the model, even at the cost of large computational and interpretability losses. Therefore, a well-adjusted and accurate financial distress prediction model can show long-term benefits.

5. Conclusion and future work

This paper proposes a new ensemble prediction model based on LightGBM and clustering under-sampling technique, which can be used to predict the financial distress of small- and medium-sized enterprises with imbalanced data to obtain better prediction performance. The proposed model has two main contributions. Firstly, five feature selection methods are used to remove redundant or irrelevant feature attributes, which improves the accuracy and stability of model prediction. Secondly, the clustering under-sampling method is used to solve the

imbalanced data problem, and the advanced LightGBM is combined to establish a new ensemble model, and the performance of the model is evaluated with three comprehensive evaluation indexes, Accuracy, F-measure and AUC. By comparing the results, the proposed ensemble classification model is better than the other nine models for small- and medium-sized enterprises' financial attribute dataset, which indicates that the model is more practical in financial distress prediction problem. In the practical application of financial risk assessment, the problem is divided into two categories, one is the enterprise with good financial rating, the other is the enterprise with poor financial rating. the proposed method obtained the highest AUC score in this task. Experimental results show that the ensemble strategy combining LightGBM classifier and clustering under-sampling technique is better than the single classification model discussed in this paper and ensemble model without imbalanced processing.

In the future, our research will consider using data from different countries to develop financial distress prediction models for small- and medium-sized enterprises to further improve the effectiveness and feasibility of the models. In addition, we will consider using other feature selection methods to preprocess the training data, such as outlier testing method. Other AI methods are also considered to explore the evaluation capabilities of potential hybrid models.

CRediT authorship contribution statement

Sumei Ruan: Conceptualization, Writing - Review & Editing, Validation, Funding acquisition.

Jiayong Zhang: Writing - Original Draft, Software.

Wei Li: Supervision, Methodology, Funding acquisition.

Compliance with ethical standards

Conflict of interest - All authors have no conflict of interest.

Ethical approval - This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent - Informed consent was obtained from all individual participants included in the study.

Acknowledgments

This work was funded by the Key Project of Philosophy and Social Sciences Planning in Anhui Province [No. AHSKZ2018D14], Key Projects of Natural Science Research of Universities in Anhui Province [No. KJ2019A0651; No. KJ2020A0008] and Natural Science Foundation of Anhui Province [No. 2008085MG234].

References

Antunes F, Ribeiro B, Pereira F (2017) Probabilistic modeling and visualization for bankruptcy prediction. Appl Soft Comput J 60:831–843. https://doi.org/10.1016/j.asoc.2017.06.043

Bao L, Juan C, Li J, Zhang Y (2016) Boosted Near-miss Under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. Neurocomputing 172:198–206. https://doi.org/10.1016/j.neucom.2014.05.096

Blagus R, Lusa L (2017) Gradient boosting for high-dimensional prediction of rare events. Comput Stat Data Anal 113:19–37. https://doi.org/10.1016/j.csda.2016.07.016

Boyu W, Joelle P (2016) Online Bagging and Boosting for Imbalanced Data Streams. IEEE Trans Knowl Data Eng 28:3353–3366. https://doi.org/10.1109/TKDE.2016.2609424

Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785

Cheng F, Zhang J, Wen C (2016) Cost-Sensitive Large margin Distribution Machine for classification of imbalanced data. Pattern Recognit Lett 80:107–112. https://doi.org/10.1016/j.patrec.2016.06.009

Du X, Li W, Ruan S, Li L (2020) CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection. Appl Soft Comput J 97:. https://doi.org/10.1016/j.asoc.2020.106758

Gong J, Kim H (2017) RHSBoost : Improving classification performance in imbalance data. Comput Stat Data Anal 111:1–13. https://doi.org/10.1016/j.csda.2017.01.005

Haixiang G, Yijing L, Yanan L, et al (2016) BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. Eng Appl Artif Intell 49:176–193. https://doi.org/10.1016/j.engappai.2015.09.011

He H, Zhang W, Zhang S (2018) A novel ensemble method for credit scoring : Adaption of different imbalance ratios. Expert Syst Appl 98:105–117. https://doi.org/10.1016/j.eswa.2018.01.012

Hu YC (2020) A multivariate grey prediction model with grey relational analysis for bankruptcy prediction problems. Soft Comput 24:4259–4268. https://doi.org/10.1007/s00500-019-04191-0

Huang X, Liu X, Ren Y (2018) Enterprise credit risk evaluation based on neural network algorithm. Cogn Syst Res 52:317–324. https://doi.org/10.1016/j.cogsys.2018.07.023

Karabadji NEI, Khelf I, Seridi H, et al (2019) A data sampling and attribute selection strategy for improving decision tree construction. Expert Syst Appl 129:84–96. https://doi.org/10.1016/j.eswa.2019.03.052

Ke G, Meng Q, Finley T, et al (2017) LightGBM : A Highly Efficient Gradient Boosting Decision Tree. Adv Neural Inf Process Syst 30 (NIPS 2017) 3149–3157. https://doi.org/10.1145/1731903.1731925

Kruppa J, Schwarz A, Arminger G, Ziegler A (2013) Consumer credit risk: Individual probability estimates using machine learning. Expert Syst Appl 40:5125–5131. https://doi.org/10.1016/j.eswa.2013.03.019

Li K, Niskanen J, Kolehmainen M, Niskanen M (2016) Financial innovation: Credit default hybrid model for SME lending. Expert Syst Appl 61:343–355. https://doi.org/10.1016/j.eswa.2016.05.029

Ma X, Sha J, Wang D, et al (2018) Electronic Commerce Research and Applications Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to di ff erent high dimensional data cleaning ☆. Electron Commer Res Appl 31:24–39. https://doi.org/10.1016/j.elerap.2018.08.002

Maurya CK, Toshniwal D, Vijendran Venkoparao G (2016) Online sparse class imbalance learning on big data. Neurocomputing 216:250–260. https://doi.org/10.1016/j.neucom.2016.07.040

Ng WWY, Zeng G, Zhang J, et al (2016) Dual autoencoders features for imbalance classification problem. Pattern Recognit 60:875–889. https://doi.org/10.1016/j.patcog.2016.06.013

Olson DL, Delen D, Meng Y (2012) Comparative analysis of data mining methods for bankruptcy prediction. Decis Support Syst 52:464–473. https://doi.org/10.1016/j.dss.2011.10.007

Papouskova M, Hajek P (2019) Two-stage consumer credit risk modelling using heterogeneous ensemble learning. Decis Support Syst 118:33–45. https://doi.org/10.1016/j.dss.2019.01.002

Raghuwanshi BS, Shukla S (2018) Class-specific kernelized extreme learning machine for binary class imbalance learning. Appl Soft Comput J 73:1026–1038. https://doi.org/10.1016/j.asoc.2018.10.011

Rayhan F, Ahmed S, Mahbub A, et al (2017) CUSBoost: Cluster-Based Under-Sampling with Boosting for Imbalanced Classification. In: 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS). IEEE, pp 1–5

Schwab L, Gold S, Reiner G (2019) Exploring financial sustainability of SMEs during periods of production growth: A simulation study. Int J Prod Econ 212:8–18. https://doi.org/10.1016/j.ijpe.2018.12.023

Shen F, Zhao X, Li Z, et al (2019) A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. Phys A Stat Mech its Appl 526:. https://doi.org/10.1016/j.physa.2019.121073

Sigrist F, Hirnschall C (2019) Grabit : Gradient tree-boosted Tobit models for default prediction. J Bank Financ 102:177–192. https://doi.org/10.1016/j.jbankfin.2019.03.004

Sun J, Lang J, Fujita H, Li H (2018) Imbalanced enterprise credit evaluation with DTE-SBD : Decision tree ensemble based on SMOTE and bagging with differentiated sam pling rates. Inf Sci (Ny) 425:76–91. https://doi.org/10.1016/j.ins.2017.10.017

Sun J, Li H, Fujita H, et al (2020) Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Inf Fusion 54:128–144. https://doi.org/10.1016/j.inffus.2019.07.006

Tsai CF, Lin WC, Hu YH, Yao GT (2019) Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Inf Sci (Ny) 477:47–54. https://doi.org/10.1016/j.ins.2018.10.029

Verikas A, Kalsyte Z, Bacauskiene M, Gelzinis A (2010) Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: A survey. Soft Comput 14:995–1010. https://doi.org/10.1007/s00500-009-0490-5

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1:67–82. https://doi.org/10.1109/4235.585893

Xia Y, Liu C, Da B, Xie F (2018) A novel heterogeneous ensemble credit scoring model based on bstacking approach. Expert Syst Appl 93:182–199. https://doi.org/10.1016/j.eswa.2017.10.022

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

Zoričák M, Gnip P, Drotár P, Gazda V (2020) Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets. Econ Model 84:165–176. https://doi.org/10.1016/j.econmod.2019.04.003