The Journal of Supercomputing (2019) 75:862–884
https://doi.org/10.1007/s11227-018-2619-8

Transfer learning-based default prediction model for consumer credit in China

Wei Li[1,2] · Shuai Ding[1,2] · Yi Chen[1,2] · Hao Wang[1,2] · Shanlin Yang[1,2]

Published online: 22 September 2018

Abstract

Financial institutions in China, such as banks, are encountering competitive impacts from Internet financial businesses. To address these impacts, financial institutions are seeking business innovations, such as an automatic credit evaluation system that is based on machine learning. Abundant new credit data are required in the implementation of new businesses to establish related risk evaluation models; however, new businesses lack data. Based on these insights, this paper innovatively proposes the idea of transfer learning, determines the similarity between traditional businesses and new businesses and transfers the data of traditional bank businesses to new business data to construct new training sets and to train small data sets. The reconstructed training data sets are used to train default risk prediction models, compare them with the benchmark models in the tests and validate the performance and adaptation of the default prediction model based on transfer learning technique. Our study highlights the commercial value of the transfer learning concept in the financial risk field and provides practitioners and management personnel with a decision basis.

Keywords Default prediction · Transfer learning · Consumer credit · Small sample · Data driven

✉ Shuai Ding
dingshuai@hfut.edu.cn

✉ Shanlin Yang
yangsl@hfut.edu.cn

Wei Li
lw5522010@163.com

Yi Chen
chenyihfut@163.com

Hao Wang
waynehfut@mail.hfut.edu.cn

1 School of Management, Hefei University of Technology, Hefei 23009, Anhui, China

2 Key Laboratory of Process Optimization and Intelligent Decision-Making (Ministry of Education), Hefei University of Technology, Hefei 23009, Anhui, China

1 Introduction

The consumer market in China has a bright future. The gross retail sales of social consumer goods of China in 2017 were RMB 36,630 billion, which increased by 10.2% compared with that of the previous year. As the absolute scale continues to increase, the per capita disposable income of urban residents also increases annually, which provides a solid foundation for a consumption upgrade. Consumer credit maintains rapid growth in China. The data in the "2017 China Consumer Credit Market Development Report" indicate that the market scale of consumer credit (excluding mortgage loans) was RMB 9800 billion at the end of 2017, which is 12.32% of the gross domestic product (GDP). Although the market has a bright future, traditional banks and financial institutions will unavoidably face an increasing amount of competition after the implementation of consumer credit business, especially small-scale banks and financial institutions. With an increase in the consumer credit market scale, this competition is intensifying in this industry. Consumer finance enterprises with licenses, peer-to-peer (P2P) platforms and Internet financial enterprises that are based on electronic commerce are continuously joining this industry, which will create competition with traditional banks and financial institutions. An increasing number of traditional financial institutions have established automatic risk control mechanisms based on machine learning using financial technology to improve the traditional artificial risk identification mode. By introducing a novel machine learning algorithm, the data become an important asset of banks and financial organizations, and data analysis becomes an important and revolutionary approach to thinking about reform. The algorithm will ultimately drive the transformation of the bank business mode and realize intelligent risk control. Therefore, data-driven risk control models based on advanced machine learning algorithms have realistic significance.

Consumer credit in consumer finance refers to a small cash loan that is based on technical means, such as the Internet, and features no mortgage, no guarantee and no scenario. Generally, the loan period is three to six months. The traditional consumer credit default risk prediction model, which is based on machine learning, can only effectively run when it satisfies the following assumed conditions. A suitable classification model can be learned from a sufficient number of workable training samples, and the training and testing data are obtained from the same feature space and the same distribution. However, data samples for training are lacking in the implementation of the consumer credit business and the establishment of a default risk prediction model for a real application environment. When the data distribution changes, a substantial number of learning models will need to be reestablished using the recollected training data. This process is costly, and collecting all required training data and reestablishing the prediction model in real financial applications is difficult. To solve this problem of insufficient training samples, namely the small sample issue, credit card data samples that are similar to consumer credit data samples can be employed. By discovering the similarity between two samples [1], the data in the auxiliary training sample sets can be transferred to the objective training sample set via the transfer learning mechanism. Based on the previously mentioned content, the constraint conditions of the learning model should be reduced as much as possible to ensure that the learning model can

effectively adapt to the training data. Therefore, transfer learning between the data fields is required in this case.

The traditional default prediction methods [2] include the statistical analysis method and the machine learning method, which have been successfully applied, e.g., Z-score model, ZETA model and logit model [3]. These three credit evaluation methods aim to establish the association between the credit level and the influence factors and then to accurately assess the credit conditions [4]. The method based on statistical analysis features has a long development history, relatively mature technology and an extensive application [5]. In addition, the default prediction aims to assess the risk level of the credit customers and predict the default possibility [6]. When the default prediction issue is transformed to a binary classification issue, the default prediction issue can be easily solved. To improve the model design defects in the traditional statistical analysis method and its insufficient nonlinear processing capabilities, machine learning methods are introduced for the default prediction issue [7]. For example, a support vector machine (SVM) is used to generate the determination and analysis functions [8]. However, these functions are not affected by assumptions and have fewer restrictions. Multiple machine learning models are used as individual classifiers, and default prediction methods, such as bagging, boosting and the random forest model, are integrated. The combination of these machine learning algorithms has a bright future. In addition, the integration of individual classifiers (logic regression, decision tree, artificial neural network (ANN) and SVM) provides better prediction performance than the implementation of a single classifier [9, 10].

Issues at the data level, such as scarce samples, imbalanced classes and high-dimensional features, will reduce the performance of the default risk prediction model [11]. In particular, the research and development (R&D) aspect of a new financial credit product can be expected to be supported by a massive amount of data. Different testing models need to be established, especially the construction of the default risk prediction model. Although a massive amount of training data are required to establish an excellent default risk prediction model, new businesses lack massive data sets or model training frameworks. Thus, the performance when implementing the traditional default risk prediction model may be poor. Based on this analysis, the default prediction model of business B is established by learning the data of business A (which indicates the credit card business as the auxiliary training data set) and business B (which indicates the consumer credit business explored in this paper as the objective training data set) according to the idea of transfer learning. Although business A differs from business B, they are associated. The knowledge of business A (it indicates data) can be transferred to business B using the transfer learning technique, which can enhance the performance of the default prediction model of business B. The classification results of three individual classifiers can be fused to improve the prediction performance of the prediction model via the ensemble learning strategy.

The remainder of this paper proceeds as follows: In Sect. 2, we provide a simple review the transfer learning technique and default risk prediction model. In Sect. 3, we present the prediction model and data based on the transfer learning technique. In Sect. 4, we discuss the model prediction results. In Sect. 5, we present final comments, discuss the implications of the study, including the strengths and weaknesses of the paper, and offer suggestions for future research.

2 Related works

The default model is designed using the initial financial ratio, within the range of the credit risk management. These financial ratios are calculated using the data on the balance sheet and income account [12]. The ratios reflect their availability and standardization degree prediction capabilities. Generally, they can distinguish the default enterprises from the nondefault enterprises [3]; thus, they are easily obtained and homogeneous because they are calculated using a given supervision framework in a similar manner. Other variants can improve the model accuracy. With the exception of the financial method, however, the remaining dimensions are seldom utilized [13]. Therefore, the accounting-based model dominates default prediction. However, the main weakness of the designed models is that they apply few variant estimates [12, 13].

The main idea of the default prediction task is to establish a quantified model and predict or assess the credit level of the loan customers according to a group of explanation variants [14]. The task aims to estimate the default probability and can be regarded as a general classification task. In past decades, different classification algorithms have been applied to study default prediction based on traditional statistical methods or machining learning technology [15]. An individual classifier primarily establishes a credit scoring model using a statistical method or a machine learning approach. The statistical methods include linear discriminant analysis, multiple discriminant analysis, logic regression or Bayesian networks. However, numerous studies show that the machine learning approach has a prediction precision that is higher than that of the traditional statistics method. These methods include neural networks [16], decision trees [17], SVMs [18], genetic algorithms [19] and naive Bayes classifiers [20]. The decision tree has been extensively applied in the construction of classification models because it is similar to the reasoning process invoked by humans and is easily understood. Sun et al. proposed a new decision ensemble model based on the synthetic minority oversampling technique (SMOTE) and a bagging ensemble algorithm to process the imbalanced enterprise credit assessment. This algorithm is better than other algorithms [21]. SVMs do not require a prior hypothesis and can solve the high-dimensional data issue. High-dimensional data have been extensively applied in the credit scoring domain [22]. With its multilayer network and nonlinear transfer function, the neural network is more discriminatory than logistic regression and other statistical methods and demonstrates its excellence in the establishment of the credit scoring model [23]. The basic principle and optimization function of the machine learning method is diversified.

Numerous other methods based on multiple regression [24] have been extensively applied. If a debtor evaluates the risks, the default prediction technology is very important. The traditional approach to evaluate the default probability employs logic regression in bank departments. Recently, researchers who study default probability are primarily evaluating machine learning approaches (closely associated with statistical data). Their feature selection capability eliminates some features with minimal prediction capabilities, reduces dimensionalities of the feature space and deletes unrelated data. Boosting is a machine learning algorithm that can reduce the variance and deviation in supervised learning [25]. Boosting and feature selection can be combined

as one method for default prediction [10]. In addition, the prediction models applied in the empirical literature can be divided into parametric models and nonparametric models [18]. The linear regression model is a popular parametric model that is robust and effective in the prediction and explanation of default risks [26]. Conversely, nonparametric methods are more flexible in default prediction models. In particular, no prior hypothesis is required when nonparametric methods are used to fit a regression model [27].

The structures and features of the credit data have significantly changed in past years; thus, traditional prediction methods cannot effectively solve encountered problems [28]. With the presentation of the transfer learning framework [29], some computational intelligence approaches, such as neural networks, Bayes networks and fuzzy logic, have been applied in real applications. These methods have been successfully applied in fields such as natural language processing, computer vision and biologics [30]. In addition, they are also applied in finance and commerce management fields [31]; however, related scientific research seldom occurs. Therefore, the transfer learning technique is chosen to solve the insufficient training data set problem and effectively improve the performance of the prediction model in credit risk modeling.

3 Methods and materials

We introduce the credit data set and the proposed method in this section. We predict the customer's default level for individual consumer credit according to two different credit data sets. We establish the default prediction model of business B by learning the credit card business (represented as business A) and the consumer credit loan business (represented as business B). Although business A completely differs from business B, they are associated. The transfer learning technique is used to transfer the knowledge of the business A (which indicates the data, and the association/associated characteristics between the data can be discovered by computing the similarity between data/features) to business B. In this paper, to enhance the default prediction model of business B, three classifiers are trained, and the classification results of a single classifier are fused by an ensemble learning strategy to improve the performance of the prediction model. We compare the default risk prediction model based on the transfer learning framework and other traditional default prediction models based on machine learning techniques without processing the transfer learning technique. The flowchart of this paper is shown in Fig. 1.

3.1 Credit data set

The experimental data in this paper come from a consumer finance company in China, in which the credit data set contains three data sets: A_train.csv data file, B_train.csv data file and B_test.csv data file. The credit card business (A_train.csv) consists of credit loans that belong to noncash transaction payment methods. These payment methods are characterized by credit loans in which the debtor does not need to provide collateral and can obtain loans with only his credit, and the credit level of the borrower

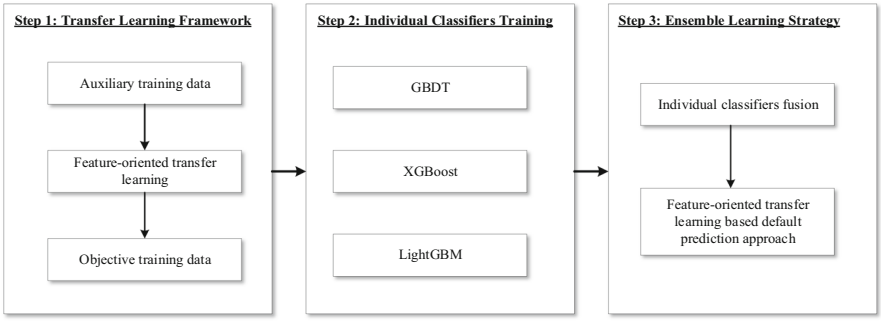

Fig. 1 General framework for default prediction model based on transfer learning

serves as the repayment guarantee. The maximum period generally does not exceed 50 days. The consumer credit business (B_train.csv) is also a credit loan; it issues commodity currencies to consumer credit users in an unsecured manner, which is similar to the credit card business in some aspects. However, the consumer credit business differs from the credit card business in several aspects: First, the maximum amount of consumer credit may be RMB 200,000, whereas credit cards generally range from thousands to tens of thousands; second, the manner of use is different. Credit cards can only be used for shopping, whereas consumer credit can be used to withdraw money for multidimensional consumption and purchase behavior. Third, the interest calculation methods are different. Credit card users can enjoy an interest-free discount as long as they repay the credit card balance within the stipulated time, whereas consumer credit borrowers need to repay their loans on time every month, and the interest is calculated from the effective date of the loan. Due to the correlation between the two services, knowledge of credit card services (referred to herein as data) is transferred to consumer credit services via the transfer learning technique to enhance the default risk prediction model for consumer credit services. Specifically, the data in A_train.csv, including features and labels, consist of 40,000 credit samples and 490-dimensional features with one-dimensional labels. The loan period is 30–50 days, and the average loan amount is several thousand to tens of thousands of medium credit loan business A training data. The data in B_train.csv, including features and labels, consist of 4000 credit samples, 490-dimensional features with one-dimensional labels. The loan period is three to six months, and the average loan amount is tens of thousands to hundreds of thousands of credit loans. B_test.csv is the testing data of 13,463 consumer credit businesses with 490-dimensional features.

3.2 Transfer learning technique

This paper proposes the default risk prediction method based on the feature-oriented transfer learning using three individual classifiers based on the tree structure to predict the default level of consumer credit customers. The transfer learning technique is also referred to as inductive transfer and field adaptation, which is an important area of study in the machine field [32]. Transfer learning aims to apply knowledge or pattern

learning in a field or task to a different related field or problem. Transfer learning facilitates learning by an analogy similar to learning in humans, e.g., learning to walk skills can be used to learn to run, and learning to identify a sedan can be used to learn to identify trucks. The core of transfer learning is to determine the similarity between the source domain and the target domain and reasonably utilize it [33]. The target domain indicates the consumer credit business data to study in this paper, which is also referred to as the objective data. The source domain indicates the credit card business data and is also referred to as auxiliary data. The consumer credit business data are insufficient and cannot be used to train an effective default prediction model; thus, similar data in the credit card business are transferred to form a new training data set and assist the training of the prediction model [34]. Therefore, this paper aims to determine the similarity between two data sets, transfer partial data in the credit card business to the consumer credit business via the feature-oriented method and form a new training data set. This method can enhance the robustness of the training data set and avoid underfitting.

First, the data in the credit card business and consumer credit business are transferred and fused to form the new training data set for data preprocessing and feature engineering. Second, the extreme gradient boosting (XGBoost) model is used to score the importance of the features and select suitable feature combinations for the prediction model. Third, the new training data set is used to train the gradient boosting decision tree (GBDT), XGBoost and light gradient boosting machine (LightGBM) classifiers based on the tree and optimize the optimal parameter combination of the prediction model. Last, the trained GBDT, XGBoost and LightGBM individual classifiers are fused via an ensemble learning strategy.

3.3 Transfer learning-based default prediction model

Before individual classifiers are trained, the similarity between the sample data from business A (credit card business) and business B (consumer credit business) is calculated using the dot product to transfer the auxiliary training data to the objective training data using the feature-oriented transfer learning technique; the new training data set is constructed, and the prediction method in this paper implements it to accurately and effectively predict the default level of credit customers [35].

If vector $\mathbf{a} = [a_1, a_2, \ldots, a_n]$ and vector $\mathbf{b} = [b_1, b_2, \ldots, b_n]$, then the dot product equation of $\mathbf{a}$ and $\mathbf{b}$ is

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

The one-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ have the same rows and columns.

The dot product can geometrically represent or calculate the angle between two vectors and the projection of vector b in the direction of vector a. The equation is described as follows:

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$$

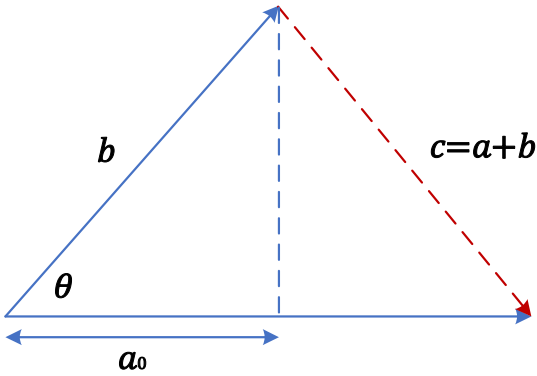

Fig. 2 Geometric representation of the dot product

Table 1 Relationship between two vectors

$\mathbf{a} \cdot \mathbf{b}$	θ	Angles	Orientation
> 0	$0° < \theta < 90°$	Acute angle	Unidirectional
$= 0$	$\theta = 90°$	Right angle	Perpendicular
< 0	$90° < \theta < 180°$	Obtuse angle	Contrary

In Fig. 2, $\mathbf{c} = \mathbf{a} - \mathbf{b}$. According to the triangular cosine theorem, we can obtain $\mathbf{c}^2 = \mathbf{a}^2 + \mathbf{b}^2 - 2|\mathbf{a}||\mathbf{b}| \cos\theta$. According to the relation $\mathbf{c} = \mathbf{a} - \mathbf{b}$ ($\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ indicate the vectors), we can obtain

$$(a - b) \cdot (a - b) = a^2 + b^2 - 2a \cdot b = a^2 + b^2 - 2|a||b| \cos\theta$$

Namely,

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$$

The length of the vectors $\mathbf{a}$ and $\mathbf{b}$ are the known values to calculate; thus, the angle θ between vector $\mathbf{a}$ and vector $\mathbf{b}$ is

$$\theta = \arccos\left(\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}\right)$$

Based on this equation, the angle between vector a and vector b can be calculated. Thus, we can determine whether two vectors have the same direction or are orthogonal to each other (perpendicular). For the specific corresponding relation, refer to Table 1.

If the cosine angle (θ) calculated for the feature similarity between two samples satisfies the condition $0° < \theta < 90°$ in this paper, a strong similarity exists between the two samples, and the data in the two samples can be transferred to form new training data according to the transfer learning technique. If the cosine angle (θ) calculated for

the feature similarity between two samples satisfies the condition $90° \leq \theta < 180°$ in this paper, a weak similarity exists between the two samples. If the data are transferred, a negative transfer will occur. These data in business A shall be deleted and will not be transferred to business B to avoid constructing an inaccurate prediction method due to the data.

The GBDT, XGBoost and LightGBM machine learning models serve as the individual classifiers in the construction of the default risk prediction method for the feature-oriented transfer learning in this paper [36]. Machine learning methods based on ensemble learning have been shown to perform better than a single classifier, as indicated in several references [22, 28, 37, 38]. Boosting is an ensemble learning method that belongs to classification algorithms. The weak classifiers can be enhanced to become strong classifiers and realize precise classification by training. The weak classifier is the submodel generated in each iteration. The strong classifier is the final prediction model. After each iteration, the classifier generated in each iteration will be added to the final model with a certain weight. The GBDT is the gradient boosting decision tree; it is an extensively applied algorithm in the current machine learning field. For XGBoost, to efficiently implement the GB algorithm, the classification and regression tree (CART, gbtree) or a linear classifier (gblinear) can be applied as the base learner. LightGBM is an open-source algorithm that was released by the DMTK team of Microsoft Research Asia, which implements the GBDT algorithm framework and features quick training and low memory consumption. The three classifiers are introduced as follows.

3.3.1 GBDT

The decision tree is a basic classification and regression method that features quick classification and visual explanation of the model; however, it sometimes causes overfitting. Although the decision tree is pruned, the result is not satisfactory. The boosting method can be used to learn multiple classifiers, linearly combine these classifiers and improve the classifier performance by changing the weights of the training samples in the classification. (The weights of incorrectly classified samples are increased, and the weights of the correctly classified samples are decreased.) Decision trees are combined with boosting to generate a variety of algorithms, including the boosting tree and the GBDT. The GBDT can implement a data classification or regression algorithm using the additive model (namely a linear combination of base functions) and continuously decreasing the residual errors in training [39].

After multiple iterations of the GBDT, a weak classifier is generated for each iteration. Each classifier is trained based on the residual errors of the previous classifier. Generally, the weak classifiers are sufficiently simple and include low variance and high deviation because the precision of the final classifier is continuously improved by decreasing the deviation. The CART is selected as the weak classifier. Each classification regression tree will not be too deep due to the previously mentioned high deviation and simplicity requirement. The general classifier is obtained by calculating the weighted sum of the weak classifier of each training model (namely the additive model). The final model form is described as follows:

$$F_m(x) = \sum_{m=1}^{M} T(x;\theta_m)$$

This model is trained for M rounds, and a weak classifier $T(x;\theta_m)$ is generated for each round. The loss function of the weak classifier is described as follows:

$$\hat{\theta}_m = \underset{\theta_m}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + T(x_i;\theta_m))$$

$F_{m-1}(x_i)$ indicates the current model. The GBDT identifies the parameters of the next weak classifier by minimizing the experience risks. The loss function is selected based on L, including the square loss function, the 0–1 loss function and the logarithm loss function.

3.3.2 XGBoost

Similar to the GBDT model, the XGBoost model is an ensemble model that is generated via continuous iterations of weak classifiers. For a given credit data set $D = \{x_i, y_i\}$ with n samples and m features, x and y indicate the characteristic variant and the label variant, respectively [40]. The prediction value $\hat{y}_i$ of the ith sample can be represented as the additive model of K decision trees and is represented as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F}$$

In this equation, each iteration will generate a decision tree, f_k indicates the kth decision tree generated in the kth iteration, and $\mathcal{F}$ is the set of all decision trees.

The XGBoost model differs from the GBDT objective function: It includes the new regularization items based on the original objective function and solves the overfitting problems by penalizing the model complexity. The minimal objective function is described as follows:

$$\text{Obj} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega(f_k)$$

In this equation, the loss function l can be diversified, including the square loss and the logarithm loss. $\sum \Omega(f_k)$ is the penalty term for the complexity of the entire model, and $\Omega(f_k)$ represents the penalty term of the kth number and can be represented as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\omega^2$$

In this equation, γ is the complexity parameter, λ is the constant coefficient, and ω is the sample weight of the leaf node, namely the split of the leaf. ω^2 is the $L2$ norm square of the left weight.

The target function of the tth iteration is calculated by using the result of the $t-1$ iteration and the tree $f_t(x)$ of the tth fusion model. Each iteration will generate a tree; thus, the target function [41, 42, 43] can be changed and expanded by using the second-order Taylor series to obtain

$$L^{(t)} \simeq \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}^{(t-1)} + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right)\right] + \Omega(f_t)$$

3.3.3 LightGBM

The GBDT is an ensemble model of the decision tree and is trained by order. The GBDT learns the decision tree by fitting the negative gradient (namely the residual error) in each iteration. The main cost of the GBDT is caused by decision tree learning. The time-consuming learning process of the decision tree involves finding the best segmentation point. A popular algorithm for determining the segmentation point is the presorting algorithm, which enumerates all possible segmentation points of presorted feature values. This algorithm is simple and can be applied to determine the optimal segmentation point. However, this algorithm has low efficiency in training speed and memory consumption. Another popular algorithm is based on the histogram. The histogram-based algorithm can place a continuous feature value in a discrete box and construct the feature histogram using this box in training. This algorithm does not determine the segmentation points from the sorted feature values. Because the histogram-based algorithm is more efficient in training speed and memory consumption, we implement it in our study. The histogram-based algorithm can determine the best segmentation point according to the feature histogram diagram. The cost of the construction of the histogram diagram is $O(\#\text{data} \times \#\text{feature})$, and the cost of determining the segmentation point is $O(\#\text{bin} \times \#\text{feature})$. Generally, #bin is substantially less than #data; thus, the construction of the histogram diagram will dominate the computing complexity. If we can reduce #data or #feature, we can considerably accelerate the GBDT training [41].

Ke et al. proposed a new GBDT sampling method to balance the reduction of the data instances and the accuracy of the decision tree learning [42]. The sample weight is a suitable index for the importance of the data instances in AdaBoost. No local sample weight exists in the GBDT. The sampling method proposed in AdaBoost cannot be directly applied. The gradual change of each data instance in the GBDT provides useful data sampling information; namely, if one instance is associated with a small gradient, the training error of this instance is very small, and this instance is trained well. A simple method is to discard the data instances with small gradients. However, this action will change the data distribution and damage the accuracy of the learning model. To avoid this issue, this paper proposes a gradient-based one-side sampling method.

Gradient-based One-Side Sampling (GOSS) reserves numerous instances with higher gradients and can randomly collect samples for instances with smaller gradients. To compensate for influences on the data distribution, when the calculating the information gains, GOSS introduces a constant multiplier for data instances with small gradients. GOSS sorts data instances according to the absolute gradients of the data instances and selects the top 100% instances. However, 100% instances are randomly sampled from other data. GOSS subsequently amplifies the sampled data with small gradients using a constant. When calculating the information gains, we can focus on instances that are not sufficiently trained, which only slightly change the distribution of the original data.

In addition, Ke et al. proposed a new method to effectively reduce the features [42]. Generally, the high-dimensional data are very sparse. Feature space sparsity enables the reduction of features without loss. Many features are mutually exclusive in the spare feature space; namely, they do not simultaneously assume nonzero values. We can securely bundle exclusive features into a single feature. (We refer to it as an exclusive feature bundle.) With a precisely designed feature scanning algorithm, we can establish the same feature histogram diagram from the feature bundle. For #bundle $\ll$ #feature, the construction complexity of the histogram diagram changes from O(#data $\times$ #feature) to O(#data $\times$ #bundle), and we can rapidly accelerate the training of the GBDT without an influence on accuracy.

3.4 Ensemble learning strategy

The fused individual classifiers can introduce some strengths in the following three fields. First, statistical information shows that multiple hypotheses may attain an equal performance in the training set due to the large hypothesis space of the learning tasks. When this occurs, if a single classifier is employed, low-performance generalization may occur due to incorrect selection. This risk can be reduced by combining multiple individual classifiers. Second, calculation information shows that the learning algorithm frequently falls into a local minimum, and the corresponding generation performance of some local minima may deteriorate. After multiple runs, the algorithm may fall into the worst local minimum. Third, the representation shows that the true hypothesis of some learning tasks may not fall within the hypothesis space of the current learning algorithm. When this situation occurs, a single classifier is ineffective. Multiple classifiers can be combined to expand the corresponding hypothesis space and learn more approximate results.

Based on this analysis, we select the linear weighting method as the integration strategy of the individual classifiers because it is simple, transparent, easily executed and performs well in empirical applications. The mathematical form is described as $H(x) = \sum_{i=1}^{T} w_i h_i(x)$, where w_i is the weight of the individual classifier h_i. Generally,$w_i > 0$ and $\sum_{i=1}^{T} w_i = 1$.

4 Experiments and results analysis

4.1 Data preprocessing and feature engineering

4.1.1 Data preprocessing

In this paper, the default prediction problem is essentially a binary classification problem. The features are divided into user features, user network behavior features and user product behavior features. The missing conditions of different types are counted, and exploratory data analysis is performed. The user features include the type features and continuous features. The product features and network features only include the type features. Some features are extensively missing.

Missing values or missing features are processed differently for different problems. Sometimes, we delete them or fill them. (Filling modes are different.) The abundance of missing values in the loan business for the credit data set of this paper has a practical interpretation. Generally, either values are missing due to users' purposes or they do not exist. As a result, we cannot collect these data, which consequently enable a simple understanding of the business. For this case, an effective mode is to directly fill in a value. (It can be filled in differently according to different data types.)

Different numbers of missing cases exist between features; e.g., UserInfo_1 only includes one data sample, and UserInfo_3 includes 1515 data samples. Thus, the data have a considerable number of missing values. UserInfo_12 does not have any missing data. For different missing cases, we delete features or samples. Because extensive missing features admit a large amount of noise to the model, the model learning is severely disturbed during learning. To enhance the model robustness, we delete data with higher noise.

Based on the analysis in Figs. 3, 4 and 5, we observe that partial features are severely missing, and the maximum miss rate is 99%. In addition, one significant boundary line is located at approximately 60%; thus, we delete the samples with more than a 60% miss rate. When the missing values are processed, we start to transfer features and perform transfer learning among the data.

4.1.2 Feature engineering using transfer learning

Not all data are normalized; e.g., the minimum and maximum of UserInfo_270 are 7000 and 401,000, respectively. The minimum of ProductInfo_216 is 0. Although the ranges of different features differ, the models selected in this paper are based on the tree model; thus, normalization is not required.

The first 25 features trained by the XGBoost model are listed and sorted in Fig. 6. Feature represents the feature name. The data set includes 490-dimensional features, and the label feature is the flag. (One represents the default customer, and zero represents the nondefault customer.) Score indicates the feature score after training. (Top 25 feature attributes with the maximal feature score are displayed.)

To design the transfer learning algorithm, we select the data to transfer. Partially important features in three data sets are visually analyzed. For example, for the feature UserInfo_82, Figs. 7, 8 and 9 show that the distribution of UserInfo_82 is similar in

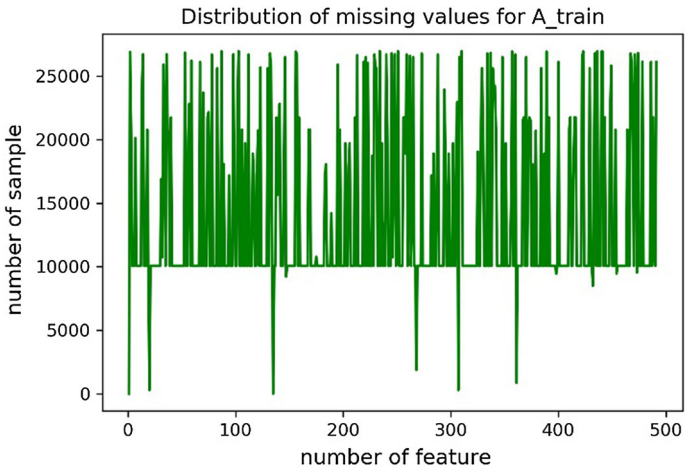

Fig. 3 Distribution of missing values for A_train

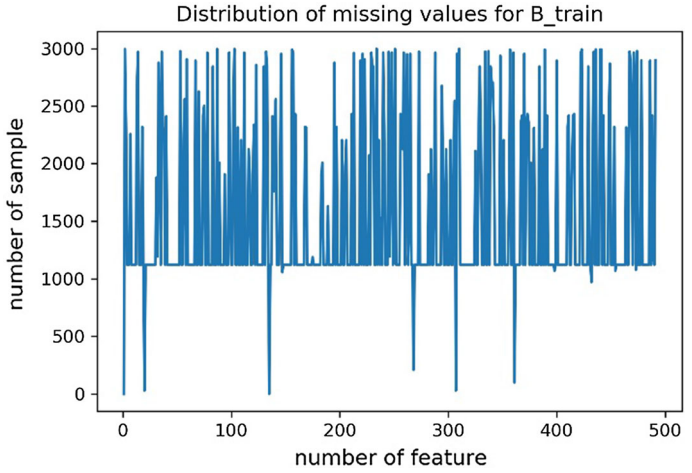

Fig. 4 Distribution of missing values for B_train

the three data sets. (The XGBoost model yields the maximal significance score for this feature.) Feature data in the auxiliary training data set A_train that are similar to those in B_train are transferred to the B_train target training data set for training to improve the performance of the prediction model. The new UserInfo_82_new obtained by the dot product of the feature UserInfo_82 in the A_train and B_train is used as the feature after the B_train transfer. The new target training data set B_train is input to the classifier for training after transfer leaning.

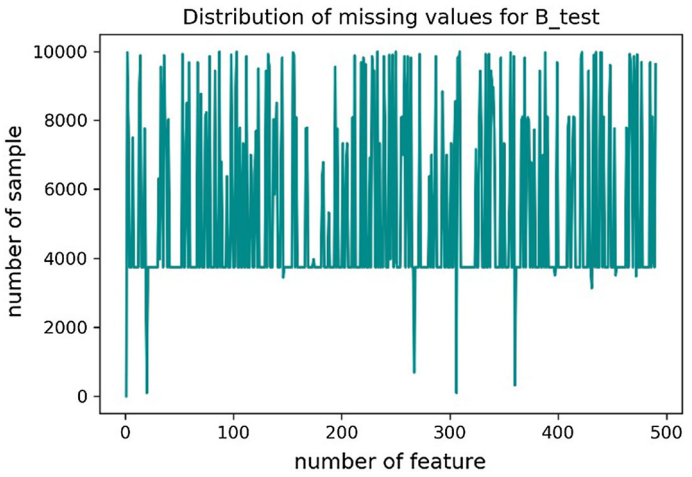

Fig. 5 Distribution of missing values for B_test

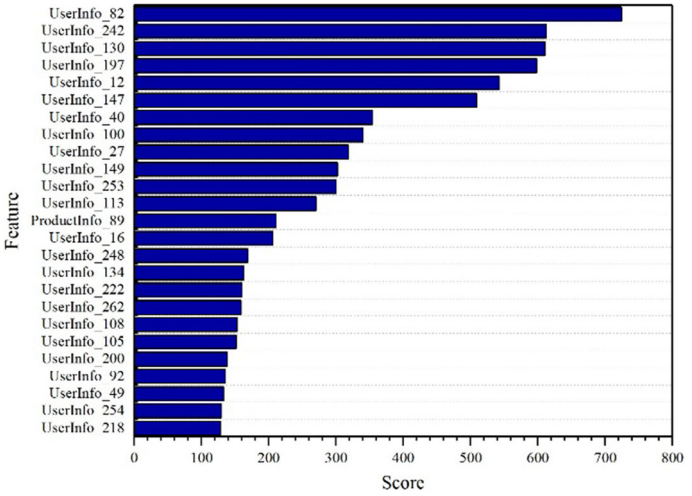

Fig. 6 Feature importance scores for XGBoost

4.2 Experimental results

We train the proposed prediction model according to the auxiliary training data set and target training data set in this section and add the transfer learning skills. The training results of the prediction model are tested and compared with different benchmark models. The benchmark models are divided into ensemble prediction models without transfer leaning skills (linear integration of GBDT, XGBoost and LightGBM) and a single benchmark model (including GBDT, XGBoost and LightGBM models with transfer skills and GBDT, XGBoost and LightGBM models without transfer skills). We compare the testing results of seven benchmark models.

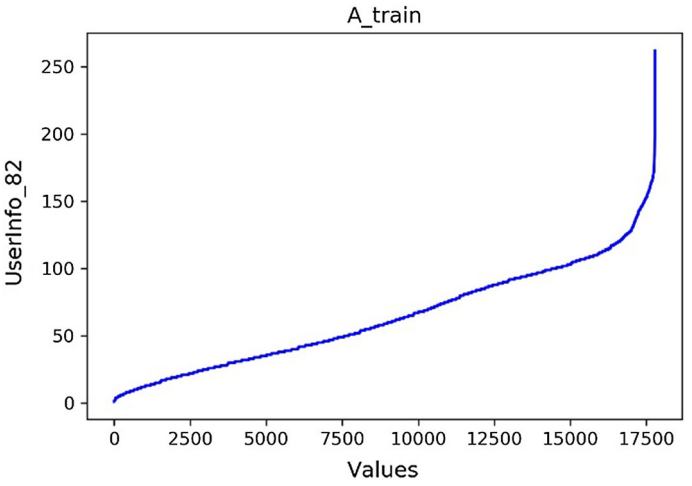

Fig. 7 Data distribution for UserInfo_82 in A_train

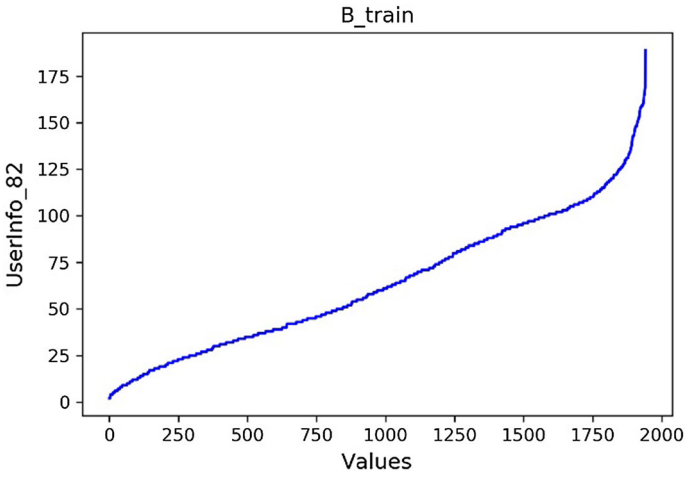

Fig. 8 Data distribution for UserInfo_82 in B_train

In addition, hyperparameters of all machine learning algorithms are optimized using 50% cross-validation in the training set. Python (version 3.6.5), which is an open-source program language, is used in the test. Python is an object-oriented explanatory computer programming language and includes rich and powerful library functions. All experiments are performed on a notebook computer with a 2.8 GHz Intel i7 central processing unit (CPU) and 16 GB RAM on a Windows 10 operating system.

4.2.1 Evaluation metrics

To accurately evaluate the default conditions of the consumption lenders, we consider some evaluation indexes in the machine learning algorithm [44]. For example, speci-

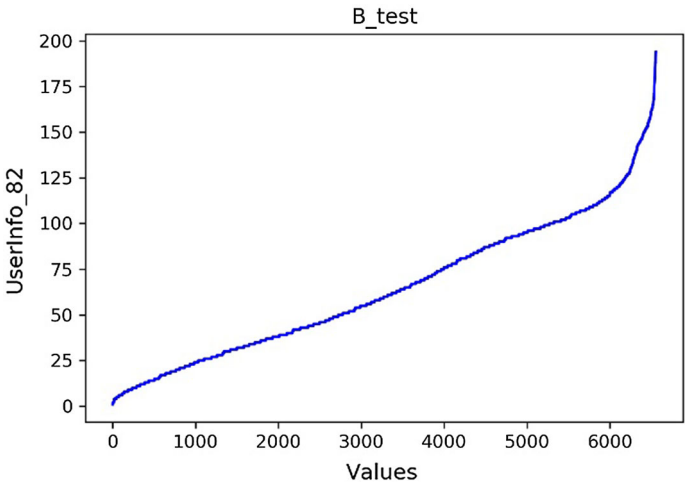

Fig. 9 Data distribution for UserInfo_82 in B_test

ficity (SPE, referred to as the true negative rate, TNR), sensitivity (SEN, referred to as the true positive rate, TPR), receiver operating characteristic (ROC) and the area under the ROC curve (AUC) are evaluated. The SPE and SEN are used to measure the correctness of the classification prediction and are defined as follows:

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

where TP indicates true positive, FP indicates false positive, TN indicates true negative, and FN indicates false negative.

We measure the discrimination capability of the model using ROC and AUC. AUC is based on the area under the ROC curve. The ROC curve is a complete sensitivity and specificity report for model evaluation. The false positive rate (FPR) is the lateral axis, the TPR is the vertical axis, $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$ and $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ in the ROC curve. For a classifier, a group (FPR, TPR) is obtained from the above coordinate axis by adjusting the threshold of the classifier. To connect these points, an ROC curve is plotted. The classifier threshold indicates the probability output of the classifier. Because a classifier cannot be quantitatively evaluated by an ROC curve, the AUC value calculated from the ROC curve is utilized as the evaluation standard. The AUC value ranges from zero to one. A higher value indicates a better classification effect.

4.2.2 Comparison of default prediction models

To validate (analyze) the prediction effect of the model proposed in this paper, we compare the performance of the prediction model with transfer skills with that of

Table 2 Performance comparison on default prediction model and benchmark models

Model	AUC	Specificity	Sensitivity
Default prediction with transfer learning	0.7170[a]	0.6142	0.7039
Ensemble-based default prediction	0.7039	0.6320	0.6669
Individual classifier with transfer learning			
GBDT	0.7001	0.7095[b]	0.5937
XGBoost	0.7158	0.6638	0.6567
LightGBM	0.6994	0.5754	0.7134[c]
Individual classifier without transfer learning			
GBDT	0.6770	0.6490	0.6157
XGBoost	0.6928	0.6640	0.6260
LightGBM	0.6662	0.6576	0.5795

[a] (0.7170) refers to the AUC value of the prediction result, which indicates the best prediction result
[b] (0.7095, specificity (SPE), also known as true negative rate (TNR)) refers to the correct prediction of non-defaulting users as a percentage of non-defaulting users, which is also the best SPE
[c] (0.7134, sensitivity (SEN), also known as true positive rate (TPR)) refers to the correct prediction of default users as a percentage of default users, which is also the best SEN

the remaining seven benchmark models. Seven benchmark models are divided into three groups. Group 1 includes three individual classifiers without transfer learning skills, namely GBDT, XGBoost and LightGBM. Group 2 includes three individual classifiers without the transfer learning skills in group 1. Group 3 is the integrated GBDT, XGBoost and LightGBM default prediction model without transfer learning skills. The performance of the prediction model and the benchmark model is shown in Table 2. Compared with the three groups of benchmark models, the default prediction model with the transfer learning skills can attain the top performance.

As shown in Fig. 10, the AUC of the default prediction model with the transfer learning skill is 0.7170. Compared with the three groups of benchmark models, its prediction accuracy is ranked in the first position. Figure 11 shows the AUC of the benchmark model in group 1. This model is the default prediction model based on the integrated fusion of GBDT, XGBoost and LightGBM without transfer learning skills, its value is 0.7039. Figure 12 shows the benchmark models in group 2, which include three individual classifiers with the transfer learning skill, namely GBDT, XGBoost and LightGBM. Their AUC values are 0.7001, 0.7158 and 0.6994, respectively. Figure 13 shows the benchmark models in group 3. Similar to the three individual classifiers in group 2, these benchmark models in group 3 do not include transfer learning skills, and their AUC values are 0.6770, 0.6928 and 0.6662.

Based on Table 2, the sensitivity of the default prediction model is 0.7039, which is higher than that (0.6669) of the benchmark model without the transfer learning skill. The prediction model based on the transfer learning skill can identify additional default customers (default targets). These features are very useful for consumer finance enterprises.

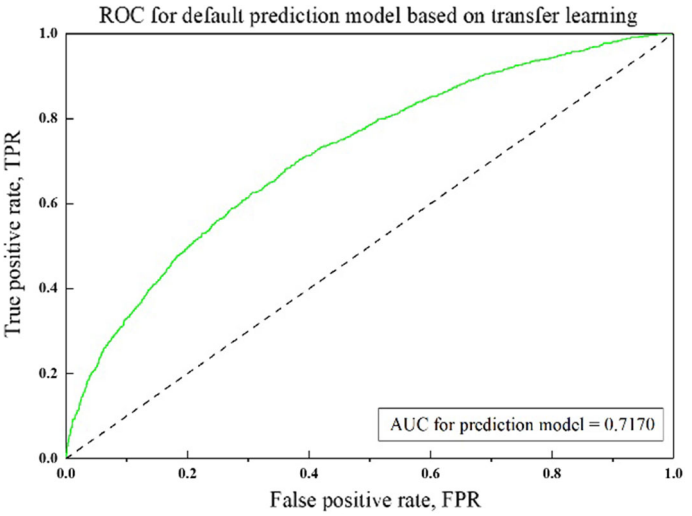

Fig. 10 ROC for default prediction model with transfer learning

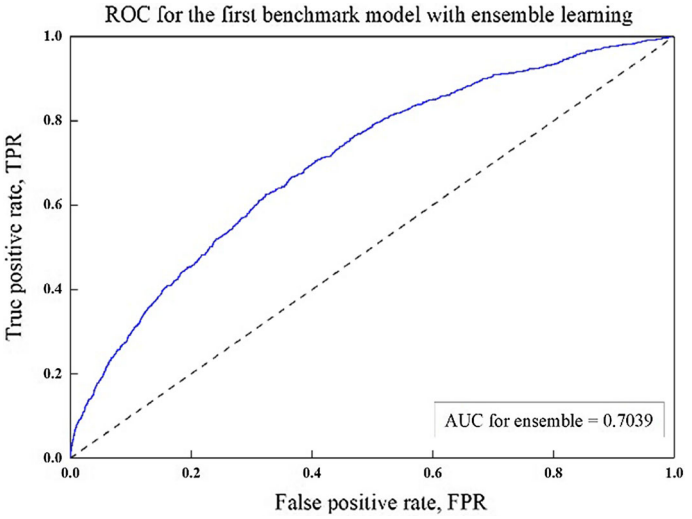

Fig. 11 ROC for the first benchmark model with the ensemble

4.3 Discussion

Loan institutions and consumer finance enterprises determine the loan issuance based on a credit score and other related information. However, determining the loan issuance criteria and evaluating the truth of the credit data sources via the Internet may be difficult for some borrowers. Some borrowers cannot provide sufficient evidence to provide their confidence level. For example, some people did not have bank accounts, and some countries have a limited credit scoring system. Therefore, the proposed

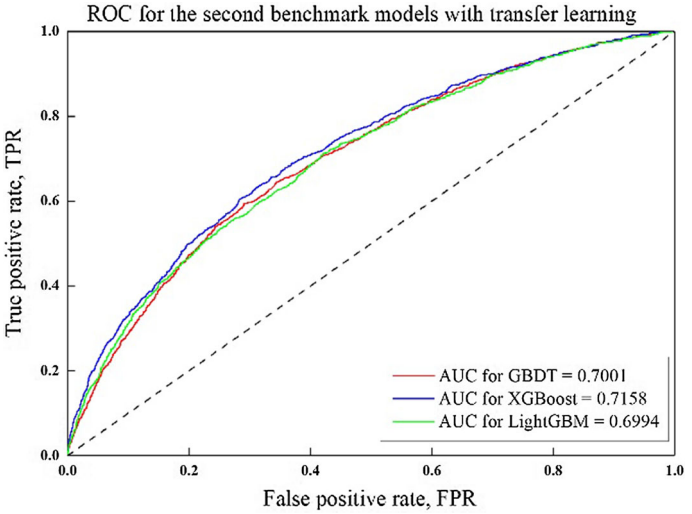

Fig. 12 ROC for the second benchmark models with transfer learning

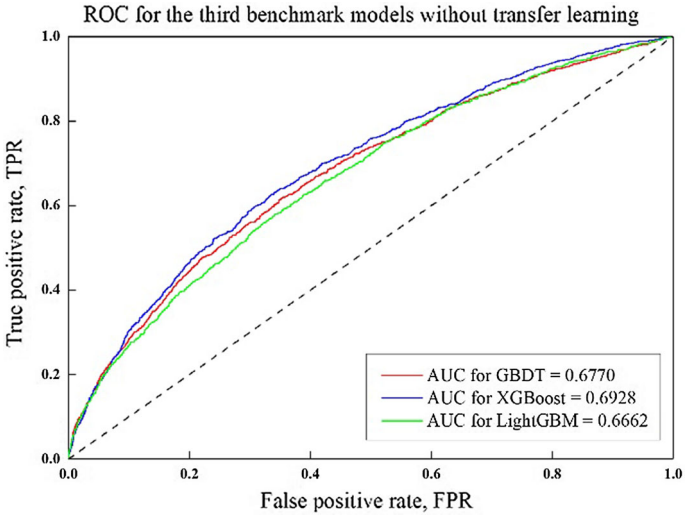

Fig. 13 ROC for the third benchmark models without transfer learning

prediction model can predict the default probability of the loan using the transfer learning skills to solve this problem. The transferred and more reliable credit data are employed in this method to make the model more universal, which can ensure the accuracy of the prediction model.

Consumer finance enterprises make decisions about numerous loans from a vast number of loan applications. However, traditional financial institutions will predict a loan decision via a manual review, which is characterized by high labor and time costs. These low-efficiency and high-cost predictions cannot satisfy the loan deci-

sion requirement of consumer finance enterprises. The default prediction method in this paper can assist consumer finance enterprises to solve this problem and can be automatically implemented based on the machine learning method. Therefore, consumer finance enterprises can instantly obtain the default prediction results and make a fast decision. Financial institutions can acquire additional data from the Internet network. The data mining methods, driven by the data, can clearly display the credit risk level of the users and improve the accuracy of the default risk prediction and the risk management capabilities of the financial institutions.

5 Conclusion

This paper proposes a novel default prediction model, namely a transfer learning model. When the training data are insufficient in credit risk evaluation and modeling, we identify feature similarities in samples using the transfer learning technique, calculating the feature similarity by the cosine angle of the dot product, and effectively transferring the auxiliary training data to the objective training data via feature similarities that form the new training data set. Therefore, the model can effectively solve the issue of insufficient training samples, effectively train the default prediction models and avoid underfitting. We combine GBDT, XGBoost and LightGBM individual classifiers based on tree structures and linear weight and fuse them to obtain the prediction results of the testing data set. The experimental results show that the default prediction method proposed in this paper can achieve a better prediction effect using the transfer learning technique. By comparing it with the benchmark models, the proposed model in this paper can achieve the optimal performance.

Our study aims to introduce the transfer learning technique to the default risk prediction domain. When the training data are insufficient, the transfer learning approach does not directly train the default risk prediction model but expands the objective training data using the auxiliary training data. This distinction enables the provision of sufficient training data to better resolve small sample issues that occur in machine learning methods. In the future, we will avoid the negative transfers in the data transfer, study the transferability between the source domain and the target domain and demonstrate these considerations to ensure that negative learning does not occur. Although data drive the technology development in the age of big data, the data remain insufficient in many aspects of the financial field. A better application of the transfer learning technique to undertake these issues will be discussed.

Acknowledgements This work was funded by the National Natural Science Foundation of China under Grant Nos. 71571058 and 71690235, and Anhui Provincial Science and Technology Major Project under Grant Nos. 16030801121 and 17030801001.

References

1. Ding S, Wang Z, Wu D, Olson DL (2017) Utilizing customer satisfaction in ranking prediction for personalized cloud service selection. Decision Support Syst 93:1–10. https://doi.org/10.1016/j.dss.2016.09.001

2. Miao H, Ramchander S, Ryan P, Wang T (2018) Default prediction models: the role of forward-looking measures of returns and volatility. J Empir Finance 46:146–162. https://doi.org/10.1016/j.jempfin.2018.01.001
3. Barboza F, Kimura H, Altman E (2017) Machine learning models and bankruptcy prediction. Expert Syst Appl 83:405–417. https://doi.org/10.1016/j.eswa.2017.04.006
4. Maldonado S, Peters G, Weber R (2018) Credit scoring using three-way decisions with probabilistic rough sets. Inf Sci (Ny) 0:1–15. https://doi.org/10.1016/j.ins.2018.08.001
5. Zheng C, Xia C, Guo Q, Dehmer M (2018) Interplay between SIR-based disease spreading and awareness diffusion on multiplex networks. J Parallel Distrib Comput 115:20–28. https://doi.org/10.1016/j.jpdc.2018.01.001
6. Jeon J, Yoon JH, Park CR (2018) The pricing of dynamic fund protection with default risk. J Comput Appl Math 333:116–130. https://doi.org/10.1016/j.cam.2017.10.031
7. Wei X, Luo X, Li Q et al (2015) Online comment-based hotel quality automatic assessment using improved fuzzy comprehensive. IEEE Trans Fuzzy Syst 23:72–84
8. Jiang H, Ching WK, Yiu KFC, Qiu Y (2018) Stationary Mahalanobis kernel SVM for credit risk evaluation. Appl Soft Comput J 71:407–417. https://doi.org/10.1016/j.asoc.2018.07.005
9. Maldonado S, Bravo C, López J, Pérez J (2017) Integrated framework for profit-based feature selection and SVM classification in credit scoring. Decision Support Syst 104:113–121. https://doi.org/10.1016/j.dss.2017.10.007
10. Pang X, Zhou Y, Wang P et al (2018) An innovative neural network approach for stock market prediction. J Supercomput. https://doi.org/10.1007/s11227-017-2228-y
11. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359. https://doi.org/10.1109/TKDE.2009.191
12. Tian S, Yu Y (2017) Financial ratios and bankruptcy predictions: an international evidence. Int Rev Econ Finance 51:510–526. https://doi.org/10.1016/j.iref.2017.07.025
13. Ciampi F (2015) Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. J Bus Res 68:1012–1025. https://doi.org/10.1016/j.jbusres.2014.10.003
14. 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 Syst. https://doi.org/10.1016/j.dss.2018.05.001
15. Tkáč M, Verner R (2015) Artificial neural networks in business: two decades of research. Appl Soft Comput 38:788–804. https://doi.org/10.1016/j.asoc.2015.09.040
16. Krauss C, Do XA, Huck N (2017) Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. Eur J Oper Res 259:689–702. https://doi.org/10.1016/j.ejor.2016.10.031
17. Zhou L, Si YW, Fujita H (2017) Predicting the listing statuses of Chinese-listed companies using decision trees combined with an improved filter feature selection method. Knowl Based Syst 128:93–101. https://doi.org/10.1016/j.knosys.2017.05.003
18. Yao X, Crook J, Andreeva G (2015) Support vector regression for loss given default modelling. Eur J Oper Res 240:528–538. https://doi.org/10.1016/j.ejor.2014.06.043
19. Gordini N (2014) A genetic algorithm approach for SMEs bankruptcy prediction: empirical evidence from Italy. Expert Syst Appl 41:6433–6445. https://doi.org/10.1016/j.eswa.2014.04.026
20. Arar ÖF, Ayan K (2017) A feature dependent Naive Bayes approach and its application to the software defect prediction problem. Appl Soft Comput J 59:197–209. https://doi.org/10.1016/j.asoc.2017.05.043
21. Sun J, Lang J, Fujita H, Li H (2017) Imbalanced enterprise credit evaluation with DTE-SBD: decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Inf Sci (Ny) 425:76–91. https://doi.org/10.1016/j.ins.2017.10.017
22. Yu L, Zhou R, Tang L, Chen R (2018) A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data. Appl Soft Comput J 69:192–202. https://doi.org/10.1016/j.asoc.2018.04.049
23. Wang D, Zhang Z, Bai R, Mao Y (2018) A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring. J Comput Appl Math 329:307–321. https://doi.org/10.1016/j.cam.2017.04.036
24. Sohn SY, Kim DH, Yoon JH (2016) Technology credit scoring model with fuzzy logistic regression. Appl Soft Comput J 43:150–158. https://doi.org/10.1016/j.asoc.2016.02.025

25. Guo Y, Zhou W, Luo C et al (2016) Instance-based credit risk assessment for investment decisions in P2P lending. Eur J Oper Res 249:417–426. https://doi.org/10.1016/j.ejor.2015.05.050
26. Hurlin C, Leymarie J, Patin A (2018) Loss functions for loss given default model comparison. Eur J Oper Res 268:348–360. https://doi.org/10.1016/j.ejor.2018.01.020
27. Pan W-T, Huang C-E, Chiu C-L (2016) Study on the performance evaluation of online teaching using the quantile regression analysis and artificial neural network. J Supercomput 72:789–803. https://doi.org/10.1007/s11227-015-1599-1
28. Feng X, Xiao Z, Zhong B et al (2018) Dynamic ensemble classification for credit scoring using soft probability. Appl Soft Comput J 65:139–151. https://doi.org/10.1016/j.asoc.2018.01.021
29. Ye R, Dai Q (2018) A novel transfer learning framework for time series forecasting. Knowl Based Syst 156:74–99. https://doi.org/10.1016/j.knosys.2018.05.021
30. Nasiri M, Minaei B (2016) Increasing prediction accuracy in collaborative filtering with initialized factor matrices. J Supercomput 72:2157–2169. https://doi.org/10.1007/s11227-016-1717-8
31. Lu J, Behbood V, Hao P et al (2015) Transfer learning using computational intelligence: a survey. Knowl Based Syst 80:14–23. https://doi.org/10.1016/j.knosys.2015.01.010
32. Zhu Y, Hu X, Zhang Y, Li P (2018) Transfer learning with stacked reconstruction independent component analysis. Knowl Based Syst 152:100–106. https://doi.org/10.1016/j.knosys.2018.04.010
33. Ding S, Li Y, Wu D et al (2018) Time-aware cloud service recommendation using similarity-enhanced collaborative filtering and ARIMA model. Decision Support Syst 107:103–115. https://doi.org/10.1016/j.dss.2017.12.012
34. Wang J, Ding S, Song M et al (2018) Smart community evaluation for sustainable development using a combined analytical framework. J Clean Prod 193:158–168. https://doi.org/10.1016/j.jclepro.2018.05.023
35. Wang Y, Zhai J, Li Y et al (2018) Transfer learning with partial related "instance-feature" knowledge. Neurocomputing 310:115–124. https://doi.org/10.1016/j.neucom.2018.05.029
36. Xia C, Ding S, Wang C et al (2017) Risk analysis and enhancement of cooperation yielded by the individual reputation in the spatial public goods game. IEEE Syst J 11:1516–1525. https://doi.org/10.1109/JSYST.2016.2539364
37. 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
38. 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
39. Haixiang G, Yijing L, Shang J et al (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239. https://doi.org/10.1016/j.eswa.2016.12.035
40. 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
41. Ma X, Sha J, Wang D et al (2018) Study on a prediction of P2P network loan default based on the machine learning lightGBM and XGboost algorithms according to different high dimensional data cleaning. Electron Commer Res Appl. https://doi.org/10.1016/j.elerap.2018.08.002
42. Ke G, Meng Q, Wang T et al (2017) LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3148–3156
43. Fujita K, Takewaki I (2011) An efficient methodology for robustness evaluation by advanced interval analysis using updated second-order Taylor series expansion. Eng Struct 33:3299–3310. https://doi.org/10.1016/j.engstruct.2011.08.029
44. Diwakaran S, Perumal B, Vimala Devi K (2018) A cluster prediction model-based data collection for energy efficient wireless sensor network. J Supercomput. https://doi.org/10.1007/s11227-018-2437-z