Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.037287
Article

MSEs Credit Risk Assessment Model Based on Federated Learning and Feature Selection

Zhanyang Xu[1], Jianchun Cheng[1,*], Luofei Cheng[1], Xiaolong Xu[1,2] and Muhammad Bilal[3]

[1]School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[2]State Key Laboratory Novel Software Technology, Nanjing University, Nanjing, 210023, China
[3]Department of Computer and Electronics Systems Engineering, Hankuk University of Foreign Studies, Yongin-si, Gyeonggi-do, 17035, Korea
*Corresponding Author: Jianchun Cheng. Email: 20201221008@nuist.edu.cn
Received: 29 October 2022; Accepted: 03 March 2023

Abstract: Federated learning has been used extensively in business innovation scenarios in various industries. This research adopts the federated learning approach for the first time to address the issue of bank-enterprise information asymmetry in the credit assessment scenario. First, this research designs a credit risk assessment model based on federated learning and feature selection for micro and small enterprises (MSEs) using multi-dimensional enterprise data and multi-perspective enterprise information. The proposed model includes four main processes: namely encrypted entity alignment, hybrid feature selection, secure multi-party computation, and global model updating. Secondly, a two-step feature selection algorithm based on wrapper and filter is designed to construct the optimal feature set in multi-source heterogeneous data, which can provide excellent accuracy and interpretability. In addition, a local update screening strategy is proposed to select trustworthy model parameters for aggregation each time to ensure the quality of the global model. The results of the study show that the model error rate is reduced by 6.22% and the recall rate is improved by 11.03% compared to the algorithms commonly used in credit risk research, significantly improving the ability to identify defaulters. Finally, the business operations of commercial banks are used to confirm the potential of the proposed model for real-world implementation.

Keywords: Federated learning; feature selection; credit risk assessment; MSEs

1 Introduction

The expansion of credit to MSEs is not only a regulatory policy objective but also a requirement imposed on commercial banks for sustainable business development to promote supply-side reform. However, given the large number of MSEs and their characteristics, they usually face problems such as a smaller amount of operating capital, limited access to information, difficulty in obtaining guarantees, and high financial risk. The amount and ratio of non-performing loans in China's commercial banks

show an increasing trend annually, resulting in higher risk pressure and a high loss rate for banks [1]. Information asymmetry is the main cause of the risk associated with lending to MSEs [2]. Commercial banks have difficulty collecting sufficient data on MSEs and their owners, which makes it difficult for banks to make reliable assessments of the financial situation and growth trajectory of MSEs. As a result, one of the most serious challenges that banks need to study and address in order to expand their lending business is the establishment of a scientific and efficient system for assessing the credit risks associated with MSEs.

According to the financing practices of MSEs in many countries, bank loans are the most prominent source of external financing for these businesses [3]. As the economy becomes increasingly digitalized, the traditional way of making credit decisions for MSEs is primarily based on information such as audit opinions, which are collected manually, and the process is lengthy, repetitive, and requires complex data verification work. This makes it difficult to extract information with real analytical value to meet actual needs and even more challenging to support the rapid growth of the credit business.

Financial inefficiencies resulting from the increasing inability of the traditional processing model to adapt to the evolving environment are becoming more apparent. Many scholars have successfully applied machine learning techniques to credit risk assessment, including neural networks (NNs) [4], genetic algorithms (GAs) [5], and decision trees (DTs) [6–8]. However, there are still some issues that require further discussion.

On the one hand, there is still a problem of limited data or poor data quality in the field of financial credit, which is not enough to support the realization of artificial intelligence technology. As a reference data reflecting social development and economic operation, electricity data has the characteristics of "fine accuracy" and "wide coverage", which can effectively reflect the business situation and development trend of various industries. Through collation and modeling based on power file data and power consumption data, from the dimensions of power consumption scale, power consumption stability, power consumption characteristics and power consumption reputation, we can effectively identify stagnant enterprises, empty shell enterprises and enterprises with poor development in various industries, and draw accurate portraits for small and micro enterprises, so as to timely understand the operation status of enterprises and position them for support. Given the lack of symmetry in financial data and the difficulty in obtaining high-quality financial statements for micro and small enterprises, power and credit data should be used as reference dimensions for assessing the credit risk of enterprises. At present, applied research to estimate credit risk based on an enterprise's electricity consumption and credit data is still insufficient.

On the other hand, features need to be filtered, because when machine learning models are trained using existing credit rating indicators, there are redundant and irrelevant features that could cause the "curse of dimensionality" and hinder model performance. Despite increasing prediction accuracy, the usual feature selection techniques of filters and wrappers do not offer a quantitative opinion on how important a feature is, and are less interpretable. Additionally, current research rarely focuses on both feature selection and model construction.

On the other hand, features need to be filtered because when machine learning models are trained using existing credit indicators, there are redundant and irrelevant features that can cause the "curse of dimensionality" and hinder model performance. Despite increasing predictive accuracy, the common feature selection techniques of filters and wrappers do not provide a quantitative opinion on the importance of a feature and are less interpretable. In addition, current research rarely focuses on both feature selection and model construction.

To address the above issues, one option is to fully utilize external data from the People's Bank of China (PBC) credit reference and electricity systems to compensate for the lack of subjective empowerment and improve the information asymmetry situation between banks and enterprises. However, it also introduces unprecedented problems such as "data silos" [9], which are impenetrable barriers between data sources that provide data support. Data confidentiality measures are necessary because of market competition, privacy and security concerns, and regulatory requirements [10,11]. Incalculable damage can be caused if sensitive data is compromised. Global attention is focused on data security [12], and the emergence of numerous privacy laws has made "data silos" more commonplace.

How to design a machine learning framework that allows AI systems to work together more efficiently and accurately with their respective data while meeting the privacy, security, and regulatory requirements for corporate financial, credit, and power data is an important topic at hand. We are shifting the focus of our research to how to solve the "data silos" problem. A feasible solution that satisfies privacy and data security is federated learning (FL) [13,14]. As a growing artificial intelligence technique, federated learning has been actively researched and applied in various fields and also provides a new way to build credit models for MSEs.

In this research, we suggest a credit risk assessment model for MSEs based on federated learning and feature selection. This paper's main points are:

- Based on the bank's internal loan data and the addition of external data, such as the electricity system and the credit system of the PBC, a model architecture based on federated learning technology was creatively applied to MSEs' credit scenarios.
- We present a two-stage feature selection algorithm: XGBoost-based mRMR-PCA (XMP) for constructing the optimal feature and a hybrid filter-wrapper feature selection algorithm that combines the benefits of high accuracy and efficiency.
- We propose a local update screening strategy based on a dual subjective logic model to filter trustworthy model parameters for aggregation at each local model update to ensure the quality of the overall model.

The following sections of the paper are structured as follows: In Section 2, we outline earlier initiatives to improve MSEs' credit model performance. In Section 3, we provide a comprehensive description of the specifics of the proposed framework. As a result, Section 4 provides details on the experimental setups used to produce the findings of this study. Finally, Section 5 concludes our analysis by summarizing its findings.

2 Related Work

2.1 Enterprise Credit Risk Assessment Model

From early expert systems and multivariate statistical analysis models to today's use of machine learning and artificial intelligence in the credit assessment indicator system, there has been extensive research into credit risk assessment methodologies. The ability to effectively use existing technologies to mitigate the impact of bank-enterprise information asymmetry is the key to tackling the problem of financing for MSEs.

The most commonly used algorithm in credit risk research is Logistic Regression (LR) because of its simple structure, high interpretability, and excellent accuracy [15]. Jones et al. [16,17] conducted a thorough investigation of the predictive accuracy of several classifiers using a large number of samples, ranging from the most sophisticated techniques such as Probit, LR, and Linear Discriminant

Analysis (LDA) to more sophisticated methods such as Support Vector Machines (SVM), NNs, and statistical learning models such as Random Forests (RF) and Generalized Boosting. They propose that Generalized Boosting and RF outperform conventional LR, Probit, and LDA models and even the popular AI methods NNs and SVM.

Although many cutting-edge AI models have demonstrated exceptional accuracy, their limited interpretability and the data inadequacy of real-world lending scenarios prevent their widespread use in credit assessment. However, machine learning algorithms represented by DT, RF, Gradient Boosting Decision Tree (GBDT), and Extreme Gradient Boosting (XGB) perform better on smaller data sets. They can produce better predictions in a relatively short training period. Nguyen [18] compared XGB with LR, DT, and NNs in recent years to demonstrate its more remarkable results in credit risk analysis. Li et al. [19] explored the theoretical modeling of the XGB algorithm for the big data-based credit assessment classification problem. By comparing the XGB model with the LR, DT, RF, and GBDT models, they found that it performed significantly better in feature selection and data classification. Recent studies have shown that ensemble learning algorithms can effectively help banks reduce credit risk.

The use of financial data alone is far from sufficient to effectively predict credit risk for MSEs. Non-financial considerations are equally important. According to Yang et al. [20], big data credit reference is a powerful complement to traditional credit business and can be used in a wider range of business scenarios. In addition, Ala'raj et al. [21] demonstrated that selecting the optimal subset of features can significantly improve prediction accuracy by relying on the results of experiments. To speed up computation and improve prediction accuracy, Zhang et al. [22] found that the feature selection procedure is crucial. Cui et al. [23] introduced a new multiple structural interaction elastic net model for feature selection that embodies the structural connections between pairs of samples by transforming the initial vector features into a structure-based feature map representation and sets information-theoretic criteria to maximize relevance and minimize redundancy. The approach effectively detects critical elements for credit assessment in Internet finance. However, it does not deal well with noisy data in the credit risk assessment.

2.2 Federated Learning

Although there have been significant developments in the study of AI-based credit risk assessment models, they have never solved the problem faced by the MSEs credit scenarios, which is the asymmetry of bank-enterprise information. Federated learning technology provides a viable solution to this problem.

Federated learning is defined as a machine learning process in which each participant can build a shared machine learning model using data from other participants. The data within each participant is not local, and no data resources need to be shared. The constraints of the federated learning system are:

$$\left|E_{fed} - E_{tra}\right| < \delta \tag{1}$$

where E_{fed} denotes the effect of the federated learning model, E_{tra} denotes the effect of the traditional method of modeling, and δ denotes a bounded integer.

Federated learning is a machine learning model in which multiple clients collaborate to solve machine learning problems under the coordination of a central server or service provider. Each client's raw data is stored locally and is not exchanged or transmitted. Under such a federal mechanism, the joint enterprise financial, credit and power data modeling can effectively solve the asymmetry of bank-enterprise information. In turn, the exploration of multi-party data calculation and credit risk

prediction can be carried out to deeply explore the value of data. Thus, federated learning technology is a "win-win" model that is extremely valuable for business interests.

With the development of research on federated learning, the projects of federated learning in various application scenarios are coming to the ground. Chen et al. [24] proposed a communication fraud detection model based on federated learning, which allows the federation to jointly model the data sets of telecom operators and public security bureau. In view of the new coronavirus pneumonia, Xu et al. [25] used edge learning and federated learning techniques to design a management model for the prevention and control of the new coronavirus pneumonia epidemic in colleges and universities, and rapidly analyzed the data of teachers and students collected by colleges and universities, in order to arrange the corresponding preventive measures in time to prevent the spread of new pneumonia. According to the current data operation situation of banks and other financial institutions, Zheng [26] actively explores the application of "federated learning + financial recommendation". Wang et al. [27] also take the insurance industry as the background, under the premise of legal compliance, to build a data fusion architecture based on federated learning and applicable privacy protection tools. Federated learning has been applied to financial, insurance, medical, and other fields.

Based on previous research, this paper analyzes and expands the indicators that affect enterprise credit, adds electricity consumption data and credit data through federated learning, and investigates the SecureBoost algorithm [28] to design a credit evaluation model for MSEs. The SecureBoost algorithm is introduced by WeBank, which performs better in terms of accuracy, differentiation, and stability. In essence, SecureBoost is just a model of the XGBoost algorithm using a federated learning technique. To increase the accuracy of the proposed model, a hybrid mRMR-PCA feature selection strategy based on XGBoost feature importance is also explored.

3 Framework of the Credit Risk Assessment Model Based on Federated Learning and Feature Selection for MSEs

Since the credit evaluation scenario involves data interactions among multiple banks and grids, there are discrepancies in the accuracy and authenticity of the data held by each bank. We hope that each participant can achieve data value fusion through federated learning technology while protecting the privacy of the data. Fig. 1 illustrates how this model uses the electric power system as the federated learning participant and trains the enterprise credit risk assessment model in cooperation with each bank participant.

The data provider, the model user, and the central server are the three primary participants in the proposed credit model for MSEs. In addition to providing the data matrix and class labels needed for training as a data provider, each bank participant also serves as a model user. The grid side only serves as a data provider, providing the matrix of power-related data needed for training. Model aggregation, dissemination, validation, and other services are handled by the central server. The proposed model includes four key processes: encrypted entity alignment, hybrid feature selection, secure multi-party computation, and global model updating.

3.1 Encrypted Entity Alignment

Since the data samples between the power company and the bank do not fully overlap, the shared users' data is extracted using encryption-based sample alignment technology, and the corresponding keys are obtained from the unified key management platform to complete the alignment with the power company's encrypted sample ID. This ensures that the data of the same person in different dimensions on each node can be correctly matched during training.

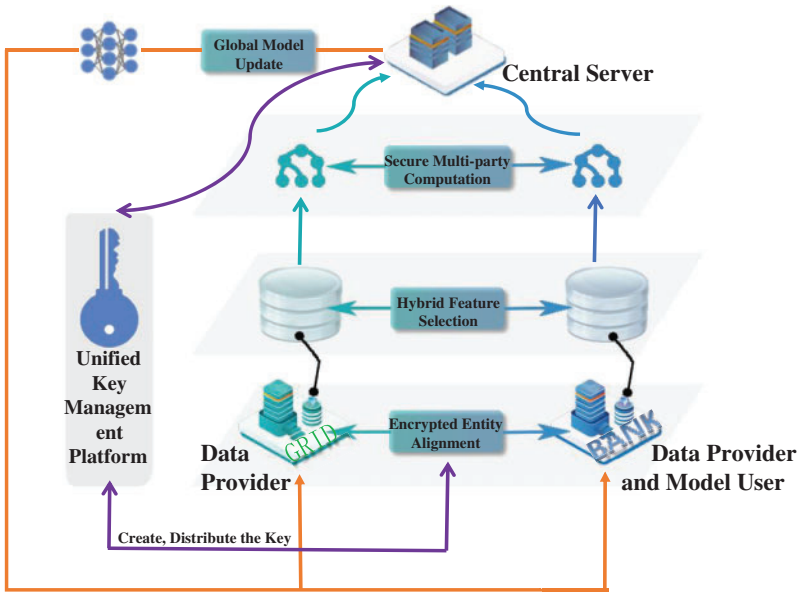

Figure 1: Architecture of the credit assessment model based on federated learning and feature selection

The model uses a secure intersection scheme based on RSA and a hashing mechanism. The grid side hashes the ID and multiplies it by a random number mask before sending the encrypted data back to the bank side. The key is first distributed by the unified key management platform, and then the bank side delivers the public key to the grid side. The data is then sent to the grid side, which then performs the power of d operation on the accepted result according to Euler's or Fermat's law. After removing the random number mask, the grid side hashes its own data to the power of d, performs another hash, intersects the results with the data that was previously sent, and sends the output to the bank side.

3.2 Hybrid Feature Selection

After determining the shared users on both sides, the raw data must be processed to remove redundant data features. Then, using two different feature selection techniques, i.e., integrated filters and wrappers, the best feature set for corporate credit evaluation is created from the raw data. We propose a two-stage feature selection method called XMP, in which the first stage on the bank side generates a candidate feature set using the max-relevance and min-redundancy (mRMR) algorithm [29], and the principal component analysis (PCA) method [30] is used on the grid side to reduce the dimension of power data. In the second stage, the feature importance is determined using the XGBoost average gain. For each training cycle with the XGBoost model, the average gain of the features is calculated, sorted, and the current loss value is recorded. The features corresponding to the most recent minimum average gain are dropped before the next training, and the next round of training is performed with the new feature set until it is completely dropped.

For the data on the grid-side, let there be n samples, and each sample corresponds to the data matrix X with m variables for normalization. $\overline{x}_j$ is the average sample value of the j-th indicator:

$$X_{ij} = \frac{x_{ij} - \frac{1}{m}\sum_{i=1}^{m} x_{ij}}{\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(x_{ij} - \overline{x}_j\right)^2}} i = 1, 2, \ldots, m \;\; j = 1, 2, \ldots, n \tag{2}$$

Then establish the covariance matrix R. The original variables x_i and x_j's correlation coefficients are determined as follows:

$$r_{ij} = \frac{\sum_{k=1}^{m}\left(x_{ki} - x_i\right) - \left(x_{kj} - x_j\right)}{\sqrt{\sum_{k=1}^{m}\left(x_{ki} - \overline{x}_i\right)^2 \sum_{k=1}^{m}\left(x_{kj} - \frac{1}{m}\sum_{i=1}^{m} x_{ij}\right)}} \tag{3}$$

Then calculate the eigenvalue $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ of R, and the corresponding eigenvector. The feature vectors are arranged in the same order to form a feature matrix, so the cumulative contribution determines how many principal components are preserved. The Contribution rate C is the ratio of a particular eigenvalue to all other eigenvalues, and its calculation formula is:

$$C = \frac{\lambda_j}{\sum_{i=1}^{n} \lambda_i} \tag{4}$$

We need to establish a set of feature subsets from the original feature set of the bank side, so that the target output features have a maximum correlation with the class variables and the features have a minimum redundancy with each other. The correlation of the category variables with the feature subsets, and the redundancy of all features in the set are defined as:

$$\max D\left(S, c\right), D = \frac{\sum_{x_i \in S} I\left(x_i; c\right)}{|S|} \tag{5}$$

$$\max R\left(S\right), R = \frac{\sum_{x_j \in S} I\left(x_i; x_j\right)}{|S|^2} \tag{6}$$

The ultimate objective of mRMR is to compute the set that has the maximum correlation minus the minimum redundancy, which can directly optimize the following equation directly:

$$mRMR = \max\ \Phi\left(D, R\right), \Phi = D - R \tag{7}$$

In these equations, x_i is the i-th feature, The feature subset is expressed by S, $c = \{c_1, c_2, \ldots, c_A\}$ is the categorical variable, A is the total number of categories, target category c and feature i's mutual information is represented as $I\left(x_i; c\right)$. This pattern goes on: for example, $I\left(x_i; x_j\right)$represents the mutual information between feature i and feature j. Algorithm 1 describes the first stage feature selection process.

Algorithm 1: mRMR-PCA

Input: $D_{n\times d}$, $X_{n\times p}$, Preselected feature set F, Class label set C, Percentage of feature selection k, PCA threshold value t

Output: Grid-side feature subset S_{grid}, Bank-side feature subset S_{bank}

(Continued)

Algorithm 1: Continued

$x_i \leftarrow x_i - \frac{1}{p}\sum_{j=1}^{p} x_j$
Compute covariance matrix $R \leftarrow \frac{1}{p}\sum_{j=1}^{p} X_i X_i^T$
Compute eigen vectors, eigen values of R: $[E_{val}, E_{vec}] \leftarrow eig\,(R)$
Extracting eigen values: $value_{set} \leftarrow dig\,(E_{val})$
Getting $sort\,(value_{set}, descend)$
Getting n by $\frac{\sum_{i=1}^{n}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \geq t$
for $i = 1 \rightarrow n$ **do**
    $Z_i \leftarrow W^T x_i$
end for
$S, S_{bank} \leftarrow \{\}$
for $each\,\forall x_i \in F$ **do**
    $MI\_set \leftarrow I\,(x_i; c)$
end for
$f_{\max} \leftarrow \max\_sort\,(MI\_set)\,, F \leftarrow F - f_{\max}, S \leftarrow f_{\max}$
While F**do**
    select f
    for $m = 2 \rightarrow Size\,(F)$**do**
        select f by $\max_{x_i \in F_{m-1}} \left\{ I\,(x_i; c) - \frac{1}{m-1}\sum_{j=1}^{m-1} I\left(x_i; x_j\right) \right\}$
    end for
    $S \leftarrow S \cup \{f\}\,, F \leftarrow F - \{f\}$
end while
$S_{grid} \leftarrow (Z_1, Z_2, \ldots, Z_n)$, $S_{bank} \leftarrow S\,(1; k \times d)$

Since mRMR only considers the local optimum, the XGBoost feature importance evaluation method is used to select the optimal set of unique features for corporate credit assessment by multiple rounds of training based on mRMR-PCA. When the model finally has the K decision trees, it needs to add up the $Gain_i$ from each tree and take the average to determine the important metrics for each feature:

$$Ave\_Gain = \frac{\sum_{i=1}^{K} Gain_i}{K} \tag{8}$$

Assuming that the feature dimension of the candidate feature set on the bank side is m, and similarly, on the power grid side is N, the original training sample set is constructed with the feature dimension $j = m+n$, and Ave_Gain is calculated by training the XGBoost model, sorting in descending order and recording the loss value. When the feature elimination process is complete, the dimension of the feature which corresponds to the minimum loss value is given, and then the optimal set of features is obtained. Algorithm 2 describes the second stage feature selection process.

Algorithm 2: XGBoost based mRMR-PCA (XMP)

Input: Grid-side feature subset S_{grid}, Bank-side feature subset S_{bank}
Output: Optimal feature subset F

(Continued)

Algorithm 2: Continued

1. *feature_list* $\leftarrow S_{grid} \cup S_{bank}$
2. **While** *feature_list*) **do**
3. Train $XGBoost\,(X, y)$
4. Record *Loss*
5. Calculate $Ave\_Gain \leftarrow \dfrac{\sum_{i=1}^{K} Gain_i}{K}$
6. Get *ranked_featureImportance*
7. Del *ranked_featureImportance*$[-1]$
8. *feature_list* $\leftarrow$ *ranked_featureImportance*
9. **end while**
10. $F \leftarrow \arg\min_{F} Loss$

3.3 Secure Multi-Party Computation

The model can be trained with the optimal feature set after the hybrid feature selection. Encrypted training with a central server's assistance is necessary to guarantee the data's secrecy throughout the training. The SecureBoost algorithm is used for training in this paper, which includes a regularization component in the loss function to reduce the complexity and to prevent overfitting. The loss function is expressed as:

$$L = \sum_{i=1}^{m} l\left(y_i + \hat{y}_i\right) + \sum_{K=1}^{K} \left(\gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2\right) \tag{9}$$

Among them, i reflects the dataset's i-th sample, the predicted and actual values are represented by $\hat{y}_i$ and y_i, respectively, L is the loss function, K is the number of all trees established, m denotes the total data volume at the moment the k-th tree was imported. The first term in the equation calculates how much the actual value deviates from the predicted value. The second term is defined as the complexity, where γ and λ are manually set parameters, ω represents the weight of each leaf node, and T represents the number of leaf nodes.

When the model is updated for each bank participant node, it is assumed that the t-th loss function is:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(x_i\right)\right) + \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} \omega_j^2 \tag{10}$$

Take the second-order Taylor expansion for $L^{(t)}$:

$$L^{(t)} \cong \sum_{i=1}^{n} \left[g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right] + \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} \omega_j^2 \tag{11}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ is the first derivative and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ is the second derivative. The convergence of the model can be accelerated and the optimal solution can be found by using the second order Taylor expansion. To obtain the optimal loss function, it is necessary to introduce the tree structure.

$$L^{(t)} = \sum_{i=1}^{n} \left[G_j \omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2\right] + \gamma T_t \tag{12}$$

Among them, $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$ represents the first derivative and the second derivative of the leaf node. The final loss function is a quadratic function about ω_j, therefore:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{13}$$

$$L^* = -\frac{1}{2} \sum\nolimits_{j=1}^{T_t} \frac{G_j^2}{H_j + \lambda} + \gamma T_t \tag{14}$$

The tree's structure is better when L^* has a lower value, and the minimal value of the loss function can be obtained by G and H. The following formula determines the optimal split for the leaf nodes:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \tag{15}$$

When the nodes are split, SecureBoost determines the objective function and information gain depending on G_j and H_j. The Paillier homomorphic encryption [31,32] algorithm with asymmetric keys is utilized to ensure privacy protection in gradient operations, considering the peculiarities of data security and application scenarios.

During objective function optimization, the tree structure is introduced into the loss function. When building a new tree with features on the bank side, the information gain of the split point is computed directly without synchronization with the grid side. If the feature is on the grid side, the bank side computes G_j and H_j with the current prediction and the class label, and interacts with the data on the grid side via homomorphic encryption. The optimal split point is determined by the bank side by using the split gain calculation formula. If the best splitting point is on the local side, there is no need to send the splitting information to the grid side while determining the best splitting point. On the contrary, the split information is homomorphically encrypted by the grid side and sent to the bank side for resolution.

Due to the requirement of confidentiality of power data, the privacy protection for grid-side features consists of two parts. First, the homomorphic encryption technology ensures the security of the power data. Second, PCA is used to downscale the power data and convert the multivariate into a few key variables that can capture the vast majority of the original data information.

It is clear from the previous study that the gradient updates of each participant's involvement in the model aggregation affect the goodness of the overall model. As a result, we expect that trustworthy and high-quality model parameters can be selected for aggregation each time to produce superior training results. As a result, when the global model is updated, the central server can help verify the local updates of each member.

3.4 Global Model Updating

After each participant uploads the gradient update model parameters, an aiding node must use its own data as a verification set to confirm each participant's local update, and the central server determines each participant's update score. Once a certain qualifying update threshold is reached, the central server aggregates them into a new global model. Training is repeated until it converges or the accuracy of the model reaches a predetermined threshold.

The reputation value (R-score) of each participant in the model training is determined by a dual subjective logic model, which is used to validate each participant's gradient updates. Each iteration's

participant model updates are verified by a test set, and the quality of their uploaded parameters (Q-score) is directly measured by using the accuracy results of the test. The participants' R-score and qualitative Q-score are combined to produce the final update score, which also incorporates the effect of the time element.

The R-score attempts to assess the participants' Credibility. First, it is determines whether the model updates uploaded by each participant in each iteration are beneficial to the overall model. This is the most directly relevant question at hand. Credibility, implausibility, and uncertainty are three vectors that are used to quantify and describe the impact.

In the subjective logic model, the R-score of the central server c to the participant n_j is represented by the vector $q_{c\to n_j} = \left(b_{c\to n_j}, d_{c\to n_j}, u_{c\to n_j}\right)$, and it satisfies $b_{c\to n_j} + d_{c\to n_j} + u_{c\to n_j} = 1$. Among them, $b_{c\to n_j}$, $d_{c\to n_j}$ and $u_{c\to n_j}$ respectively represent the reliability, unreliability and uncertainty of the central server to the participant nodes.

The subjective logic model was then used to construct R-scores for the federated learning participants [33].

$$\begin{cases} b_{c\to n_j} = \left(1 - u_{c\to n_j}\right) \dfrac{\alpha_j}{\alpha_j + \beta_j} \\ d_{c\to n_j} = \left(1 - u_{c\to n_j}\right) \dfrac{\beta_j}{\alpha_j + \beta_j} \\ u_{c\to n_j} = 1 - p_{c\to n_j} \end{cases} \tag{16}$$

By establishing particular criteria, the central server confirms the dependability of the local model updates uploaded by the participant j. Model learning is regarded as a positive interaction event if the test accuracy of the participant's local model updates falls under the threshold, and vice versa for negative interaction events. The numbers of positive and negative interactions are represented by α_j and β_j, respectively. $p_{c\to n_j}$ is the probability of successfully transmitting the data model parameters.

These vectors can generate a reputation value that quantifies the credibility of the participants:

$$T_{c\to n_j} = b_{c\to n_j} + \gamma u_{c\to n_j} \tag{17}$$

Among them γ is a given constant, which represents the weight of uncertainty. The influence factors μ and θ of interaction events on reputation opinions are introduced to encourage high-quality data contributors to join the federated learning mission. μ represents the weight of positive interactions, θ represents the weight of negative interactions, among which $\mu > \theta$ and $\mu + \theta = 1$. The R-score expression is updated to:

$$\begin{cases} b_{c\to n_j} = \left(1 - u_{c\to n_j}\right) \dfrac{\mu\alpha_j}{\mu\alpha_j + \theta\beta_j} \\ d_{c\to n_j} = \left(1 - u_{c\to n_j}\right) \dfrac{\theta\beta_j}{\mu\alpha_j + \theta\beta_j} \\ u_{c\to n_j} = 1 - p_{c\to n_j} \end{cases} \tag{18}$$

It also considers the effect of time, since federated learning participants are not always trustworthy, and the more recent the interaction event, the greater the impact on the reputation score. Define a freshness fading function to describe how events affect reputation: $t(\varphi) = F^{Y-y}$. where $F \in (0, 1)$ is a given decay parameter, which is related to the freshness of the interaction. Y represents the time

slot, and $Y \in (1, Y]$. Add it to the opinion calculation, the R-score expression for a period of time is updated to:

$$\begin{cases} b_{c\to n_j} = \dfrac{\sum_{y=1}^{T} t(\varphi)\, q_{c\to n_j} \mu\alpha_j}{\sum_{y=1}^{T} t(\varphi)\left(\mu\alpha_j + \theta\beta_j\right)} \\ d_{c\to n_j} = \dfrac{\sum_{y=1}^{T} t(\varphi)\, q_{c\to n_j} \theta\beta_j}{\sum_{y=1}^{T} t(\varphi)\left(\mu\alpha_j + \theta\beta_j\right)} \\ u_{c\to n_j} = \dfrac{\sum_{y=1}^{T} t(\varphi)\, u_{c\to n_j}}{\sum_{y=1}^{T} t(\varphi)} \end{cases} \tag{19}$$

Therefore, the final R-score is expressed as:

$$T_{c\to n_j} = \frac{\sum_{y=1}^{T} t(\varphi)\left(b_{c\to n_j} + \gamma u_{c\to n_j}\right)}{\sum_{y=1}^{T} t(\varphi)} \tag{20}$$

The direct quality and the interaction time are taken into account when calculating the Q-score. The direct quality is the accuracy rate $q_{c\to n_j}$ achieved by the central server in recording the gradient update parameters in each iteration and testing them with the test set of the collaborating participants.

$$q_{c\to n_j} = \frac{\sum_{t=1}^{N} q_{c\to n_j}}{N} \tag{21}$$

where t is the number of iterations of federated learning, N is the total number of predefined iterations, and $t \in [1, N]$.

Participants' Q-scores fluctuated throughout the interaction time, which was determined in the same way as the R-score. Therefore, the Q-score is updated to:

$$q_{c\to n_j} = \frac{\sum_{y=1}^{T} t(\varphi)\, q_{c\to n_j}}{\sum_{y=1}^{T} t(\varphi)} \tag{22}$$

The acquired Q-score is then expressed as:

$$Q_{c\to n_j} = \frac{\sum_{y=1}^{T} t(\varphi)\, q_{c\to n_j}}{\sum_{y=1}^{T} t(\varphi)} \tag{23}$$

The current update score for the federated learning participant n_j is created by combining the R-score and Q-score with a specific weight, and this score is used as the evaluation index for the next round of collaborative participant selection. As a result, the total score is expressed as:

$$C_j^{final} = (1-\delta)\, T_{c\to n_j} + \delta Q_{c\to n_j} \tag{24}$$

where δ acts as a moderator to balance the Q-score and R-score, and $\delta \in [0, 1]$.

4 Empirical Analysis

4.1 Experiment Setup

The research subject uses actual enterprise credit data given by a regional branch of the Industrial and Commercial Bank of Lianyungang City to confirm the performance of the proposed model. The time period is from 2018.06 to 2020.06. It includes numerous financial and non-financial

characteristics of the lending company and the related evaluation findings. The average loan period is 6 months. Table 1 lists the credit evaluation indicators. 170 corporate loan data points were ultimately selected, including 37 non-performing loans and 133 regular loans, because data security necessitates desensitization of some indicators.

Table 1: Risk assessment indicator system

Indicator type	Indicator name	Symbolic representation
Solvency	Current ratio	X_1
	Quick ratio	X_2
	Cash ratio	X_3
	Equity ratio	X_4
	Debt ratio	X_5
Operating capacity	Total assets turnover ratio	X_6
	Fixed assets turnover ratio	X_7
	Accounts receivable turnover ratio	X_8
	Inventory turnover ratio	X_9
	Accounts receivable turnover days	X_{10}
	Inventory turnover days	X_{11}
Business growth and profitability	Operating income	X_{12}
	Primary business revenue growth rate	X_{13}
	Business taxes and surcharges	X_{14}
	Operating expenses	X_{15}
	Administrative expenses	X_{16}
	Financial expenses	X_{17}
	Total profit	X_{18}
	Operating profit margin	X_{19}
	Operating profit growth rate	X_{20}
	Net profit margin	X_{21}
	Net profit growth rate	X_{22}
	Total assets growth rate	X_{23}
	Return on net assets	X_{24}
Cash flow	Net cash flows from operating activities	X_{25}
	Net cash flows from investing activities	X_{26}
	Net cash flows from financing activities	X_{27}

(Continued)

Table 1: Continued

Indicator type	Indicator name	Symbolic representation
PBC credit reference	Number of accounts that have been overdue in the past two years	X_{28}
	Number of credit inquiries in the past month	X_{29}
	Number of institutions with credit transactions	X_{30}
	Number of institutions with current outstanding credit transactions	X_{31}
	Number of tax delinquency records	X_{32}
	Number of civil judgment records	X_{33}
	Number of enforcement records	X_{34}
	Number of administrative penalty records	X_{35}
	Number of accounts with liabilities of concern	X_{36}
	Number of accounts with non-performing liabilities	X_{37}
Enterprise foundation quality	Year of enterprise registration	X_{38}
	Type of enterprise	X_{39}
	Asset size	X_{40}
	Leadership quality	X_{41}
	Shareholders	X_{42}
	Staff quality	X_{43}
	Number of employees	X_{44}
Growth potential	Enterprise strategy	X_{45}
	Market capacity	X_{46}
	Research investment growth rate	X_{47}
	Industry output growth rate	X_{48}
	Recent year staff growth rate	X_{49}
Enterprise market position	Pricing power	X_{50}
	Market share	X_{51}
	Market competitiveness	X_{52}
Technical advantages	Maintenance and renewal of fixed assets	X_{53}
	Technological advancement	X_{54}
	Technological R&D capability	X_{55}

(Continued)

Table 1: Continued

Indicator type	Indicator name	Symbolic representation
Products competitiveness	Brand building	X_{56}
	After-sales service	X_{57}
	Market segmentation	X_{58}
	Degree of product diversification	X_{59}
Corporate governance	Shareholder control	X_{60}
	Connected transactions	X_{61}
	Corporate governance structure	X_{62}
Inventory management level	Inventory decline possibility	X_{63}
	Inventory structure rationality	X_{64}
	Inventory management policy	X_{65}
Production management level	Production and sales rate	X_{66}
	Production equipment utilization rate	X_{67}
	Quality management	X_{68}
Business development status	Exchange rate risk	X_{69}
	Raw material price risk	X_{70}
	Management expenses	X_{71}
	Sales scale growth	X_{72}
	Sales margin growth	X_{73}
	Policy support	X_{74}
	Industry risk	X_{75}

Precision, Recall, F1 score, KS value, and AUC are chosen as the model evaluation indicators during the feature selection procedure. The KS value is the degree of separation employed in the model to discriminate between positive and negative data for prediction, and a higher KS value indicates a stronger ability to discriminate.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{25}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{26}$$

$$F1score = \frac{2 * TP}{2 * TP + FP + FN} \tag{27}$$

$$\text{KS} = \left(\frac{TP}{TP + FN} - \frac{FP}{FP + TN} \right) \tag{28}$$

$$\text{AUC} = \frac{\sum I\left(P_{pos}, P_{neg}\right)}{M * N} \tag{29}$$

where TP denotes the number of true positive records, TN denotes the number of true negative records, FN denotes the number of false positive records, FP denotes the number of false negative

records, M denotes the number of positive samples in the data set, N denotes the number of negative samples, and $I\left(P_{pos}, P_{neg}\right)$ denotes the number of samples in the $M * N$ pair of samples for which the predicted probability of a positive sample is greater than the predicted probability of a negative sample, calculated as follows:

$$I\left(P_{pos}, P_{neg}\right) = \begin{cases} 1, P_{pos} > P_{neg} \\ 0.5, P_{pos} = P_{neg} \\ 0, P_{pos} < P_{neg} \end{cases} \tag{30}$$

4.2 Identification of Credit-Related Factors

To begin with, the data gathered from the power system is dedimensionalized using the PCA algorithm. The enterprise electricity consumption data used includes voltage level e_1, electricity consumption category e_2, contract capacity e_3, quarterly average change rate of contract capacity e_4, quarterly electricity consumption e_5, quarterly average change rate of electricity consumption e_6, monthly average change rate of load fluctuation e_7, monthly average load curve e_8, proportion of electricity consumption in valley section e_9, average load of electricity consumption in valley section e_{10}, line loss level e_{11}, electricity bill settlement e_{12}, importance level e_{13}, breach of contract electricity stealing record e_{14}, etc. There are fourteen features in total.

The original data sets were standardized using the corresponding principal component analysis function in MATLAB software, and then PCA was applied to the matrix consisting of the 14 indicators of the power data. Table 2 displays the results of the analysis.

Table 2: Covariance matrix eigenvalues, contribution rates and cumulative contribution rates

Component	Eigenvalue	Contribution/%	Cumulative/%
F_1	1.6174	32.84	32.84
F_2	1.3751	27.51	60.35
F_3	0.9562	19.48	79.83
F_4	0.6729	12.14	91.97
F_5	0.4328	8.03	100

The rule of the cumulative contribution of 90% or more was used to calculate the number of major components. The number of major components is five, as shown in Table 2. From the characteristic roots and contribution rates of the principal components, it can be seen that the characteristic root $\lambda_1 = 1.6174$, the characteristic root $\lambda_2 = 1.3751$, the characteristic root $\lambda_3 = 0.9562$, the characteristic root $\lambda_4 = 0.6729$. The cumulative variance contributions of the first four principal components reached 91.97%, so the first four indicators can be extracted, and they are recorded as F_1, F_2, F_3, and F_4, respectively. Through the eigenvectors corresponding to the first four characteristic roots, the linear expression of each principal component factor can be obtained respectively.

$$F_1 = 0.0104e_1 - 0.2163e_2 + 0.6428e_3 + 0.1125e_4 + 0.7813e_5 - 0.2840e_6 + 0.4601e_7 + 0.0072e_8 - 0.0905e_9 + 0.4239e_{10} - 0.1322e_{11} + 0.2342e_{12} + 0.1730e_{13} - 0.0726e_{14}$$

$$F_2 = 0.3626e_1 - 0.0283e_2 + 0.2071e_3 + 0.5026e_4 - 0.2004e_5 + 0.7037e_6 - 0.6483e_7 + 0.0701e_8 + 0.3302e_9 - 0.3140e_{10} + 0.3811e_{11} + 0.2003e_{12} - 0.0316e_{13} + 0.0726e_{14}$$

$$F_3 = 0.4733e_1 + 0.6702e_2 - 0.0480e_3 + 0.1042e_4 + 0.3123e_5 + 0.2394e_6 - 0.0362e_7 + 0.2129e_8 + 0.5539e_9 - 0.0230e_{10} + 0.6920e_{11} - 0.1385e_{12} + 0.0042e_{13} + 0.0726e_{14}$$

$$F_4 = -0.2173e_1 + 0.0479e_2 + 0.1930e_3 - 0.3028e_4 + 0.02042e_5 + 0.0953e_6 - 0.2215e_7 + 0.0350e_8 - 0.1102e_9 + 0.1047e_{10} - 0.4811e_{11} + 0.64e_{12} + 0.3051e_{13} + 0.8306e_{14}$$

Through the analysis of the above principal component factors, it is found that: In the principal component F_1, contract capacity, quarterly electricity consumption, monthly average load curve and valley section electricity consumption average load have higher weights, indicating that this principal component is significantly related to electricity consumption level, which can be defined as "electricity consumption scale factor". In the principal component F_2, the quarterly average change rate of contract capacity, the quarterly average change rate of electricity consumption and the monthly average change rate of load fluctuation have higher weights, indicating that this principal component is significantly related to electricity consumption fluctuations, which can be defined as "electricity stability factor". In the Principal Component F_3, the voltage level, the electricity consumption category, the proportion of electricity consumption in the valley section and the importance level have higher weights, indicating that this principal component is significantly related to the electricity consumption characteristics, which can be defined as "electricity consumption characteristic factor". In the principal component F_4, electricity bill payment and contract violation (including, for example, electricity stealing) records have a higher weight, indicating that this principal component is significantly related to the behavior of electricity consumption, which can be defined as "electricity consumption reputation factor". Therefore, it is feasible to use the first four principal components as the credit-related factors of the enterprise's energy consumption.

Second, based on the obtained internal bank data set, the mRMR algorithm is applied to extract characteristics from the information of the borrowing companies and select the indicators that can most comprehensively reflect the financial status of the companies to construct the mRMR feature set. After obtaining the pre-selected feature set using the mRMR-PCA hybrid feature selection algorithm suggested in this research, the important measure of the feature variables is then measured by the XGBoost model. The experiment uses ten-fold cross-validation to determine the optimal parameters of the XGBoost model. The following parameters have been set: max $\_depth = 6$, $n\_estimators = 142$, $learning\_rate = 0.2$. Taking the Ave_gain of features as the feature importance measure, the top 30 features are as follows: X_{40}, X_{35}, X_5, X_{22}, X_8, X_9, X_3, X_{13}, X_2, X_{24}, F_2, X_{28}, F_4, X_1, X_6, X_{48}, X_{52}, F_1, X_{21}, X_{75}, X_{25}, X_{54}, X_{23}, X_{41}, X_7, X_{30}, X_{20}, X_{38}, X_{47}, X_{67}.

The pre-selected feature set is fed into the XGBoost model, and the best feature set is selected using the AUC, an assessment metric for the binary classification model. Fig. 2 depicts the relationship between feature dimension j and AUC. The 24 features with the highest relevance rankings are used as the ideal feature set for evaluating corporate credit, as shown in the Fig. 2, where the AUC of the XGBoost model is maximum when the feature dimension is 24. This indicates that the best effect is achieved at this point.

In this research, RF and LR are used as controls to indicate the ability of the XGBoost model to identify the best set of features for corporate credit scoring via the feature significance ranking approach. After model training, both classifiers can produce feature significance rankings. The current data sets are randomly sampled into training and test sets, and seventy percent of the data sets are fed into LR, RF, and XGBoost to train the models. Then, the trained models are used to predict the default cases of the test set samples, and the evaluation indicators for the three models under each sample set are recorded. Table 3 shows the final results.

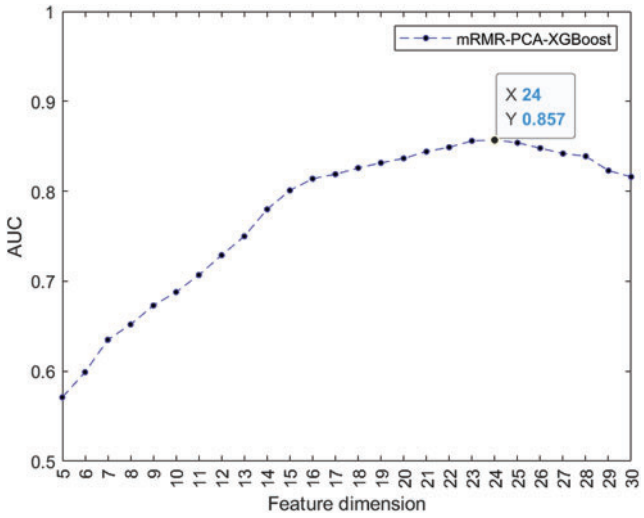

Figure 2: Variation of AUC with feature dimension

Table 3: Comparison of model predictions outcomes

Algorithm	Accuracy	Precision	Recall	F1 score	KS value
XGBoost	0.8693	0.7468	0.7542	0.7397	0.513
Random Forest	0.8525	0.7213	0.7020	0.7114	0.495
Logistic Regression	0.8047	0.6413	0.7726	0.7028	0.461

Table 3 suggests that the XMP feature selection method outperforms the RF and LR models in terms of accuracy, precision, and F1 score. The KS value demonstrates that the XGBoost model is more effective than RF and LR in distinguishing between positive and negative samples, and it has a higher degree of discrimination in judging whether a user is in default or not. The comprehensive analysis above shows that the XMP feature selection algorithm performs well.

4.3 Credit Evaluation Model Prediction

The credit assessment model for MSEs based on the SecureBoost algorithm is built using the 24 features of the most optimal collection after screening the indicators using the feature selection method provided in this study. The probability of customer default is predicted using the model from this study as well as the conventional credit risk prediction methods employed by commercial banks and the current methodologies RF [34] and LS-SVM [35]. The prediction results are displayed in Fig. 3.

The experiments suggest that the model proposed in this research has a superior predictive ability than the established credit risk prediction techniques. The higher recall rate shows that the model is more effective in identifying defaulters. The higher recall rate and accuracy of the proposed model for high-risk enterprise samples showed the validity of the credit risk assessment model we developed for the joint modeling of electricity data, bank data, and enterprise credit data.

The results of comparing the suggested model with the traditional federated learning framework [36] are shown in Fig. 4. It can be observed that there isn't any discernible difference in the overall

accuracy between the model which is trained by the scheme proposed in this paper and the typical federated learning scheme. However, the local update screening strategy proposed in this research, which selects trustworthy and high-quality model parameters for aggregation each time, can significantly improve the efficiency of federated learning iterations.

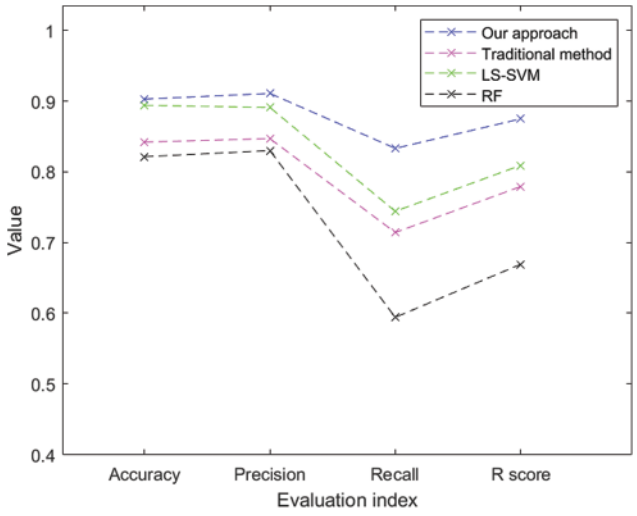

Figure 3: Performance comparison of various models for credit assessment

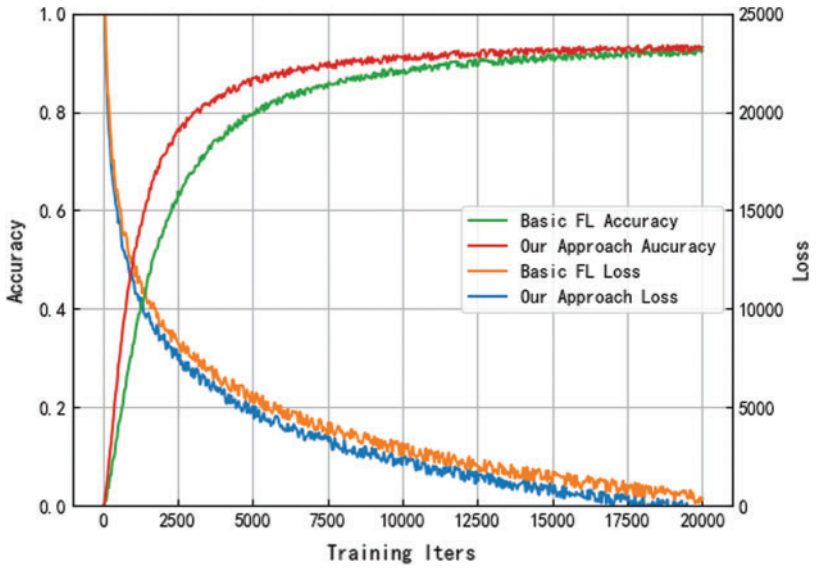

Figure 4: Variation of model accuracy and loss function with the round of iterations

The analysis is conducted by contrasting the actual lending business models of commercial banks, which helps to further demonstrate the effectiveness of the proposed model in predicting the credit rating of lending firms. Ten-fold cross-validation was employed numerous times and the following parameters have been set: max $\_depth = 9$, $n\_estimators = 174$, $learning\_rate = 0.1$. Table 4 displays the outcomes of the partial predictions for the test set.

Table 4: Enterprise credit assessment under different models

Enterprise	Actual model	Proposed model
Agriculture, forestry, livestock and fishery E1	75.84	73.37
Agriculture, forestry, livestock and fishery E2	64.49	60.26
Industrial manufacturing E3	63.87	59.04
Industrial manufacturing E4	61.64	57.83
Industrial manufacturing E5	74.47	72.29
Information technology service industry E6	68.55	69.84
Information technology service industry E7	86.92	85.57
Information technology service industry E8	87.95	86.22
Information technology service industry E9	74.17	71.32
Culture and education industry E10	80.74	79.76
Culture and education industry E11	79.17	67.81
Culture and education industry E12	66.71	62.03
Wholesale and retail industry E13	77.71	75.73
Wholesale and retail industry E14	65.76	61.15
Wholesale and retail industry E15	73.09	73.22

As shown in Table 4, the proposed model and the actual model's enterprise credit scores are generally consistent in ranking, indicating that the enterprise credit rating predicted by the model in this work is relatively objective. For the default samples of E3 and E4, we found through the subsequent empirical study, we found that the two enterprises' own market competitiveness and technological progress are lower than the same level in the industry, and the electricity consumption stability factor is relatively low. The XMP feature selection integrates enterprise power consumption factors, so the credit scores of the above enterprises are lower than the traditional method. For the defaulting enterprises E2 and E14, the proposed model also identifies them well and assigns a relatively low credit score.

For commercial banks, effective prediction of enterprise credit risk is a crucial issue for efficient credit operations. Through the above analysis, the proposed model can provide more accurate results on the credit risk assessment of MSEs and effectively discriminate the risks.

5 Conclusion

This paper develops a privacy preserving credit risk assessment model for MSEs based on the currently popular federated learning to address the issues of bank-enterprise information asymmetry and weak risk identification in the credit scenario of MSEs. For the model aggregation phase, we propose a global model update strategy that can filter high-quality local models. For the feature selection phase, we provide the XMP feature selection algorithm appropriate for corporate credit scenarios. Through the empirical analysis, the results illustrate that the credit risk assessment model for MSEs proposed in this research, which is based on federated learning and feature selection, achieves a higher accuracy rate while protecting the privacy of each participant. The results of this research

will serve as a reference for commercial banks in developing a credit risk assessment system, and for electric power companies in deriving value from data related to marketing.

Federated learning and blockchain have a common application foundation, and a trusted network of multi-party cooperation is realized through technical consensus, which has good complementarity. Future related work will further integrate federated learning with blockchain and design a more complete federated learning mechanism that protects user privacy. This will help solve the problem of single point dependency and contribution allocation in federated learning. This work will also explore business innovation scenarios that are broadly applicable to different industries while fully exploiting the value of data.

Acknowledgement: We appreciate NUIST providing us with the chance to carry out this research.

Funding Statement: This research was funded by the State Grid Jiangsu Electric Power Company (Grant No. JS2020112) and the National Natural Science Foundation of China (Grant No. 62272236).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] N. Gulsoy and S. Kulluk, "A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3, pp. e1299, 2019.

[2] X. Wang and J. Zhang, "On the bank credit rationing and loan of small and medium-sized enterprises," *Economic Research Journal*, vol. 4, no. 7, pp. 68–75, 2003.

[3] T. X. Sheng and C. L. Fan, "Fintech, optimal banking market structure, and credit supply for SMEs," *Journal of Financial Research*, vol. 480, no. 6, pp. 114–132, 2020.

[4] X. Huang, X. Liu and Y. Ren, "Enterprise credit risk evaluation based on neural network algorithm," *Cognitive Systems Research*, vol. 32, no. 1, pp. 317–324, 2018.

[5] P. Golbayani, D. Wang and I. Florescu, "Application of deep neural networks to assess corporate credit rating," arXiv preprint arXiv:2003.02334, 2020.

[6] Y. Xia, C. Liu, Y. Li and N. Liu, "A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring," *Expert Systems with Applications*, vol. 78, pp. 225–241, 2017.

[7] Y. Zhang, G. Chi and Z. Zhang, "Decision tree for credit scoring and discovery of significant features: An empirical analysis based on Chinese microfinance for farmers," *Filomat*, vol. 32, no. 5, pp. 1513–1521, 2018.

[8] J. Sun, J. Lang, H. Fujita and H. Li, "Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates," *Information Sciences*, vol. 425, no. 4, pp. 76–91, 2018.

[9] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua and S. Guo, "Protection of big data privacy," *IEEE Access*, vol. 4, no. 4, pp. 1821–1834, 2016.

[10] C. P. Ge, W. Susilo, J. Baek, Z. Liu, J. Y. Xia *et al.,* "A verifiable and fair attribute-based proxy re-encryption scheme for data sharing in clouds," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 7, pp. 1–12, 2021.

[11] Y. Ren, K. Zhu, Y. Q. Gao, J. Y. Xia, S. Zhou *et al.,* "Long-term preservation of electronic record based on digital continuity in smart cities," *Computers, Materials & Continua*, vol. 66, no. 3, pp. 3271–3287, 2021.

[12] K. A. Houser and W. G. Voss, "GDPR: The end of Google and Facebook or a new paradigm in data privacy," *Rich JL & Tech*, vol. 25, no. 3, pp. 1, 2018.

[13] J. Konečný, H. B. McMahan, D. Ramage and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," arXiv preprint arXiv:1610.02527, 2016.

[14] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh *et al.,* "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.
[15] B. W. Chi and C. C. Hsu, "A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model," *Expert Systems with Applications*, vol. 39, no. 3, pp. 2650–2661, 2012.
[16] S. Jones, D. Johnstone and R. Wilson, "An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes," *Journal of Banking & Finance*, vol. 56, no. 11, pp. 72–85, 2015.
[17] S. Jones, D. Johnstone and R. Wilson, "Predicting corporate bankruptcy: An evaluation of alternative statistical framework," *Journal of Business Finance & Accounting*, vol. 44, no. 1–2, pp. 3–34, 2017.
[18] C. Nguyen, "The credit risk evaluation models: An application of data mining techniques," in *Proc. SAIS*, Georgia, USA, pp. 36, 2019.
[19] H. Li, Y. Cao, S. Li, J. Zhao and Y. Sun, "XGBoost model and its application to personal credit evaluation," *IEEE Intelligent Systems*, vol. 35, no. 3, pp. 52–61, 2020.
[20] Y. X. Yang and W. J. Pang, "The development status, problems and countermeasures of big data credit investigation industry in China," *Credit Reference*, vol. 38, no. 2, pp. 49–52, 2020.
[21] M. Ala'raj and M. F. Abbod, "A new hybrid ensemble credit scoring model based on classifiers consensus system approach," *Expert Systems with Applications*, vol. 64, pp. 36–55, 2016.
[22] W. Zhang, H. He and S. Zhang, "A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring," *Expert Systems with Applications*, vol. 121, pp. 221–232, 2019.
[23] L. X. Cui, L. Bai, Y. C. Wang, X. Jin and E. R. Hancock, "Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection," *Pattern Recognition*, vol. 114, pp. 107835, 2021.
[24] G. R. Chen, M. R. Mu and R. Zhang, "Realization of communication fraud identification model based on federated learning," *Telecommunications Science*, vol. 36, no. S1, pp. 304–310, 2020.
[25] P. Xu, J. J. He and X. Y. Yue, "Study on management model of prevention and control of new coronary pneumonia (COVID-19) in colleges and universities based on marginal learning and federal learning," *Forum on Contemporary Education*, vol. 19, no. 2, pp. 76–82, 2020.
[26] L. Z. Zheng, "The exploration of data security based on federal learning in banking," *China Financial Computer*, vol. 20, no. 9, pp. 22–26, 2020.
[27] C. K. Wang and J. Feng, "An applied study of federal learning in the insurance industry," *Journal of Vocational Insurance College*, vol. 34, no. 1, pp. 13–17, 2020.
[28] K. Cheng, T. Fan, Y. Jin, Y. Liu, T. Chen *et al.,* "Secureboost: A lossless federated learning framework," *IEEE Intelligent Systems*, vol. 36, no. 6, pp. 87–98, 2021.
[29] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
[30] S. Wold, K. Esbensen and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987.
[31] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. of the Forty-first Annual ACM Symp. on Theory of Computing*, Maryland, USA, pp. 169–178, 2009.
[32] Z. Brakerski, "Fully homomorphic encryption without modulus switching from classical GapSVP," in *Annual Cryptology Conf.*, Berlin, Heidelberg, Springer, pp. 868–886, 2012.
[33] Y. L. Lu, X. H. Huang, Y. Y. Dai, S. Maharjan and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4177–4186, 2019.
[34] W. Y. Qiu, S. W. Li, Y. M. Cao and L. Hua, "Credit evaluation ensemble model with self-contained shunt," in *2019 5th Int. Conf. on Big Data and Information Analytics (BigDIA)*, Kunming, China, IEEE, pp. 59–65, 2019.

[35] F. T. Wang, L. H. Ding, H. X. Yu and Y. J. Zhao, "Big data analytics on enterprise credit risk evaluation of e-Business platform," *Information Systems and e-Business Management*, vol. 18, no. 3, pp. 311–350, 2020.
[36] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," *Artificial Intelligence and Statistics. PMLR*, vol. 54, pp. 1273–1282, 2017.