Research and Practice of Financial Credit Risk Management Based on Federated Learning

Yan Li, Guihua Wen

***Abstract*—*Currently, the application of big data and* artificial intelligence (AI) in financial credit risk management is a research hot spot. Most research has focused on using data and AI algorithms to improve modeling performance, ignoring the difficulty of data sharing. Federated learning has become an effective solution for achieving the "available and invisible" goal of data sharing under the premise that the data do not come out of the local area and meet global data compliance. However, we must find a solution for the poor modeling performance caused by data heterogeneity to truly achieve the goal of credit risk management in the financial industry. This paper proposes an application scheme of federated learning in the field of financial credit risk management. It presents an experimental comparison and analysis to verify the feasibility of the application. The antagonistic verification results show the existence of non-independent and identically distributed (non-IID) data. Moreover, using the Chi-square test method, we detail the modeling, analysis, and evaluation of heterogeneous data with non-IID characteristics. The experimental results demonstrate a performance improvement of 14% by the proposed application framework and method in the financial credit risk management of small banks with few data samples.**

***Index Terms*—Credit risk management, federated learning, financial engineering, non-IID**

I.INTRODUCTION

THE financial industry's ever-growing wealth of data has continually presented new opportunities to develop and improve the industry [1]. However, obtaining and utilizing financial data are difficult. First, with the increasing importance of data privacy worldwide, information security regulations have made it difficult to fulfill the legal requirements necessary to obtain financial data [2]. Furthermore, to protect customers and avoid losses, the financial industry has been unwilling to share customer information, leading to "data silos" [3]. Federated learning is an effective solution to these issues, meeting the legal requirements and ensuring customer privacy and protection [4, 5]. Through the application of federated learning technology, the data elements that financial institutions seek can be shared across departments, institutions, and industries.

For federated learning to be used in credit risk management, however, a solution is needed for the problem of heterogeneity [6-8]. This term refers to statistical heterogeneity and system heterogeneity, that is, equipment and model heterogeneity [9]. Because financial institutions have extremely strict regulations in the interest of business continuity security, it can be assumed that the experiment environment is ideal, that is, clients are always reachable, clients' local data are unchanged, and the model can be kept consistent, assuming that there is no system heterogeneity. However, statistical heterogeneity exists widely owing to the huge differences in the amount of data owned by different clients. This is caused by the differing scale and geographic locations of different institutions and ultimately results in a large amount of data being non-independent and identically distributed (non-IID). This presents some challenges in the convergence of data modeling and analysis model training, which may significantly affect the training performance of the model. This is the core problem that federated learning needs to overcome in the field of credit risk management. It is also a key problem that must be solved for the further development and utilization of federated learning.

The main contributions of this study are as follows: first, we use federated learning to solve the credit risk identification problem of multiple banks. We clarify the technical steps and methods of using federated learning to solve real problems, such as obtaining an effective method of data preprocessing. The second contribution is to design a specific federated learning scheme, which has achieved a similar effect of centralizing bank data for modeling. In practice, it has been proved that federated learning can not only protect privacy data but also achieve a better risk identification effect, which strongly supports such typical applications. The third contribution is to prove that the effect of federated learning modeling is better than that of any single bank modeling, which provides a feasible method for bringing more banks into federated learning.

II. RELATED WORK

The concept of federated learning was first proposed by Google in 2016 [10]. It is a framework based on distributed machine learning, in which the clients cooperatively train the model under the coordination of the central server, and the training data are kept locally without being uploaded to the data center [11]. Unlike general machine learning models, federated learning can use datasets distributed on participating clients to cooperatively train the sharing model, which has the default attribute of confidentiality [12]. According to different datasets, federated learning can be divided into three categories: horizontal federated learning [13], vertical federated learning [13, 14], and federated transfer learning [15].

Federated learning has been widely used in finance,

Manuscript received April 25, 2022; revised November 28, 2022.

Yan Li is a PhD student of the School of Computer Science and Engineering, the South China University of Technology, Guangzhou, China.(email：13760815161@139.com)

Guihua Wen is a professor of the South China University of Technology, Guangzhou, China. (email: crghwen@scut.edu.cn)

medical care, keyboard prediction, and other fields [16-18]. For example, in the field of credit card anti-fraud and cross-bank anti-money laundering, companies such as JP Morgan Chase, IBM, and various academic institutions have conducted theoretical discussions and modeling tests of federated learning.

In China's financial industry, the application practice of federated learning mainly focuses on two types of business scenarios: intelligent marketing and intelligent risk control. Through dimension expansion and data feature splicing, the ability of precise marketing and risk quantification can be improved. In recent years, many domestic enterprises have collaborated with financial institutions, medical companies, and local governments to carry out research and applications for federated learning [19-21]. At present, federated learning platforms that have entered the pilot stage include the Tencent Secure Federated Learning Application Service Platform (FLAS); the open-source federated learning framework FATE, from WeBank; and the open-source federated learning framework PaddleFL, based on Baidu Paddle [9]. These platforms have been jointly tested in marketing and risk control scenarios with many commercial banks, internet enterprises, consumer financial institutions, and insurance companies.

There are still many problems in the application scenario of federated learning, including incentive mechanisms [22], security [23-25], and data heterogeneity [26]. This paper focuses on solving the problem of data heterogeneity. Individual clients and their devices generate, process, and collect data on the network in different distributed methods, creating large variations in data volume and characteristics. Thus, the training data generated locally by each client are non-IID. However, mainstream machine learning and artificial intelligence algorithms are mainly based on the assumption that data are independent and identically distributed (IID). If the training datasets on different clients participating in federated learning are non-IID, challenges arise in the process of data modeling and model analysis [27-29]. The non-IID data distribution characteristics slow the model training speed, decrease model accuracy, and make more communication overhead necessary, thus increasing the difficulty of federated learning model training. Therefore, finding solutions to model, analyze, and evaluate the heterogeneous data with the characteristics of non-IID data has become an important research topic.

III. Theory and Methods

China's large and medium banks usually have relatively mature credit risk assessment capabilities and ample financial data samples. If the data and modeling capabilities of large banks can be exported to small banks through federated learning, without the outflow of data assets and model assets, it can help improve small banks. This is because small banks have less data, fewer fraud samples, and inadequate risk control abilities; so, they can benefit from using the federated learning modeling platform to improve the overall credit risk control ability of the banking industry.

To meet the above aims, we established a scheme designating Bank A as a joint-stock bank, belonging to a large-sized bank, and Bank B as a city commercial small bank. Banks A and B have jointly explored cooperation in the field of credit risk control and carried out the practice of federated learning to establish a federated learning platform.

This scheme uses the federated learning modeling method to help banks with credit risk identification. The selected information is distributed in different banks and includes customer information, credit contracts, loan receipts, loan extension information, and deposit and loan account information. By building a federated learning model for corporate loan risk identification, banks can predict whether a normal corporate loan will be risky in the coming month. An overview of the proposed system architecture is shown in Fig. 1.

Although there may be only a few overlapping customers between different banks, the overall customer data are similar enough to cooperatively establish the machine learning model through horizontal federated learning to realize risk prevention and control [30, 31]. According to the data characteristics of corporate loans and the principle of federated learning technology, we chose the logistic regression modeling scheme based on horizontal federated learning.

To verify the scheme, we devised two experimental hypotheses: (1) the performance of federated learning modeling with data distributed in two banks is similar to or better than the performance of centralized modeling (data aggregated in one place), proving the effectiveness of the federated learning modeling scheme; (2) the performance of federated learning modeling is better than that of local independent modeling (i.e., local modeling by each individual bank) because federated learning modeling can utilize more data and features and has high practical value.

Because Bank A is a national joint-stock bank and Bank B is a city commercial bank with a business scope in a specific province, the customer groups and sample size vary. Therefore, the data of the two banks are heterogeneous. The proposed scheme attempts to solve the problem of heterogeneity and show that the federated learning modeling results are similar to or better than the local independent modeling results [32].

IV. Experiment and Result Analysis

In this experiment, Banks A and B selected transaction flow data in the same time period and meeting the same filter criteria. The effectiveness of the design scheme was verified through data preprocessing, algorithm model design, and modeling experiments.

A. Overview of data

According to the established experiment plan, nearly 600,000 samples of transaction data were collected, analyzed, and processed (over 540,000 samples from Bank A and over 40,000 samples were obtained from Bank B), and data showing a strong correlation with corporate loan business were screened out, as shown in Table 1. Features were divided into four dimensions: loan-related static information, dynamic change information of the loan, loan customer sub-account capital flow, and total capital flow of the loan customer's other accounts. After a series of various

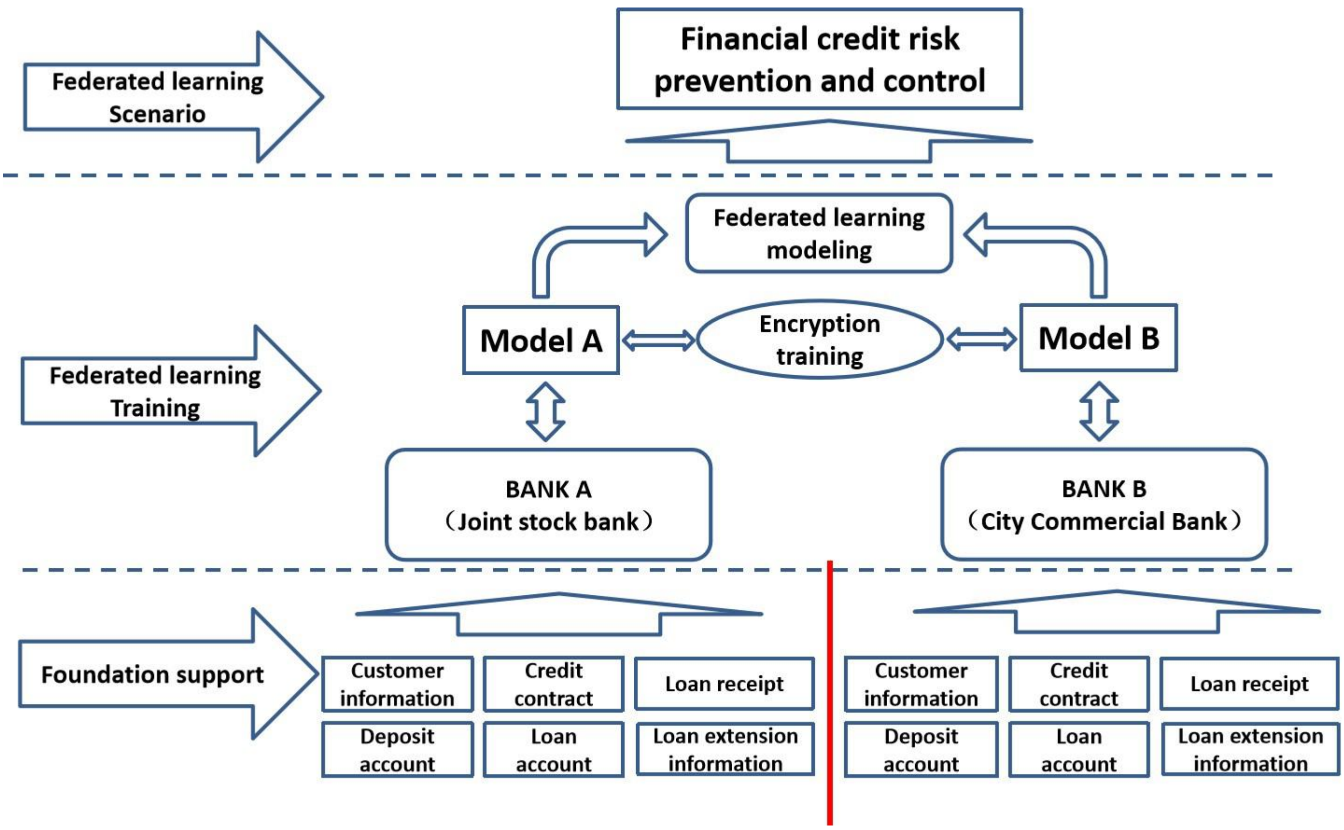

Fig. 1. System architecture of the proposed model of federated learning

TABLE I

DATA TABLE INFORMATION

Loan-related static information	Dynamic change information of the loan	Sub-account capital flow information of loan customers	Total capital flow information of other accounts of loan customers
Credit contract table	Statement of IOUs for corporate credit business	Sub-account table of corporate credit	Sub-account of corporate demand deposits
Corporate customer table	Loan extension table	Sub-account detail record of corporate credit	Sub-account detail of corporate demand deposits

operations involving filtering and joining in the time period of January to June 2020, 103 valid fields were selected from 263 fields in eight tables.

B. Preprocessing of data

2.1. Data cleaning

(a) **Standardization of data.** Owing to different data definition standards between banks, slight variations in data descriptions lead to different statistical values for describing the same feature. For example, in the Uniform Credit Flag field, data from Bank A are denoted with "Yes," and data from Bank B are denoted with "Have"; so, "Yes" and "Have" were uniformly mapped to "Yes."

(b) **Null value processing.** There is inevitably some missing information when collecting data; so, it is necessary to process null values. For numeric null values, if 0 was meaningless, it was filled with 0, and if 0 was meaningful, it was filled with special values. For non-numeric null values, we filled them with "None."

(c) **Avoiding sample duplication.** To avoid the influence of duplicate samples on the model, we removed duplicate data, and only one copy of identical data was reserved.

2.2. Characteristic engineering

(a) **Numericalization of features**. For features whose values were non-numeric, the non-numeric values were converted into numeric values by mapping relationships. For example, according to the mapping relationship of {Agriculture: 0, IT: 1, Finance: 2, Manufacturing: 3}, a set of values of industry characteristics {Agriculture, IT, Finance, Agriculture, Manufacturing} could be processed as {0,1,2,0,3}.

(b) **Discrete intervals of features**. We segmented the value of features into discrete intervals. For example, for a group of value results in the income feature denoted as {3000, 4500, 5000, 9000, 10000, 12000, 15000}, we created three discrete intervals of low, medium, and high, and the division results are thus {low, low, low, medium, medium, high, high}.

(c) **One-hot coding**. One-hot coding is a method of converting category variables into simple data forms for machine learning models. By representing category variables as binary vectors, we could represent each numerical value as a binary vector. For example, for a group of classification variables {0,1,2}, three coded values of "100," "010," and "001" could be obtained through one-hot coding.

2.3. Data sampling

In this modeling scenario, Bank A had 1,514 positive samples and 542,920 negative samples, while Bank B had 82 positive samples and 45,792 negative samples. Between the two banks and within the positive and negative samples of each individual bank, the data were extremely unbalanced. The positive-to-negative sample ratio of Bank A was 1:10 and that for Bank B was 1:12. Therefore, when building the model, we proportionally under-sampled negative samples to balance the data and meet the needs of algorithm modeling.

C. Federated learning model design

Based on the business scenario of corporate loan risk identification, a model was constructed to predict whether a normal corporate loan will have risks in the coming month. To explore the data characteristics of corporate loans and the advantages of federated learning technology, we utilized logistic regression to train the local model of Bank A and Bank B. Logistic regression is widely used in the financial industry and is one of the common methods to solve the two-category problem. Moreover, it is highly interpretable; so, it can meet the demands of regulators, customers, and bank staff in the financial scenario.

Federated learning modeling then involves the coordination of multiple clients (i.e., Bank A and Bank B in this paper) to jointly learn a global model without data sharing. We used the FedAvg as a baseline, a widely used standard method.

3.1. Local modeling: Logistic regression

Logistic regression is a generalized linear model. The original range R of linear regression is mapped to the interval [0,1] by the sigmoid function. When the value is greater than the critical value, it is classified as one type, and when it is less than the critical value, it is classified as another type [34].

Its function definition is as follows:

$$P = \frac{1}{1+e^{-\left(\beta^T x+\beta_0\right)}}, \tag{1}$$

Where β is the coefficient vector for the single input x, and β_0 is a constant term.

For the linear regression model, as shown in (2),

$$y_i = \beta_0 + \beta^T x_i + \varepsilon_i \qquad i = 1,2,3,\cdots,n, \tag{2}$$

where y_i is a response variable, $y_i \in R$; x_i is a p-dimensional variable, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$; β is a parameter of the model, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$; and ε_i is an IID random error. The value of β, estimated by the least square estimation, is commonly used, making $\sum_{i=1}^{n}\left(y_i - \beta^T x_i\right)^2$ the minimum value and yielding the estimate $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$, in which X is a matrix, and Y is the response vector, where $Y = \left(y_1, y_2, \cdots, y_n\right)$.

Dependent variables in logistic regression can be expressed in many forms, and if we move $\beta^T x+\beta_0$ to one side of the equation in (1), the expression can be expressed by independent variables, as follows:

$$\frac{p}{1-p} = e^{\beta^T x+\beta_0}. \tag{3}$$

Both sides of the equation take a logarithm at the same time, as follows:

$$\ln\frac{p}{1-p} = \beta^T x + \beta_0. \tag{4}$$

If the probability of an event happening is defined as $p = p(y=1)$, the probability of the event not happening is $1-p = 1-p(y=1) = p(y=0)$. We obtain the occurrence ratio as

$$odds = \frac{p}{1-p}. \tag{5}$$

Suppose that there are n corporate credit loan data, (x_i, y_i) $i \in 1,2,\cdots,n$, in which $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ is a feature vector of loan customers; y_i is a class label, where $y_i \in \{0,1\}$; Class 0 represents normal customers; and Class 1 represents abnormal customers.

The probability of each observed sample is given by

$$p\left(y_i, x_i\right) = p\left(y_i = 1 \mid x_i\right)^{y_i}\left(1 - p\left(y_i = 1 \mid x_i\right)\right)^{1-y_i}. \tag{6}$$

The maximum likelihood estimation method is used to estimate the parameters, and the likelihood function is given by

$$L(\beta) = \prod_{i=1}^{n} p\left(y_i = 1 \mid x_i\right)^{y_i}\left(1 - p\left(y_i = 1 \mid x_i\right)\right)^{1-y_i}. \tag{7}$$

3.2. Global model: FedAvg

To construct the global model without sharing the data of Bank A and Bank B, it is necessary to use a certain algorithm to aggregate the two banks' local models into a more accurate global model. We use the FedAvg model, the essence of which is to integrate and average the weights trained by each client. For the specific algorithm, see Algorithm 1.

D. Model evaluation index

To identify as many risky corporate loans as possible and ensure prediction accuracy, the model evaluation mechanisms are as follows:

1) Compare the recall rate of the model when the accuracy rate of the model is close to but not less than 80%. The higher the recall rate of the model is, the better the

model performance is;

2) To factor in the precision rate and recall rate of the classification model simultaneously, use the F1 score to measure the effectiveness of the model. The F1 score can be regarded as a weighted average of the model precision and recall. Its maximum value is 1, and its minimum value is 0. The larger the F1 score is, the better the model is.

Algorithm 1: Averaging for federated learning between Bank A and Bank B

Server input: role: arbiter, work_mode: 1
Client i's input: local labeled feature data A and B train data, local step size η
Client-created data:

```
    def var_data(x):
        z (derived var) created by x
    return z
    Bank A create var:
        extract wide table tableA from Bank A
        varA = var_data (tableA)
    Bank B create var:
        extract wide table tableA from Bank B
        varB = var_data (tableB)
```

Server executes:

    initialize w_0
    for each round t = 1, 2, … **do**
        m ← max(C · K, 1)
        S_t ← (random set of m clients)
        for each client $k \in S_t$ in **parallel do**

$$w_{t+1}^{k} \leftarrow \text{Client Update}(k, w_t)$$

$$w_{t+1}^{k} \leftarrow \sum_{k=1}^{k} \frac{n_k}{n} w_{t+1}^{k}$$

Client update(k, w, f): // Run on client k in [A, B], and f in [varA, varB]

    $B \leftarrow$ (Split P_k into batches of size B)
    for each local epoch *i* from 1 to *E* **do**
        for batch $b \in B$ **do**
            $w \leftarrow w - \eta \nabla \ell(w, b)$
        Return w to server

E. Experimental verification of federated modeling

5.1. Experiment I: "Special mention," positive samples only

Banks always classify loans into five categories: standard, special mention, substandard, doubtful, and loss.

In this experiment, only loans classified as "special mention" were selected as an abnormal positive sample. The training set comprised the data of Bank A and Bank B from January to May 2020, and the test set comprised the data from the month of June in 2020. We compared the results of local independent modeling, centralized modeling, and federated learning modeling with the evaluation standard of the recall rate, as shown in Fig. 2.

Under the condition of insufficient positive samples, federated modeling was unsuccessful; the performances of both centralized modeling and federated learning modeling were far worse than that of local independent modeling. Furthermore, the performances of centralized modeling and federated learning modeling were relatively close in Bank A, but could not be determined in Bank B because the result was 0.

Based on Experiment I, Experiment II expanded the positive sample data. While keeping all other conditions unchanged, we expanded the abnormal class from "special mention" to four statuses: "special mention," "substandard," "doubtful," and "loss." The recall rates are shown in Fig. 3, and the F1 scores are shown in Fig. 4.

Under the condition of sufficient positive samples, the following conclusions can be drawn:

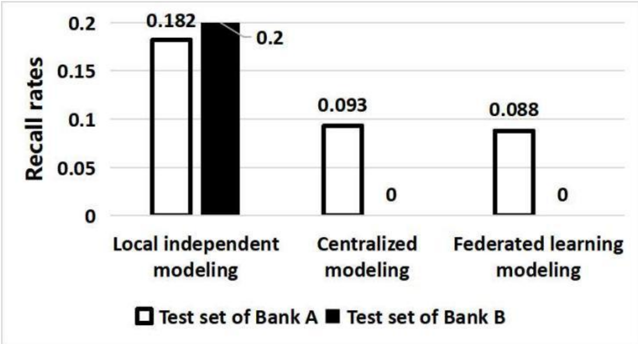

Fig. 2. Comparison of recall rates between Bank A and Bank B (Experiment I)

5.2. Experiment II: Expanding positive samples

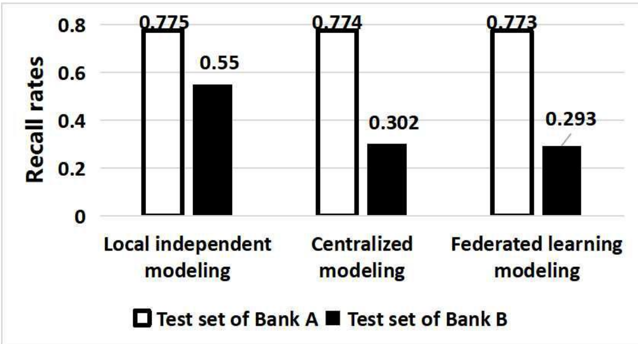

Fig. 3. Comparison of recall rates between Bank A and Bank B (Experiment II)

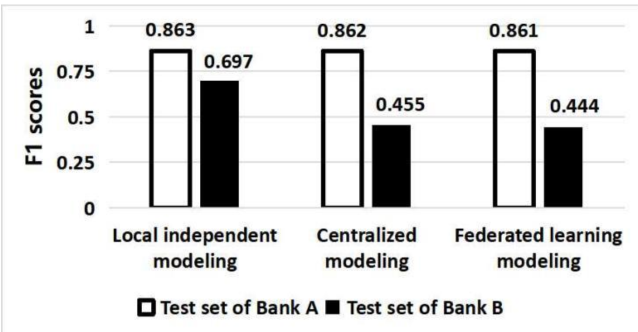

Fig. 4. Comparison of F1 scores between Bank A and Bank B (Experiment II)

1) From the perspective of recall rates, the performance of both centralized modeling and federated learning modeling is significantly improved, approaching but not exceeding that of the local independent modeling. This suggests that increasing the number of positive samples can increase the success rate of federated learning modeling,

although the modeling performance of federated learning is not significantly better than that of local independent modeling.

2) From the perspective of F1, the performance of both centralized modeling and federated learning modeling is also worse than that of local independent modeling.

3) Regardless of the recall rate or F1, the performance of federated learning modeling and centralized modeling is similar in both banks' data.

5.3. Experiment III: Test of data identically distributed

The datasets of the two banks participating in federated learning have different data structures and amounts of data, which brings about fairness issues in federated learning. For example, in the process of federated learning and training, the model may be more biased toward banks with larger data volumes. Therefore, it is challenging to explore the impact of the differing data volumes and data structures on the performance of the federated learning modeling system in order to improve the fairness problem in the framework.

Following the results of the experiments I and II, we must further analyze the findings, knowing that the data are not IID. Experiment III evaluated whether the data of two banks accord with the same distribution by the antagonistic verification method.

The antagonistic verification method is a common method to test whether the datasets are distributed identically. The experiment assumed that the data of two parties were in different feature distribution spaces (0 and 1), and a classification model was constructed to distinguish the data of two parties. If the area under the curve (AUC) value output by the model was close to 1, the distribution of the two datasets was different. If it was close to 0.5, the data were identically distributed.

Procedure:

(1) A new label column marked as 0 was added for Bank A data, and a new label column marked as 1 was added for Bank B data;

(2) Marked data of Bank A and Bank B were merged into a dataset, which was re-divided into a training set and a test set;

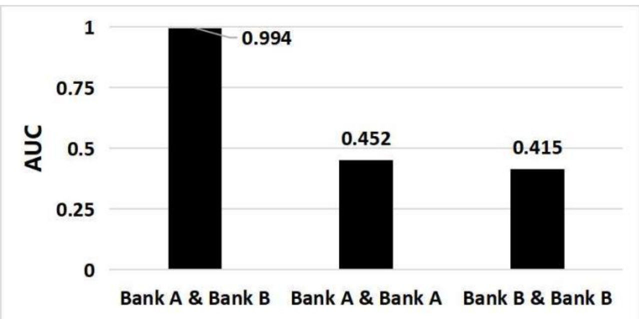

Fig. 5. Experimental results of antagonistic verification

(3) The LR model was used to train on the training set and test on the test set, with AUC as the judging standard. If the AUC was around 0.5, the model could not distinguish between the datasets of Bank A and Bank B. The larger the AUC value, the easier it was to distinguish between the datasets of Bank A and Bank B, and the larger the distribution difference. The experimental results are shown in Fig. 5.

The results show that the division values of distribution areas between Bank A and Bank B are close to 1, which shows that there is a significant difference in the data distribution. As a control, the distribution division values of each bank compared to itself are close to 0.5, indicating that the data basically have the same distribution.

5.4. Experiment IV: Chi-square test to delete high discrimination features

From the results of experiment III, we see a great difference in data distribution among institutions; so, we should consider using the Chi-square test to eliminate the features with high discrimination.

The Chi-square test is a widely used nonparametric hypothesis test method. It compares two or more sample rates and analyzes the correlation between two categorical variables. Assuming that the data of the two parties in the experiment are in different feature distribution spaces (0 and 1), if the P value of the Chi-square test output is less than 0.5, there is a strong correlation between features and classification labels. This means that the feature is highly differentiated.

The Chi-square test method was used to detect and delete features of high discrimination in the data of Bank A and Bank B, making the data on both sides more identically distributed, and the antagonistic test method was used for verification. The experimental results are shown in Fig. 6.

After the deletion of features of high discrimination, the distribution division value between Bank A and Bank B is close to 0.5, which indicates that the data of both parties are closer to being IID.

The data of Bank A and Bank B, which are IID after data processing, were next used for the federated modeling experiment. The training set comprised Bank A and B’s data from January to May 2020, and the test set comprised the June 2020 data. The recall rates are shown in Fig. 7, and the F1 scores are shown in Fig. 8.

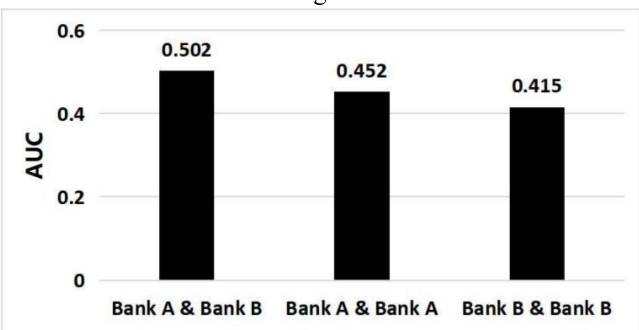

Fig. 6. Antagonistic test results after deleting strong correlation features

5.5. Experiment V: Verification of federated modeling performance

The experimental results show that the performance of the centralized modeling and federated learning modeling of the Bank A test set is only slightly different from that of the local independent modeling. The performance of centralized modeling and federated learning modeling of the Bank B test set is better than that of local independent modeling (recall rates increased by 14%, from 0.779 to 0.889, and F1 increased by 7.9%, from 0.865 to 0.933), and the performance of centralized modeling is close to that of federated learning modeling. The experimental results show that the application framework and method can significantly improve the performance of the financial credit risk management of small banks with fewer data samples.

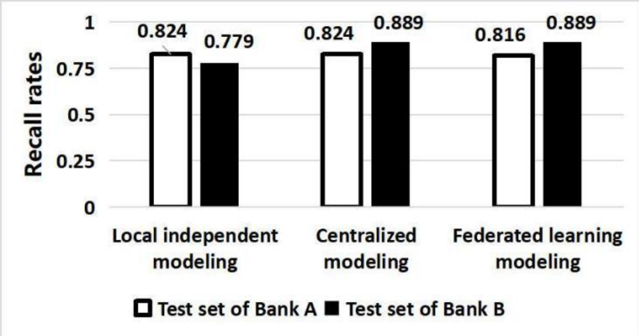

Fig. 7. Comparison of recall rates between Bank A and Bank B (Experiment V)

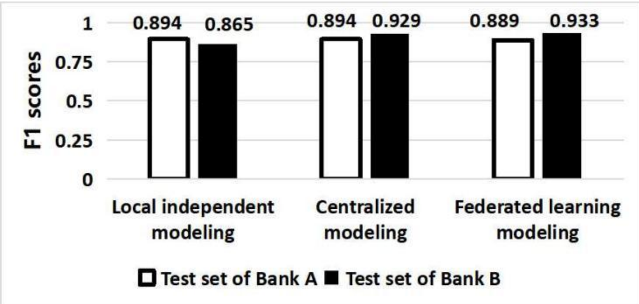

Fig. 8. Comparison of F1 scores between Bank A and Bank B

(Experiment V)

V.CONCLUSION

This study has found that the application of federated learning can effectively solve the data-sharing problem of financial institutions. Making full use of the data of different institutions for federated learning modeling is beneficial to small and medium banks and other institutions with insufficient data samples. To combat the data's non-IID features, we used the Chi-square test method to detect and delete the features with a high degree of discrimination in the data of each institution. This creates a more identically distributed dataset, allowing for ideal modeling performance. Applying this achievement to the field of financial credit can help financial institutions, especially smaller banks, solve data problems, share data, improve the identification of risky loans or credit card fraud [35], and solve the problem of data silos in the financial industry.

The methods in this study still have the issue of insufficient data features. In the future, we will continue to research the combined application of federated learning and modeling to address this problem. For all training samples, augmented feature vectors with additional features are constructed from the knowledge graph. For example, based on the original feature vector of the sample x, the graph features in the sample points are additionally extracted in addition to the existing features, such as how many of the sample points are overdue three times, as well as the frequency of nighttime transactions, forming an enhanced feature containing additional features. Next, these enhanced feature vectors are used to train a classifier. The classifier trained based on the enhanced feature not only predicts the classification result based on the original feature but also adds the feature vector to eliminate the correlation of the sample data, thus solving the non-IID problem more effectively.

Finally, we believe that blockchain is a promising future research direction to improve data sharing security when used in combination with federated learning [36].

ACKNOWLEDGEMENT

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

REFERENCES

[1] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.
[2] Jolfaei, A., Ostovari, P., Alazab, M., Gondal, I., & Kant, K. (2020). Guest editorial special issue on privacy and security in distributed edge computing and evolving IoT. IEEE Internet of Things Journal, 7(4), 2496-2500.
[3] Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. IEEE Access, 4, 1821-1834.
[4] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. Knowledge-Based Systems, 216, 106775.
[5] Hirofumi Miyajima, Noritaka Shigei, Hiromi Miyajima, and Norio Shiratori, Federated learning with divided data for BP, Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2021, 20-22 October, 2021, Hong Kong, pp94-99.
[6] Zhu, H., Xu, J., Liu, S., & Jin, Y. (2021). Federated learning on Non-IID data: a survey. Neurocomputing, 465, 371-390.
[7] Wang, L., Xu, S., Wang, X., & Zhu, Q. (2021, February). Addressing class imbalance in federated learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 11, pp. 10165-10173).
[8] Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., & Zhang, Y. (2021, January). Personalized cross-Silo federated learning on Non-IID data. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 9, pp. 7865-7873).
[9] Li S., Yang L., Li C., Zhang A., Luo R.(2021). Overview of federated learning: technology, applications and future. Computer Integrated Manufacturing Systems.
[10] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.
[11] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273-1282). PMLR.
[12] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.
[13] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021).Advances and open problems in federated learning.. Foundations and Trends in Machine Learning, 14(1–2), 1-210.
[14] Liu, Yang， Kang, Yan, Zhang, Xinwei, Li, Liping, Cheng, Yong, Chen, Tianjian, Hong, Mingyi, Yang, Qiang, (2019). A Communication efficient vertical federated learning framework. arXiv abs/1912.11187 (2019): n. pag.
[15] Sharma, S., Xing, C., Liu, Y., & Kang, Y.. Secure and efficient federated transfer learning. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 2569-2576). IEEE.
[16] Liu, B., Guo, Y., & Chen, X.. PFA: Privacy-preserving federated adaptation for effective model personalization. In Proceedings of the Web Conference 2021 (pp. 923-934).
[17] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.
[18] Liu, Q., Chen, C., Qin, J., Dou, Q., & Heng, P. A. (2021). Feddg:

Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1013-1023).
[19] Li, L., Long, S., Bi, J., Wang, G., Zhang, J., & Tao, X. A federated learning based semi-supervised credit prediction approach enhanced by multi-layer label mean. In Web Intelligence (No. Preprint, pp. 1-14). IOS Press.
[20] Li, L., Fan, Y., Tse, M., & Lin, K. Y. (2020). A review of applications in federated learning. Computers & Industrial Engineering, 149, 106854.
[21] Yang, W., Zhang, Y., Ye, K., Li, L., & Xu, C. Z. (2019, June). FFD: A federated learning based method for credit card fraud detection. In International Conference On Big Data (pp. 18-32). Springer, Cham.
[22] Fang, C., Guo, Y., Ma, J., Xie, H., & Wang, Y. (2022). A privacy-preserving and verifiable federated learning method based on blockchain. Computer Communications. 186, 1-11.
[23] Gao, L., Li, L., Chen, Y., Xu, C., & Xu, M. (2022). FGFL: A blockchain-based fair incentive governor for federated learning. Journal of Parallel and Distributed Computing. 163, 283-299.
[24] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., ... & He, B. (2021). A survey on federated learning systems: vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering.
[25] Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619-640.
[26] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip B Gibbons. The non-IID data quagmire of decentaralized machine learning. Proceedings of The 37th International Conference on Machine Learning (ICML), 2020
[27] Sattler, F., Wiedemann, S., Müller, K. R., & Samek, W. (2019). Robust and communication-efficient federated learning from non-i.i.d. data. IEEE Transactions on Neural Networks and Learning Systems, 31(9), 3400-3413.
[28] Van Lier, S., de Vries, A. P., & Herder, E. (2018). Robustness of federated averaging for non-iid data. Doctoral Dissertation, Master's Thesis, Radboud University.
[29] Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in Neural Information Processing Systems, 33, 7611-7623.
[30] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.
[31] Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., & Yu, H. (2019). Federated learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 13(3), 1-207.
[32] Kulkarni, V., Kulkarni, M., & Pant, A. (2020, July). Survey of personalization techniques for federated learning. In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4) (pp. 794-797). IEEE.
[33] AbdulRahman, S., Tout, H., Ould-Slimane, H., Mourad, A., Talhi, C., & Guizani, M. (2020). A survey on federated learning: the journey from centralized to distributed on-site learning and beyond. IEEE Internet of Things Journal, 8(7), 5476-5497..
[34] Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics, 19(1), 15-18.
[35] Ayoub Mniai, and Khalid Jebari, Credit card fraud detection by improved SVDD, Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2022, 6-8 July, 2022, London, U.K., pp32-37.
[36] Noor Afiza Mat Razali, Wan Nurhidayat Wan Muhamad, Khairul Khalil Ishak, Nurjannatul Jannah Aqilah M. Saad, Muslihah Wook, and Suzaimah Ramli, Secure blockchain-based data-sharing model and adoption among intelligence communities, IAENG International Journal of Computer Science, vol. 48, no.1, pp18-31, 2021.

Yan Li was born in Hunan, China, in 1979. In 2004, she received a M.S degree from Beihang University. She is currently a Ph.D. student at the School of Computer Science and Engineering at the South China University of Technology. Her research interests include financial technology, credit risk management, cybersecurity, and machine learning.

Guihua Wen received his Ph.D. degree in computer science and engineering from the South China University of Technology, and he is currently a professor and doctoral supervisor at the School of Computer Science and Technology of the South China University of Technology. He is also the professor-in-chief of the Data Mining and Machine Learning laboratory at the South China University of Technology and of the artificial intelligence and Traditional Chinese Medicine program at the Guangdong Research Center of Engineering and Technology. His research areas include cognitive affective computing, machine learning, data mining, and intelligent medical methods.