Hindawi
Complexity
Volume 2021, Article ID 6649605, 10 pages
https://doi.org/10.1155/2021/6649605

Research Article

A Machine Learning Approach to Evaluate the Performance of Rural Bank

Jun Wei,[1] **Tao Ye**,[2] **and Zhe Zhang**[3]

[1]*School of Economics and Management, Beijing Jiaotong University, 100044 Beijing, China*
[2]*School of Finance, Capital University of Economics and Business, 100070 Beijing, China*
[3]*School of Management Science and Engineering, Shandong University of Finance and Economics, 250014 Jinan, China*

Correspondence should be addressed to Jun Wei; 15113140@bjtu.edu.cn

Received 9 December 2020; Revised 26 December 2020; Accepted 28 December 2020; Published 13 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

In the current performance evaluation works of commercial banks, most of the researches only focus on the relationship between a single characteristic and performance and lack a comprehensive analysis of characteristics. On the other hand, they mainly focus on causal inference and lack systematic quantitative conclusions from the perspective of prediction. This paper is the first to comprehensively investigate the predictability of multidimensional features on commercial bank performance using boosting regression tree. The dimensionality in the financial-related fields is relatively high. There are not only observable price data, financial fundamentals data, etc., but also many unobservable undisclosed data and undisclosed events; more sources of income cannot be explained by existing models. Aiming at the characteristics of commercial bank data, this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation. In this method, a random subsample sampling is performed before training each regression tree. The adaptive reduction step size is used to replace the reduction step size setting of the original algorithm, which overcomes the shortcomings of low accuracy and poor generalization ability of the existing regression decision tree model. Compared to the BIRCH algorithm for classification of existing data, our proposed gradient boosting regression tree algorithm with adaptively reduced step size obtains better classification results. This paper empirically uses data from rural banks in 30 provinces in China to classify the different characteristics of rural banks' performance in order to better evaluate their performance.

1. Introduction

The traditional Malmquist index [1] examines the efficiency and productivity changes of financial institutions. For example, Paradi et al. [2] estimated that the Bank of Canada will develop a two-stage DEA to simultaneously benchmark performance in different dimensions and modify the SBM.

Machine learning technology has certain applications in performance evaluation in the financial field. Taking the fund performance analysis and evaluation model as an example [3, 4], the use of related technologies can improve the traditional model evaluation methods required in the risk model and the fund performance evaluation method, such as the adjustment and optimization of the characteristic risks of individual stocks, the summary of potential laws, and the forecast adjustment of the fund's short-term exposure. In the current performance evaluation work of commercial banks using machine learning, most of the researches only focus on the relationship between a single characteristic and performance and lack a comprehensive analysis of characteristics; on the other hand, they mainly focus on causal inference and lack systematic quantitative conclusions from the perspective of prediction.

Most of the existing bank performance evaluation models are based on the Malmquist index method, but the information dimensionality in the financial-related fields is relatively high [5–8]. There are not only observable price data, financial fundamentals data, etc., but also many unobservable undisclosed data and undisclosed events; more sources of income cannot be explained by existing models.

Based on boosting regression tree technology, this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation. Aiming at the characteristics of commercial bank data, this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation. In this method, a random subsample sampling is performed before training each regression tree. The adaptive reduction step size is used to replace the reduction step size setting of the original algorithm, which overcomes the shortcomings of low accuracy and poor generalization ability of the existing regression decision tree model. This paper empirically uses data from rural banks in 30 provinces in China to classify the different characteristics of rural banks' performance in order to better evaluate their performance. In this paper, we use predictive modeling and explanatory modeling in machine learning to evaluate the performance of rural banks and predict the possible development trend of performance. Explanatory models make assumptions about causality in advance and then use data to test them. Predictive models can unearth more complex laws in the data. However, the two are not completely opposed.

In summary, the contributions and innovations of this paper can be summarized as follows:

(1) This paper proposes the use of predictive models to evaluate the performance of commercial banks for the first time. In our opinion, compared with explanatory models, the predictive models can unearth more complex laws in the datasets.

(2) Aiming at the characteristics of commercial bank data, this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation.

(3) This study uses real commercial bank data from 30 provinces to conduct experiments. The experiment shows that the adaptively reduced step size gradient boosting regression tree algorithm proposed in this paper reveals the performance of commercial banks more objectively.

(4) This research not only uses predictive model methods to study bank performance evaluation from a more comprehensive perspective but also provides useful inspiration for commercial bank operations and management.

The rest of this paper is organized as follows: Section 2 is the related work part. Section 3 is the method proposed in this paper. Section 4 is the experimental results and analysis. We summarized a conclusion in Section 5.

2. Related Work

Ravi et al. presents a soft computing based bank performance prediction system [9]. It is an ensemble system whose constituent models are many kinds of neural network, SVM, CART, and a fuzzy rule-based classifier. Selecting a subset of favorable features is beneficial to improve the accuracy of bank performance prediction. For example, particle swarm optimization (PSO) is used to obtain suitable parameter settings for support vector machines (SVMs) and decision trees (DTs) [10], such as neural networks, support vector machines, and multicriteria decision aid that have also been used in the bank failure prediction, creditworthiness assessment, and underperformance [11].

A series of modeling techniques were employed to predict bank insolvencies on a sample of US-based financial institutions. The empirical results indicate that the method of random forests (RF) has a superior out-of-sample and out-of-time predictive performance, with neural networks also performing almost equally well as RF in out-of-time samples [12]. A sample of 3000 US banks (1438 failures and 1562 active banks) is investigated by two traditional statistical approaches (discriminant analysis and logistic regression) and three machine learning approaches (artificial neural network, support vector machines, and k-nearest neighbors) [13]. The empirical result reveals that the artificial neural network and k-nearest neighbor methods are the most accurate. An accurate risk assessment tool was proposed using unique KYC data and machine learning techniques to overcome problems in existing risk detection methods [14]. This work proposes that the bank branch is the best level at which to determine the degree of default risk and can also provide insight into patterns of suspicious transactions.

Several machine learning algorithms have been used on a real bank credit dataset for comparative analysis and to choose which algorithms are the best fit for learning bank credit data. These algorithms gave over 80% accuracy in prediction [15]. For evaluating bank efficiency and performance, a combined DEA with three machine learning approaches were used in 444 Ghanaian bank branches, decision-making units (DMUs). The results suggested that the decision tree (DT) and its C5.0 algorithm provided the best predictive model [16]. The potential usage of bagging also has been investigated which is one of the most popular ensemble learning methods, in building ensemble models, and is used to predict the determinants of Turkish IaDB profitability [17]. This empirical study indicates that bagging ensemble models are superior to their base learners and could improve the prediction accuracy of individual ML models.

There are a large number of empirical studies to analyse and evaluate machine learning techniques in the bank risk management [18]. The areas or problems in risk management also have been inadequately explored for further research. These prior empirical studies have shown that the application of machine learning in the management of banking risks such as credit risk, market risk, operational risk, and liquidity risk has been explored. For example, the combination of financial indicators, readability, sentiment categories, and bag-of-words was used to increase prediction accuracy. It shows that the quality of the prediction significantly increased when using the correlation-based feature selection of bag-of-words [19]. The supervised artificial neural network algorithm is implemented for classification purpose in customer retention and fraud detection [20].

We can clearly conclude that machine learning algorithms have been widely used in various areas of banking, including performance assessment, credit evaluation, risk management, customer retention, and fraud detection. However, when we carefully review the above work, it is easy to see that the machine learning algorithms used in the above work are mostly explanatory models, which are used to verify the causal relationships between observable variables in the theory. Unlike the previous work mentioned above, our work in this paper is based on predictive analysis, which has appeared less frequently in empirical studies of finance and banking. The method proposed in this paper does not assume a causal relationship between variables, and most of the models that fit well do not assume a specific functional form between variables (e.g., linear relationship, U-shaped relationships, and exponential relationships), and thus predictive models are able to uncover more complex patterns in the datasets.

3. Methods

3.1. Variable Selection. This paper studies the performance of China's provincial rural banks, that is, provincial rural banks represent the regional heterogeneity of rural banks. Fukuyama and Weber [5] used a two-stage network model including good and bad output to evaluate the performance of Japanese banks. They use labor, physical capital, and financial equity capital to produce loans and securities investments and use deposits as intermediate output.

In order to evaluate the performance of rural banks in different provinces, this paper selects 30 provincial rural banks across the country except Tibet as the research object and uses 4 years of data to evaluate the productivity growth and decomposition efficiency indicators of provincial rural banks in China. According to the concept in the literature [6–8], the input variables are capital and employees based on cost and the ideal output variable is profit based on revenue. In addition, this paper studies the dynamic development and risk control of rural banks in China and incorporates carry-over activities and negative externalities into the study, as shown in Table 1.

Banks use capital and human resources to make profits. Bank deposits are considered as a special resource because banks strive to attract deposits and use them as a positive indicator of performance evaluation. At the same time, they use these deposits to earn future profits. In DEA banking literature, deposit is a controversial topic. Compared with other input-output variables, deposits have the characteristics of dynamic variables. Therefore, rural bank deposits are defined as a carry-over variable. From a more comprehensive analysis, nonperforming loans (NPLs) represent bad debt risks, and there is an inevitable symbiotic relationship between bad debt risks and profits. Therefore, the nonperforming loans of rural banks are defined as the nonperforming output of rural banks.

3.2. Model Construction and Algorithm

3.2.1. Malmqusit Index Calculation. The provincial rural bank is defined as the decision-making unit of the performance evaluation of the rural bank, and it is the research object of the performance evaluation of the rural bank. In period t, provincial rural banks (DMUs) use input X and carry-over activity z to produce ideal output Y_d and bad output Y_u. Carry forward activity connection time periods $t-1$, t, and $t+1$. The variables of input, output, and carry-over activities have regional heterogeneity.

In the traditional dynamic DEA model, (X_t, Y_t) and (X_{t+1}, Y_{t+1}) are separately dealt with for obtaining catch-up effect and frontier-shift effect. However, Tone and Tsutsui (2010) introduced the carry-over into the dynamic model, called dynamic SBM (DSBM). This paper basically follows the dynamic SBM thinking. To estimate the frontier functions, upon which we compute the nonoriented measures of the efficiency, we deal with n DMUs ($j = 1, \ldots, n$) over period t ($t = 1, \ldots, T$). Using period t as a benchmark, DMUs produce s outputs ($i = 1, \ldots, s$) using m inputs ($i = 1, \ldots, m$). Moreover, we define r links ($i = 1, \ldots, r$) as carry-over activities between two consecutive periods. Then, we can obtain the pure technical efficiency for DUM j in period t as follows:

$$\rho_v^* = \min \frac{1 - (1/(m+r))\left(\sum_{i=1}^{m}\left(S_{it}^-/x_{iot}\right) + \sum_{i=1}^{m}\left(S_{ilt-1}^-/z_{iot-1}\right)\right)}{1 + (1/(s+r))\left(\sum_{i=1}^{s}\left(S_{it}^+/y_{iot}\right) + \sum_{i=1}^{m}\left(S_{ilt}^+/z_{iot}\right)\right)}, \tag{1}$$

where x_{ijt} and y_{ijt} are the inputs and outputs of DMU_j at period t, respectively, and we define z_{ijt} as links. S_{it}^-, S_{it}^+, S_{ilt-1}^-, and S_{ilt}^+ are slack variables denoting, respectively, input excess, output shortfall, link excess, and link shortfall.

Solving the above program for each DMU, we can obtain ρ_c^*, which means the variable returns to scale case. For the constant returns to scale case ρ_v^*, we only need delete the restriction $\sum_{j=1}^{n}\lambda_j = 1$ in the above model. Then, we can decompose the technical efficiency (TE) into scale efficiency (SE) and pure technical efficiency (PuTE) as

$$\begin{aligned} \overline{\mathrm{TE}} &= \rho_c^*, \\ \overline{\mathrm{PuTE}} &= \rho_v^*, \\ \overline{\mathrm{SE}} &= \rho_c^*/\rho_v^*. \end{aligned} \tag{2}$$

Finally, using the above formula, we can decompose the sources of catching-up effect as

$$\mathrm{CU} = \mathrm{TEC} = \mathrm{PUTC}^*\mathrm{SEC}. \tag{3}$$

According to the above, we can decompose the sources of frontier-shift effect as

Table 1: Data description.

Index	Capital stock	Staffs	Deposit	Profit	NPLR
Mean	71.25	21976	1210.84	1449.96	12.43
SD	61.71	15232	1163.65	1361.5	18.3
Min	0.58	2087	28.26	33.01	−29
Max	234.15	60896	5940.71	6957.89	104.29

Note. NPLR: nonperforming loan ratio.

$$\mathrm{FS} = \mathrm{DPC}^{*}\mathrm{TPC}. \tag{4}$$

In conclusion, we decomposed the dynamic Malmqusit model as

$$\overline{M}(x, y, z) = \overline{\mathrm{TEC}} \cdot \mathrm{DTPC} = \overline{\mathrm{SEC}} \cdot \overline{\mathrm{PuTC}} \cdot \mathrm{DTPC}. \tag{5}$$

In the clustering part, we use hierarchical clustering, gradient boosting regression tree algorithm, and other related algorithms to further cluster the above index results. The hierarchical clustering uses the BIRCH algorithm. This algorithm is mainly used when the amount of data is large and the data type is numerical. We use the adaptively reduced step size gradient boosting regression tree algorithm proposed in this paper to optimize, so as to make the clustering effect better.

3.2.2. Adaptively Reduced Step Size Gradient Boosting Regression Tree. The gradient boosting regression tree algorithm is widely used in clustering research in the financial field. The existing gradient boosting regression tree method has certain shortcomings. Firstly, the existing methods rely too much on data quality, which makes us often unable to achieve the desired prediction accuracy in actual modeling. Secondly, the existing methods require careful adjustment of parameters, and the training time may be relatively long. Finally, the improvement effect of existing methods is relatively limited.

Next, we will introduce the adaptively reduced step size gradient boosting regression tree algorithm. In the gradient boosting regression tree algorithm, the reduction step size is fixed, and it is determined as a parameter when starting to train the model. We now analyze the loss function of the model. Let $H_j(x)$ be the integrated learner of the first j residual trees, let $h_{j+1}(x)$ be the $j+1$ weak learner, and the learning step is λ. The probability of each training sample being selected as a random subsample is $1/n$, so the loss function can be defined as

$$L\left(y, H_j(x) + \lambda h_{j+1}(x)\right) = \sum_{i=1}^{n} \frac{1}{n}\left(y_i - \left(H_j(x_i) + \lambda h_{j+1}(x_i)\right)\right)^2. \tag{6}$$

Given $H_j(x)$ and $h_{j+1}(x)$, in order to find the corresponding reduction step λ when the loss is the smallest, let the loss function take the derivative of λ and make the derivative equal to 0, we can get

$$\frac{\partial L}{\partial \lambda} = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i\left(H_j(x_i) + \lambda h_{j+1}(x_i)\right)\right)h_{j+1}(x_i) = 0. \tag{7}$$

Then, we have

$$\lambda = \frac{\sum_{i=1}^{n}\left(y_i - H_j(x_i)\right)h_{j+1}(x_i)}{\sum_{i=1}^{n} h_{j+1}^2(x_i)}. \tag{8}$$

Therefore, the reduction step size can be automatically updated with the current learning result to adapt to the minimization of the function.

Then, we can write the improved gradient boosting regression tree Algorithm 1 steps as follows.

4. Results

4.1. Experimental Methods and Processes. The experimental data in this paper are the four-year data of 30 provincial rural banks except Tibet, including deposits, capital stock, employees, profits, and nonperforming loan rates. The five efficiency indexes decomposed by the Malmquist index method are SuEC, PuTC, SEC, DPC, and TPC. Taking Yunnan Province as an example, these five indicators are shown in Table 2:

The experimental process of this paper is shown in Figure 1.

The classification part is to divide the rural banks in 30 provinces into several groups, so that the above groups can be divided into different performance categories based on the characteristics of the efficiency of rural banks.

In clustering, we use the BIRCH algorithm and the algorithm proposed in this paper, respectively. Use the original classification results of 30 provinces as a reference to check the accuracy of clustering by these two algorithms.

4.2. Clustering of Rural Bank Performance

4.2.1. BIRCH Clustering. As shown in Figure 2, when using the BIRCH algorithm to classify existing data, we get a total of six groups of results. Since the cluster feature tree has a limit on the number of cluster features of each node, the clustering result may be different from the real category distribution. In addition, the algorithm has a poor clustering effect on high-dimensional feature data.

4.2.2. Gradient Boosting Regression Tree Clustering. As shown in Figure 3, when we use the gradient boosting regression tree algorithm, we get seven groups of provincial banks. The accuracy of the algorithm is higher, the generalization ability is stronger, and the classification result is basically consistent with the original reference.

4.2.3. Performance Type. According to the cluster analysis result and the character of decomposed efficiency in Chinese rural banks, we merge special groups for analysis, such as Group 4 and Group 6 as TPEI (traditional pure economic improved type) and Group 2, Group 5, and Group 7 as SuECI (sustainable efficiency change improved type). This grouping sounds more realistic and good to empirical analysis, so we distinguish Chinese rural banks into four type of performance as shown in Table 3.

Input:
Training samples $T = \{(x_i, y_i) = (x_{i1}, ..., x_{ip}, y_i) | i = 1, ..., n\}$
Residual tree training times is M, random sampling rate is rate, complexity parameter is cp.
Training steps:
Initialize training samples $T_1 = T$, *where* $y_j = (y_1, y_2, ..., y_n)$, *reduce step size* $\lambda_1 = 0.01$,
FOR $j = 1, 2, \ldots, M$
(1) *From* T_j *without replacement, repeat the subsample with a random ratio of rate as the training sample of the current regression tree.*
(2) *Based on the complexity parameter cp, train the* $j-$th *residual tree model* $h_j(x)$ *on the current training sample.*
(3) Update reduction step $\lambda_j = \sum_{i=1}^{n} (y_i - H_{j-1}(x_i))/h_j(x_i)$.
(4) *Give the predicted value* $\widehat{y}^j = (\widehat{y}_1^j, \widehat{y}_2^j, \ldots, \widehat{y}_n^j)$ *of the training sample* T_j *on* $h_j(x)$.
(5) *Update the output variable value* $y^{j+1} = y^j - \lambda_j \widehat{y}^j$ *on the training sample* T_j.
END FOR
Output: improved gradient boosting regression tree model $H(x) = h_M(x) + \sum_{j=1}^{M-1} \lambda_j h_j(x)$.

Algorithm 1: Gradient boosting regression tree with an adaptively reduced step size.

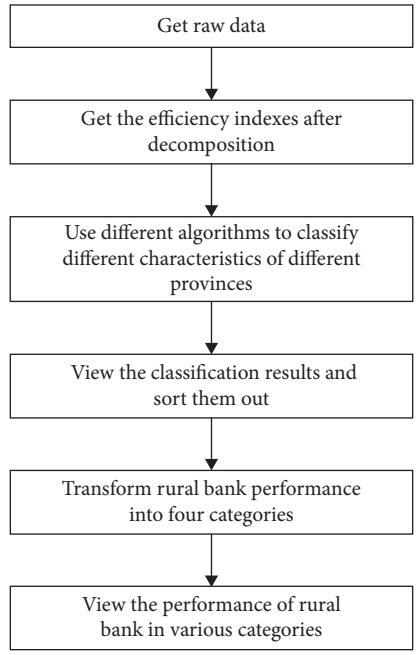

Figure 1: The experimental process outline.

Types (I) and (II) rural banks perform lower than type (III). While from the sustainable development viewpoint, types (I) and (II) belong to potential banks and type (III) exists implicit crisis. We refer to type (III) as cash cows in Boston matrix. Rural banks of type (IV) are diverse. However, the unified advantage of sustainable efficiency makes them stand out as part of a sustainable development strategy. Hereafter, we analyze the four types.

4.2.4. Rural Bank Performance in Inland Areas. Most DPCL banks are located in inland areas in China. The performances of rural banks seriously lag behind other three type banks. The main characteristic is that DPC is the only bottleneck that constrains their performance. From purely a profit viewpoint, PuTC is on the efficient frontier and TPC improves productivity growth. This suggests allocation of inputs and desirable outputs are effective and their quality growths are positive. However, the undesirable outputs and links are ineffective. That is to say, these banks aim at pursuing short-term profit and ignore long-term sustainable profit.

As shown in Figure 4, the Gansu rural bank keeps pure profit indexes effective. Meanwhile, its highly sustainable efficiency changes keep its performance rank the top seven in China. This suggests that, at the primary period of sustainable development, incorporating sustainable method into performance estimation makes greater progress. Above all, though DPCL performance lags behind others, it has the only bottleneck of carry-over activity (deposit). This presents challenges as well as opportunities.

4.2.5. Rural Bank Performance in Coastal Areas. The SuTECL banks are located in the coastal panhandle of the east area, which includes seven provinces. Besides the coastal

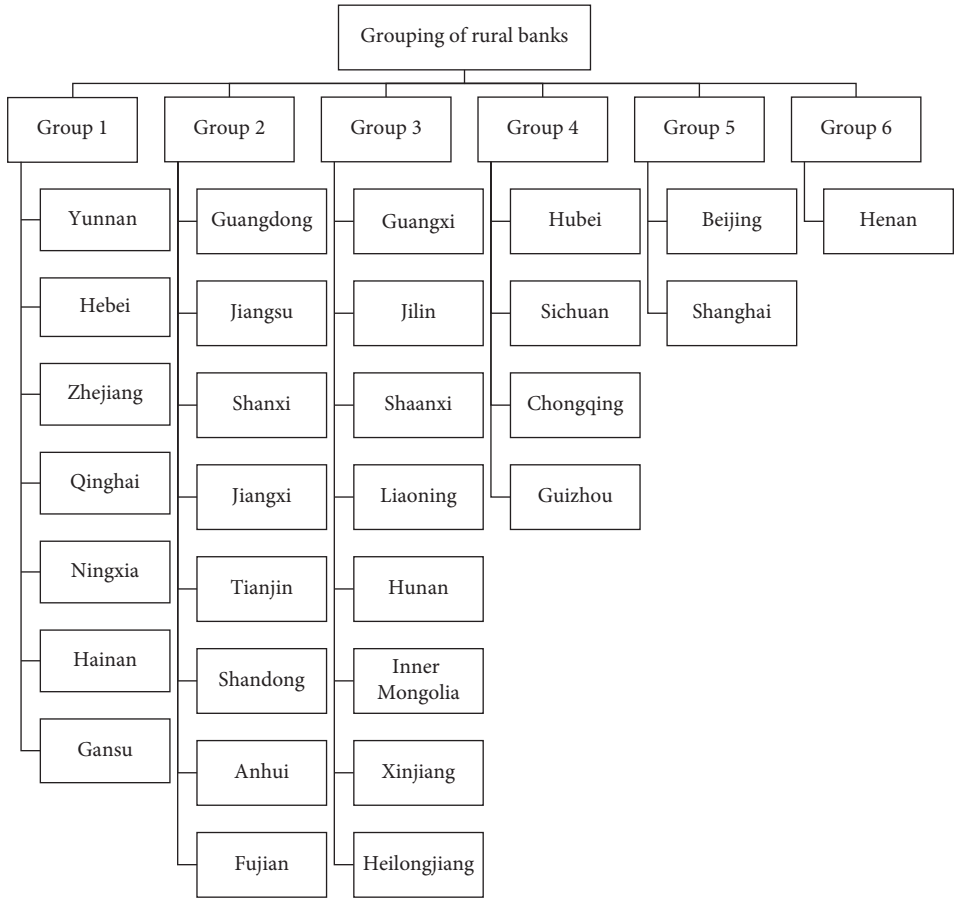

Figure 2: The clustering result of Birch algorithm.

panhandle of the east area, Shanxi rural bank also belongs to the SuTECL. The performances of these rural banks are in the bottom half in Chinese rural banks. The main characteristics are that TPC is the only benefit, and lower PuTC and medium below SuEC in SuTECL type banks limit the productivity growth of these banks. This suggests the local developed economy drives the improvement of performance. However, allocation of inputs and desirable outputs loses the customary advantage in the eastern area. That is to say, this is a big challenge because these banks ignore the basic control of factor efficiency.

As shown in Figure 5, Anhui rural bank is the only one whose DPC is effective. Although it is ineffective from a pure profit perspective, this bank focuses on a sustainable development strategy. So, the control of link and undesirable output improves its performance and puts it in the top two of this type. The Anhui rural bank is lower than that of Fujian. However, with the viewpoint of the sustainable efficiency change of Anhui rural bank, its performance will exceed Fujian's in the near future.

4.2.6. Rural Bank Performance in Central Areas. The banks of TPEI are located in the panhandle of northern and central regions of China as *T* sharp, including five provinces in northern China. Besides that, Hubei and Hunan rural banks also belong to TPEI as shown in Figure 6. The performances of these rural banks are higher and show smooth fluctuation. The main characteristics are that DPC is lower, and the performances of PuTC and TPC improve together. This suggests it has advantages from a purely technical viewpoint. The performances of these banks are in the leading position among Chinese rural banks. However, it is a big challenge to this type since lower DPCs in these banks mean ignorance of the deposit scroll effect in the long term. It will be difficult for this type of bank to keep its predominance if it continues to pursue short-term profit. This will also have a series of drawbacks.

As shown in Figure 6, the number of rural banks in TPEI is one-third of 30 banks in China. So for Chinese rural banks, it is still a long way to control carry-over activity (deposit) and undesirable output (NPLR) and the situation is severe. Rural banks in Xinjiang, Liaoning and Jilin are effective from a sustainable development strategy viewpoint. The effective situation means these banks have already focused on developing a sustainable dynamic strategy, especially deposit and NPLR control. These type banks are referred to as cash cows in Boston matrix. So, using the profit advantage, if it gradually transfers the focus into sustainable development strategy, it will be in the leading position of China.

4.2.7. Rural Bank Performance in Municipality Areas. The SuECI banks have the advantage of being new, and these type banks include Henan bank and the three municipality

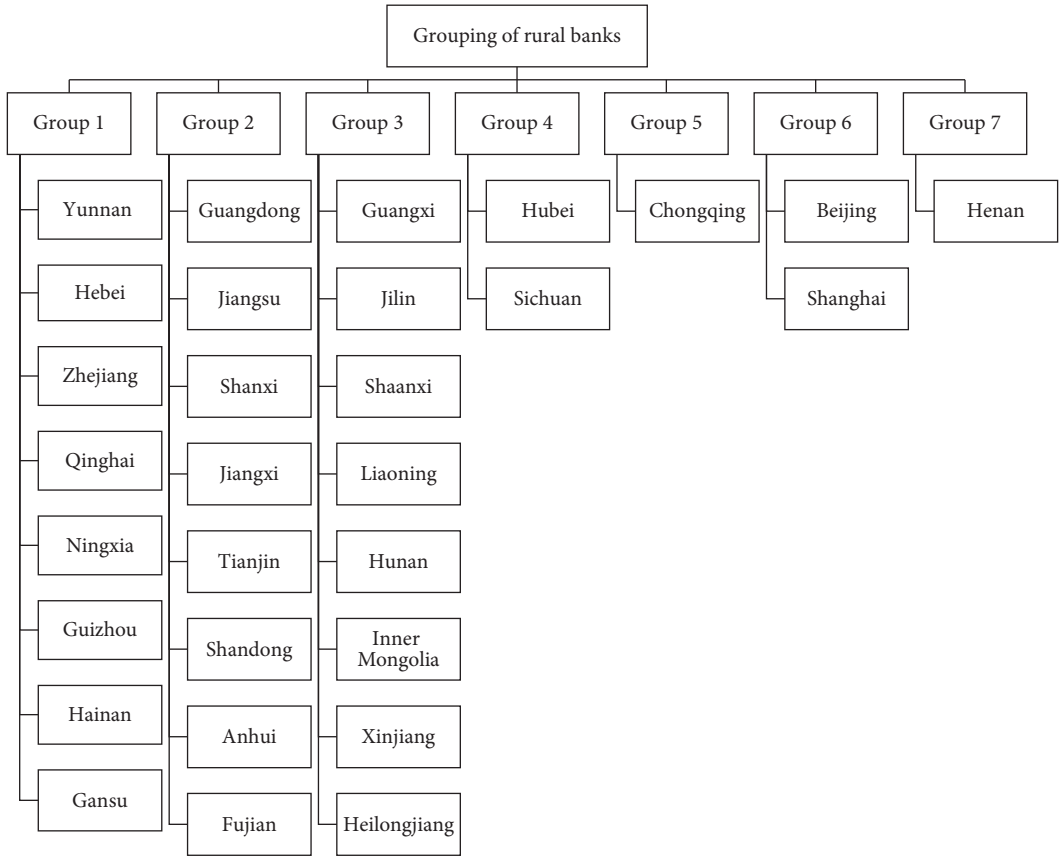

Figure 3: Gradient lift regression tree clustering.

Table 2: Five indicators of Yunnan province.

Inland	DSTFP	SuEC	PuTC	SEC	DPC	TPC
Yunnan	0.99	1.00	1	1	0.89	1.12

Table 3: Four type of performance of Chinese rural banks.

No.	Abbreviation	Norm
(I)	DPCL	Dynamic progress limited type
(II)	SuTECL	Sustainable technical efficiency changes limited type
(III)	TPEI	Traditional pure economic improved type
(IV)	SuECI	Sustainable efficiency changes improved type

banks in Chongqing, Beijing, and Shanghai. The performances of these rural banks differ significantly. The main characteristics are that SuEC and TPC are higher. The characteristics mean that the performances of SuECI banks have benefited from local economic advantages and sustainable development strategy. This is an opportunity for great performance improvement because of the sustainable advantage.

As shown in Figure 7, Chongqing is the youngest municipality in China. The sustainable dynamic performance (DSTFP) is the lowest among Chinese rural banks. The main reason is its low profit efficiency. The scroll of deposit and the control on NPLR have advantage in rural banks of China. This presents challenges as well as opportunities. The Beijing and Shanghai rural banks are in developed areas in China. Strong local economies there improve the performance of rural bank. However, the control of deposit is a drawback, especially in Beijing. The drawback means sustainable development requires a qualitative leap after quantity accumulates. Otherwise, it is difficult to continue performance improvements. The performance in Henan rural bank is the best from the viewpoints of both sustainability and allocation. This proves that Henan rural bank seizes the opportunity even if it does not have a strong economic backdrop. That is to say, at the primary period of sustainable development, incorporating sustainable methods into performance management can improve the productivity growth greatly.

4.3. Contrastive Analysis of Rural Bank Performance. After analysing the four types of rural banks in my country, the results of the model before and after using machine learning technology are compared. This can more clearly show our contribution to empirical analysis.

Based on the above model, we compared the total factor productivity of China's rural banking industry. On the whole, the use of machine learning technology has a more obvious positive effect on bank performance evaluation, especially for high-efficiency banks. It refers to provinces that are purely economically efficient, ignores sustainable

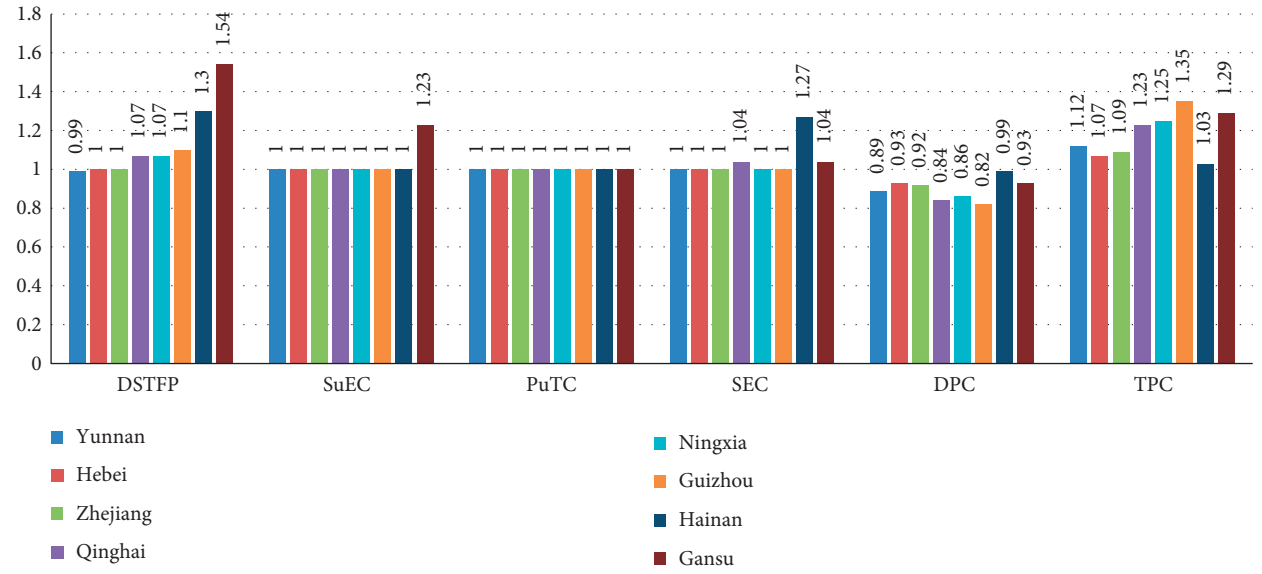

Figure 4: Decomposed efficiency indexes of DPCL.

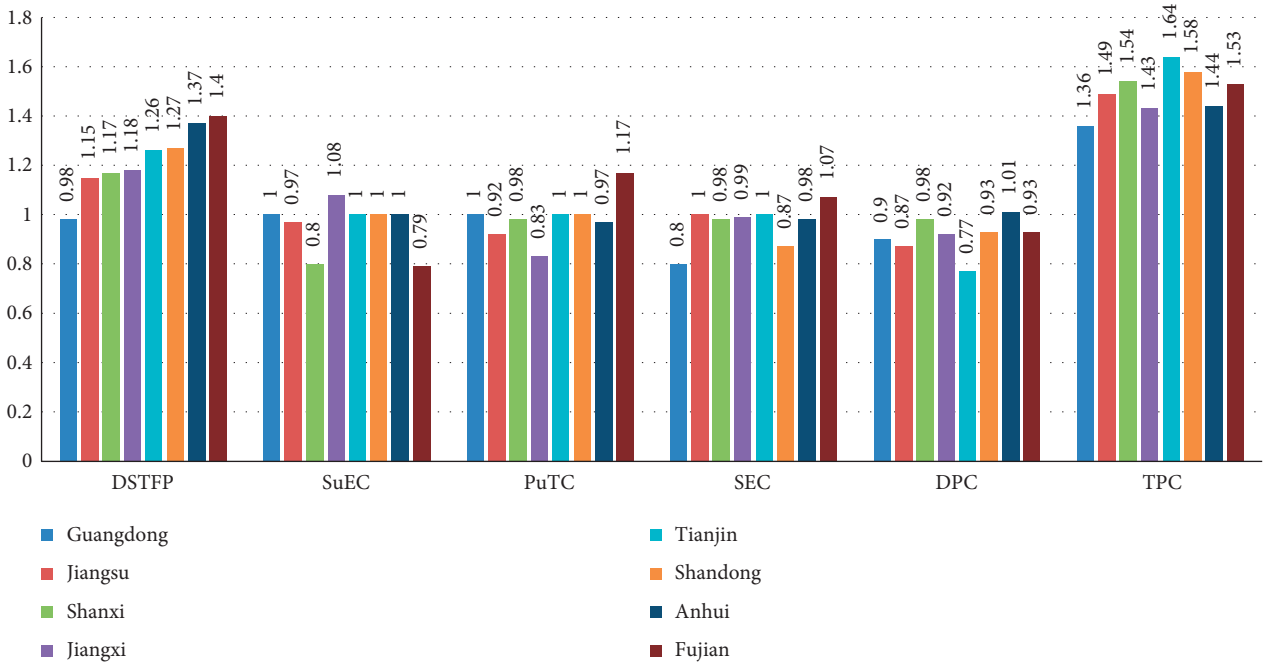

Figure 5: Decomposed efficiency indexes of SuTECL.

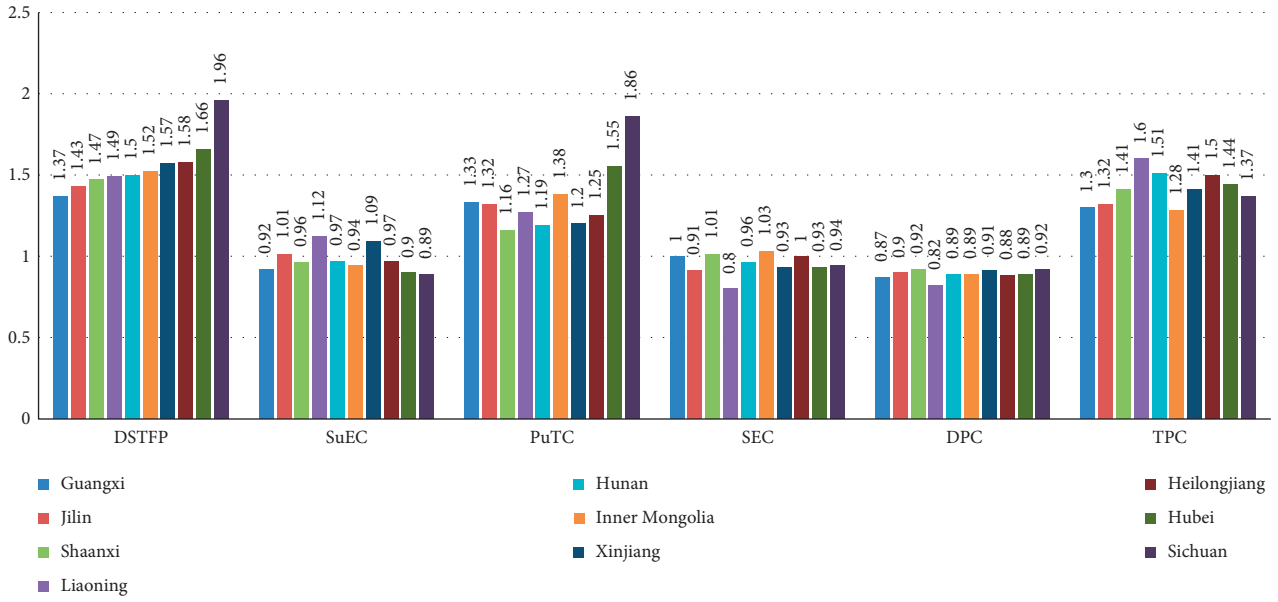

Figure 6: Decomposed efficiency indexes of TPEI.

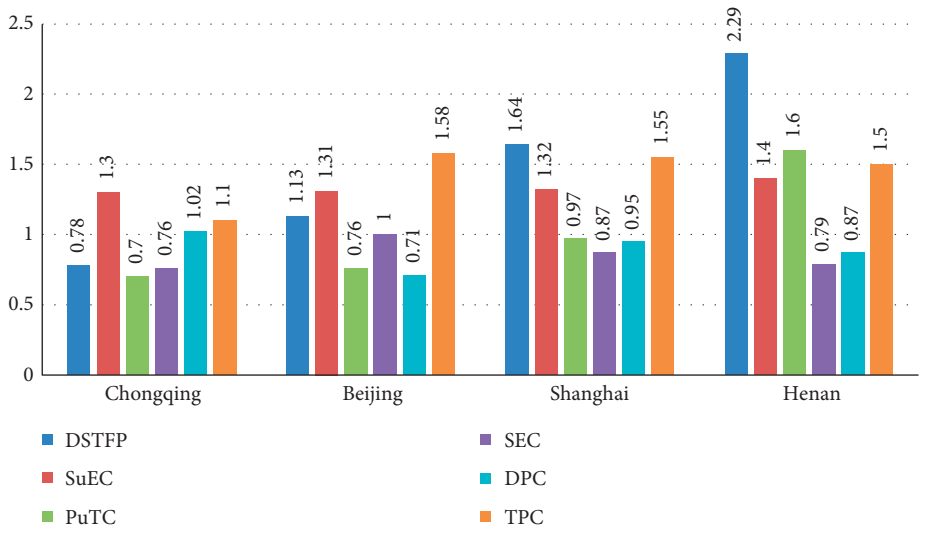

Figure 7: Decomposed efficiency indexes of SuECI.

development, and emphasizes short-term development. Among these banks, the rolling effect of efficiency and loan interest rates restricts the sustainable development of rural banks. As an inefficient bank in a purely economic sense, sustainable dynamic efficiency has a positive impact on its performance. For example, Xinjiang Rural Bank has a good performance and sustainable dynamic efficiency has played a positive role. The development model of the region is in good condition and needs attention.

In summary, it is still a process to incorporate sustainable development strategies into the operation and management of rural banks in my country. It can be seen from the above model that the productivity growth of rural banks is affected by catching up with the effective frontier and shifting from the effective frontier. We have made a comparative analysis of its performance from the perspective of pure economy and sustainable development.

5. Conclusions

In the current performance evaluation works of commercial banks, most of the researches only focus on the relationship between a single characteristic and performance and lack a comprehensive analysis of characteristics. On the other hand, they mainly focus on causal inference and lack systematic quantitative conclusions from the perspective of prediction. This paper is the first to comprehensively investigate the predictability of multidimensional features on commercial bank performance using boosting regression tree. Aiming at the characteristics of commercial bank data,

this paper proposes an adaptively reduced step size gradient boosting regression tree algorithm for bank performance evaluation. Compared to the BIRCH algorithm for classification of existing data, our proposed gradient boosting regression tree algorithm with adaptively reduced step size obtains better classification results. This paper empirically uses data from rural banks in 30 provinces in China to classify the different characteristics of rural banks' performance in order to better evaluate their performance.

Based on the hierarchical cluster analysis, the banks in China are divided into four groups: DPCL, SuTECL, TPEI, and SuECI. This paper also summarizes some interesting findings about the productivity growth of various types of rural banks in China, such as SuECI is worthy of attention; TPEI is potentially dangerous. The reason is that although this type of bank has good profit performance, it performs poorly in the evaluation of NPLR.

The follow-up research includes four aspects. First, we will apply external weights to all inputs, links, and outputs [21, 22]. Second, we will incorporate dynamic cost revenue and profit efficiency into our model [23]. Third, we will conduct sensitivity analysis and factor analysis of DSMPI [24]. Fourth, we will apply resampling methods, such as bootstrap techniques, to estimate the performance.

Data Availability

All data used in this study can be made available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] W. Bolt and D. Humphrey, "Bank competition efficiency in Europe: a frontier approach," *Journal of Banking & Finance*, vol. 34, no. 8, pp. 1808–1817, 2010.

[2] J. C. Paradi, S. Rouatt, and H. Zhu, "Two-stage evaluation of bank branch efficiency using data envelopment analysis," *Omega*, vol. 39, no. 1, pp. 99–109, 2011.

[3] D. U. A. Galagedera, I. Roshdi, H. Fukuyama, and J. Zhu, "A new network DEA model for mutual fund performance appraisal: an application to U.S. equity mutual funds," *Omega*, vol. 77, pp. 168–179, 2018.

[4] S. Cheng, R. Lu, and X. Zhang, "What should investors care about? Mutual fund ratings by analysts vs. Machine learning technique," *Machine Learning Technique*, 2020.

[5] H. Fukuyama and W. L. Weber, "A slacks-based inefficiency measure for a two-stage system with bad outputs," *Omega*, vol. 38, no. 5, pp. 398–409, 2010.

[6] T. Kaoru, M. Tsutsui, and D. E. A. Dynamic, "A slacks-based measure approach," *Omega*, vol. 38, pp. 145–156, 2010.

[7] T. Kaoru and M. Tsutsui, "Network DEA: a slacks-based measure approach," *European Journal of Operational Research*, vol. 197, pp. 243–252, 2009.

[8] T. Kaoru and M. Tsutsui, "Dynamic DEA with network structure: a slacks-based measure approach," *Omega*, vol. 42, no. 1, pp. 124–131, 2014.

[9] V. Ravi, H. Kurniawan, P. N. K. Thai, and P. R. Kumar, "Soft computing system for bank performance prediction," *Applied Soft Computing*, vol. 8, no. 1, pp. 305–315, 2008.

[10] S.-W. Kumar, Y.-R. Shiue, S.-C. Chen, and H.-M. Cheng, "Applying enhanced data mining approaches in predicting bank performance: a case of Taiwanese commercial banks," *Expert Systems with Applications*, vol. 36, no. 9, pp. 11543–11551, 2009.

[11] M. D. Cheng and F. Pasiouras, "Assessing bank efficiency and performance with operational research and artificial intelligence techniques: a survey," *European Journal of Operational Research*, vol. 204, no. 2, pp. 189–198, 2010.

[12] A. Petropoulos, V. Siakoulis, E. Stavroulakis, and N. E. Vlachogiannakis, "Predicting bank insolvencies using machine learning techniques," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1092–1113, 2020.

[13] H. H. Le and J.-L. Viviani, "Predicting bank failure: an improvement by implementing a machine-learning approach to classical financial ratios," *Research in International Business and Finance*, vol. 44, pp. 16–25, 2018.

[14] T.-H. Chen, "Do you know your customer? Bank risk assessment based on machine learning," *Applied Soft Computing*, vol. 86, p. 105779, 2020.

[15] R. E. Turkson, E. Y. Baagyere, and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness," in *Proceedings of the 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*, pp. 1–7, Shenzhen, China, September 2016.

[16] P. Appiahene, Y. M. Missah, and U. Najim, "Predicting bank operational efficiency using machine learning algorithm: comparative study of decision tree, random forest, and neural networks," *Advances in Fuzzy Systems*, vol. 2020, Article ID 8581202, 12 pages, 2020.

[17] H. Erdal and İ. Karahanoğlu, "Bagging ensemble models for bank profitability: an emprical research on Turkish development and investment banks," *Applied Soft Computing*, vol. 49, pp. 861–867, 2016.

[18] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: a literature review," *Risks*, vol. 7, no. 1, p. 29, 2019.

[19] P. Hájek, "Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns," *Neural Computing and Applications*, vol. 29, no. 7, pp. 343–358, 2018.

[20] P. S. Patil and N. V. Dharwadkar, "Analysis of banking data using machine learning," in *Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 876–881, Tirupur, India, February 2017.

[21] A. J. Ikechukwu, "Assessment of organizational performance of private manufacturing companies: the impact of supply chain management responsiveness," *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 26–44, 2019.

[22] W. Ghodbane, "Corporate social responsibility and performance outcomes of high technology firms: impacts on open innovation," *Journal of System and Management Sciences*, vol. 9, no. 4, pp. 29–38, 2019.

[23] I. Doina, "POPESCU, sebastian-ion CEPTUREANU, Adriana ALEXANDRU, eduard-gabriel CEPTUREANU, relationships between knowledge absorptive capacity, innovation performance and information technology. Case study: the Romanian creative industries SMEs," *Studies in Informatics and Control*, vol. 28, no. 4, pp. 463–476, 2019, ISSN 1220-1766.

[24] B. Lalic, M. Delic, N. Simeunovic, N. Tasic, and S. Cvetkovic, "The impact of quality management purchasing practices on purchasing performance in transitional economies," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 3, pp. 815–822, 2019.