Evaluation method of agricultural production technical efficiency based on Borderline-SMOTE and LightGBM

Jianying Feng, Yan Shi, Yunhui Su, Weisong Mu and Dong Tian*

College of Information and Electrical Engineering, China Agricultural University, Beijing,100083, China

*Correspondingauthor:td_tiandong@cau.edu.cn

Abstract. Data envelopment analysis (DEA) model is widely used to calculate the technical efficiency of agricultural production, but it is facing the defects of poor flexibility and slower speed. For this reason, we propose an evaluation method of agricultural production technical efficiency that integrates the DEA model, Borderline-SMOTE oversampling algorithm, and Light Gradient Boosting Machine (LightGBM) regression algorithm, and verify the effect of the method on the grape farmer dataset. The experimental results show that the MAE, MSE, and R^2 of the fusion model are 4.05E-02, 5.25E-03, and 0.898 respectively on the test set when the imbalanced ratio of the dataset is 4, which is better than other comparison models under the same imbalanced ratio and other fusion models under different imbalanced ratio. It indicates that the regression model of agricultural production technical efficiency based on the Borderline-SMOTE and LightGBM algorithm has superior prediction effect and can effectively make up for the limitations of the DEA model.

Keywords: technical efficiency; oversampling; regression prediction; efficiency prediction; algorithm fusion

1 Introduction

Technical efficiency reflects the relationship between input and output under certain production conditions [1]. DEA is one of the most commonly used methods for measuring the technical efficiency of agricultural production. It does not need to estimate or assume the function form of the evaluation object and avoids the interference of subjective factors in the evaluation process [2]. However, a large amount of linear programming must be done in the DEA model operation. Especially when it is necessary to deal with the technical efficiency evaluation task of new samples, the technical efficiency of all samples in the dataset must be recalculated [3], resulting in the problems of slow operation efficiency and low model flexibility in the practical application of DEA model [4,5].

Bayesian ridge regression [6], support vector regression (SVR) [7], elastic network [8], LightGBM [9], and other models in machine learning can effectively find the internal nonlinear relationship of data. At present, to make up for the lower flexibility of

A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_60

the DEA model, some researchers have combined the machine learning model and DEA model to calculate the technical efficiency [3,5,10,11]. The fusion model based on this idea has been applied to supplier selection [11], innovation efficiency of regional rural commercial banks [3], production performance evaluation of grape farmers [5], etc.

Nevertheless, the existing researches did not consider the class distribution of samples: in reality, production units located at the frontier (i.e., production units with DEA technical efficiency of 1) often account for only a small proportion, and vast majority of production units are not located at the frontier. There is an obvious class imbalance problem, which will reduce the accuracy of the evaluation method based on the machine learning model.

Therefore, this paper proposes an agricultural production technology efficiency evaluation method combining DEA, Borderline-SMOTE oversampling technology, and machine learning regression model so as to achieve rapid and accurate agricultural production technology efficiency evaluation.

2 Materials and methods

2.1 Dataset

The proposed method is applied to the evaluation of grape farmers' production technology efficiency. Usually, the cultivation and production process of grapes is shown in Figure 1. In this process, some production factors are needed, thus forming the production cost. The output value (i.e., value of production), furthermore, is an important indicator of the state of grape production [5]. Our basic data comes from the vineyard input-output survey data collected by China Agriculture Research System—Grape in the main grape-producing areas of China in 2019. The dataset of this paper includes construction cost, land cost, material cost, and labor cost as the input variables of the technical efficiency evaluation model, and selects the output value as the output variable [5]. It consists of 685 samples. Hereinafter, the dataset is referred to as the grape farmer dataset.

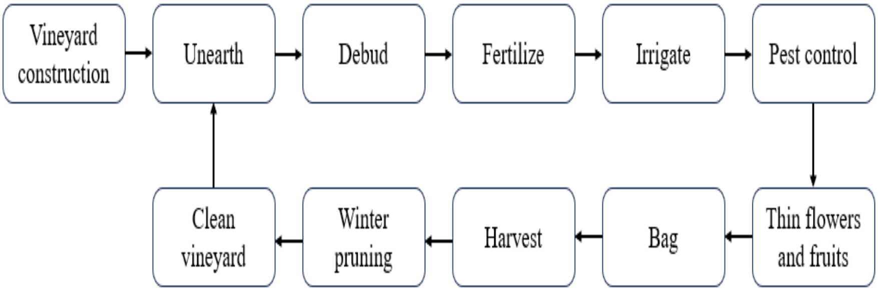

Fig. 1. Cultivation and production process of grapes

2.2 The proposed method

For the the problem of class imbalance, the imbalanced ratio (IR) of a dataset is used as an indicator, which is defined as

$$IR = {}^{\mathrm{m}_s}/_{\mathrm{m}_l} \tag{1}$$

where m_s is the number of samples in the majority class and m_l is the number of samples in the minority class.

If class imbalance is not solved, the effect of the prediction model will decline. Therefore, this paper proposes an evaluation method of agricultural technology efficiency, which is to reduce class imbalance by oversampling before training the model.

The basic process of agricultural production technical efficiency evaluation method based on Borderline-SMOTE and LightGBM is shown in Figure 2. Specifically, the details are as follows: firstly, for the samples without technical efficiency labels, we use the DEA model to calculate the technical efficiency of each production unit; divide the dataset into two classes: samples with technical efficiency =1 (minority class) and samples with technical efficiency <1 (majority class), then set the target imbalanced ratio, apply Borderline-SMOTE algorithm to oversample the minority class samples, so that the imbalanced ratio of the oversampled dataset is equal to the target imbalanced ratio, and set the DEA technical efficiency of the new samples to 1; afterwards, with input variable and output variable as independent variables and DEA technical efficiency as dependent variable, LightGBM regression algorithm is constructed and trained, and then the final technical efficiency regression model is got. For the new samples, the technical efficiency is obtained by substituting the input and output characteristics into the trained regression model without using the DEA model.

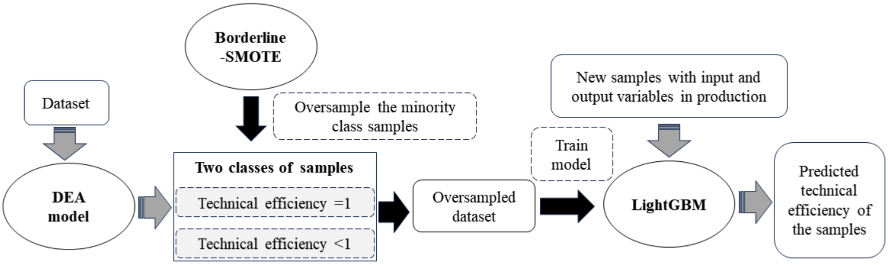
Borderline -SMOTE
Dataset
Oversample the minority class samples
New samples with input and output variables in production
Train model
DEA model
Two classes of samples
Technical efficiency =1
Technical efficiency <1
Oversampled dataset
LightGBM
Predicted technical efficiency of the samples

Fig. 2. Basic process of the proposed method

2.3 Performance evaluation index of regression model

The prediction performance of the regression model is mainly evaluated based on the error between the predicted value and the actual value. In order to compare the prediction performance of different regression prediction models for agricultural production technical efficiency, three indicators are selected to evaluate the models in this paper: mean absolute error (MAE), mean square error (MSE) and coefficient of determination (R^2).

3 Results and discussion

The efficiency prediction model proposed in this paper was applied to the grape farmer dataset. In the machine learning model, the construction cost, land cost, material cost, labor cost and output value of the sample were treated as independent variables, and the DEA technical efficiency was set as the dependent variable.

3.1 Calculation results of technical efficiency based on DEA model

The technical efficiency of each sample in the farmer dataset based on the DEA-BCC model was calculated and then the technical efficiency was taken as the new characteristic of each unit to the dataset. In the dataset of this paper, the number of production units on the efficiency frontier is only 36, accounting for only 5.26% of the total number of samples. The number of production units that are not on the efficiency frontier is 649, accounting for 94.74% of the total samples. The technical efficiency of most vineyards is relatively low, which does not reach the production frontier. The imbalanced ratio of the dataset IR=649/36=18.03. Therefore, it is necessary to oversample the dataset to reduce the modeling error that may be caused by class imbalance.

3.2 Modeling results of the technical efficiency prediction model based on Borderline-SMOTE and LightGBM

Base model selection and parameter optimization. The sample dataset was randomly divided into training set and test set according to 85%:15%. Oversampling was only used for the training set. The training set before oversampling is called original training set. The training set after oversampling is called enlarged training set. The LightGBM model was trained on the enlarged training set. When training the model on the training set, the 6-fold cross-validation method was used, and the coefficient of determination was selected as the evaluation index to determine the optimal values of the key parameters of each model.

In order to achieve better prediction results, the parameters of the selected models including SVR, BR, EN, and LightGBM need to be adjusted and optimized. The grid search method was used to traverse different parameter combinations. The optimized parameters and the optimization results of each model are shown in Table 1.

Table 1. Adjustment results of model parameters

Model	Parameter	Adjustment result	Model	Parameter	Adjustment result
SVR	Regularization parameter	1	LightGBM	boosting_type	gbdt
BR	alpha_1	0.001		learning_rate	0.1
	alpha_2	0.5		n_estimators	100
EN	alpha	1.0		max_depth	20
	l1_ratio	0.5			

Performance comparison of models. In order to explore the optimal imbalanced ratio of the enlarge training set, this paper conducted experiments with different imbalanced ratios of the enlarged training set. The performance of each model with the different imbalanced ratio is shown in Table 2.

According to Table 2, when the imbalanced ratio of the enlarged training set is equal to 4, the vineyard technical efficiency evaluation model based on Borderline-SMOTE (B-SMOTE) and LightGBM performs best, with the smallest MAE, the smallest MSE and the largest R^2. When the imbalanced ratio of the enlarged training set is larger than 4 or less than 4, the effect of the model gradually deteriorates. Therefore, in this paper, the IR was set to 4.

Different models were used to fit the input-output relationship, and the performance of each model is shown in Table 3. The data of Table 2 and Table 3 are the average values of 100 repeated experiments. According to the experimental data, the vineyard technical efficiency evaluation model based on Borderline-SMOTE and LightGBM performs well. Its MAE, MSE, and R^2 on the test set are 4.05E-02, 5.25E-03, and 0.898 respectively, which are better than other comparison models.

The experimental results show that the regression model based on Borderline-SMOTE and LightGBM has more outstanding performance in calculating the technical efficiency, can replace the DEA model to calculate the technical efficiency, and realize the rapid and flexible evaluation of the technical efficiency of agricultural production.

Table 2. Performance comparison of Borderline-SMOTE+LightGBM model with the different IR of enlarged training set

IR of enlarged	Original training set			Test set			Enlarged training set		
training set	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2
1	1.82E-02	9.81E-04	0.981	4.84E-02	7.45E-03	0.856	**1.48E-02**	**7.03E-04**	**0.994**
1.5	**1.67E-02**	**8.79E-04**	**0.983**	4.52E-02	6.65E-03	0.870	1.55E-02	7.73E-04	0.993
2	1.68E-02	9.23E-04	0.982	4.36E-02	6.23E-03	0.876	1.70E-02	9.35E-04	0.992
2.5	1.69E-02	9.39E-04	0.982	4.20E-02	5.65E-03	0.887	1.78E-02	1.01E-03	0.991
3	1.69E-02	9.58E-04	0.982	4.27E-02	5.85E-03	0.888	1.81E-02	1.08E-03	0.989
3.5	1.72E-02	9.95E-04	0.981	4.07E-02	5.40E-03	0.893	1.87E-02	1.14E-03	0.988
4	1.77E-02	1.09E-03	0.979	**4.05E-02**	**5.25E-03**	**0.898**	1.92E-02	1.24E-03	0.986
4.5	1.75E-02	1.06E-03	0.980	4.09E-02	5.51E-03	0.895	1.91E-02	1.25E-03	0.986
5	1.79E-02	1.14E-03	0.978	4.11E-02	5.61E-03	0.892	1.95E-02	1.32E-03	0.984
7	1.97E-02	1.43E-03	0.973	4.22E-02	5.95E-03	0.886	2.13E-02	1.62E-03	0.978
9	2.13E-02	1.66E-03	0.968	4.28E-02	6.11E-03	0.881	2.27E-02	1.85E-03	0.972

Table 3. Performance comparison of vineyard technical efficiency evaluation models

Model	oversampling algorithm	Original training set			Test set			Enlarged training set		
		MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2
BayesianRidge	/	1.09E-01	2.27E-02	0.564	1.10E-01	2.30E-02	0.557	/	/	/
ElasticNet	/	1.09E-01	2.25E-02	0.567	1.12E-01	2.41E-02	0.546	/	/	/
SVR	/	9.03E-02	1.89E-02	0.639	9.15E-02	1.93E-02	0.624	/	/	/

LightGBM	/	2.54E-02	2.30E-03	0.956	4.75E-02	7.10E-03	0.862	/	/	/
Bayesian-Ridge	B-SMOTE	1.38E-01	2.83E-02	0.459	1.38E-01	2.82E-02	0.444	1.56E-01	3.73E-02	0.593
ElasticNet	B-SMOTE	1.38E-01	2.82E-02	0.459	1.39E-01	2.83E-02	0.445	1.56E-01	3.72E-02	0.594
SVR	B-SMOTE	9.86E-02	1.85E-02	0.645	1.01E-01	1.94E-02	0.627	1.26E-01	3.24E-02	0.646
LightGBM	B-SMOTE	**1.77E-02**	1.09E-03	0.979	**4.05E-02**	**5.25E-03**	**0.898**	1.92E-02	1.24E-03	0.986
LightGBM	SMOTE	1.78E-02	**1.04E-03**	**0.980**	4.20E-02	5.74E-03	0.888	1.98E-02	1.25E-03	0.986
LightGBM	ADASYN	1.83E-02	1.15E-03	0.978	4.20E-02	5.76E-03	0.889	**1.90E-02**	**1.20E-03**	**0.987**

4 Conclusion

In order to make up for the defect of low flexibility of the DEA model, aiming at the class imbalance of the dataset, this paper proposes an evaluation method of agricultural production technology efficiency combining the DEA model, Borderline-SMOTE, and LightGBM. The application results in the grape farmer dataset show that when using Borderline-SMOTE to adjust the IR of the training set to 4, the LightGBM model performs excellent. The experimental results also show that the regression model based on Borderline-SMOTE and LightGBM proposed in this paper has high accuracy and can replace the original DEA model for rapid technical efficiency calculation.

The technical efficiency calculated by the DEA model is a relative value, which is used to measure the level of the technical potential of a production unit relative to other production units. For a dataset, if more production units with high or low production levels are added as new decision-making units, it may cause obvious changes in the DEA efficiency value of the original production units. Therefore, in the case of dealing with a large number of new samples, it is recommended to use the DEA model to recalculate the technical efficiency of each sample to obtain more accurate results.

Acknowledgements

Funding: this paper was supported by the earmarked fund for CARS-29.

Reference

1. Zhang X, Ma W, Vatsa P and Jiang S. (2023). Short supply chain, technical efficiency, and technological change: Insights from cucumber production. *Agribusiness*, **39**(2), 371-386. https://doi.org/10.1002/agr.21789
2. Rezitis A N and Kalantzi M A. (2016). Investigating technical efficiency and its determinants by data envelopment analysis: an application in the greek food and beverages manufacturing industry. *Agribusiness*, **32**(2), 254-271. https://doi.org/10.1002/agr.21432

3. Zhong K, Wang Y, Pei J, Tang S and Han Z. (2021). Super efficiency SBM-DEA and neural network for performance evaluation. *Information Processing & Management*, **58**(6), Article 102728. https://doi.org/10.1016/j.ipm.2021.102728
4. Hong H K, Ha S H, Shin C K, Park S C and Kim S H. (1999). Evaluating the efficiency of system integration projects using data envelopment analysis (DEA) and machine learning. *Expert Systems with Applications*, **16**(3), 283-296. https://doi.org/10.1016/S0957-4174(98)00077-3
5. Feng J, Su Y, Gong S, Wang Z and Mu W. (2021). Evaluation method of agricultural production technical efficiency based on ensemble learning. *Transactions of the Chinese Society for Agricultural Machinery*, **52**(S1), 148-155. https://doi.org/10.6041/j.issn.1000-1298.2021.S0.019
6. Imane M, Aoula E S and Achouyab E H. (2022). Using bayesian ridge regression to predict the overall equipment effectiveness performance. *2022 2nd Int. Conf. on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, https://doi.org/10.1109/IRASET52964.2022.9738316
7. Shafiee S, Lied L M, Burud I, Dieseth J A, Alsheikh M and Lillemo M. (2021). Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. *Computers and Electronics in Agriculture*, **183**, 106036. https://doi.org/10.1016/j.compag.2021.106036
8. Khosravi K, Golkarian A, Melesse A M and Deo R C. (2022). Suspended sediment load modeling using advanced hybrid rotation forest based elastic network approach. *Journal of Hydrology*, **610**, 127963. https://doi.org/10.1016/j.jhydrol.2022.127963
9. Tang M Z, Zhao Q, Ding S X, Wu H W, Li L L, Long W and Huang B. (2020). An improved LightGBM algorithm for online fault detection of wind turbine gearboxes. *Energies*, **13**(4), Article 807. https://doi.org/10.3390/en13040807
10. Skevas T and Grashuis J. (2020). Technical efficiency and spatial spillovers: Evidence from grain marketing cooperatives in the US Midwest. *Agribusiness*, **36**(1), 111-126. https://doi.org/10.1002/agr.21617
11. Cheng Y, Peng J, Zhou Z, Gu X and Liu W. (2017). A hybrid DEA-Adaboost model in supplier selection for fuzzy variable and multiple objectives. *Int. Federation of Automatic Control*, https://doi.org/10.1016/j.ifacol.2017.08.2038