An Improved Gradient Enhancement Regression Tree Method to Evaluate the Innovation Capability of Energy Companies

Zhibo Yu*, Ji Li, Chaohui Gu, Juan Zhou, Mengxi Wu,
Research Institute of Natural Gas Economy PetroChina Southwest Oil & Gasfield Company, Chengdu, Sichuan, China
**Corresponding Author.*

Abstract: In recent years, with the introduction of the innovation-driven development strategy, audit departments across the country have actively responded by arranging audit experts to assess the innovation capabilities of enterprises based on their information. They have also studied the impact of implementation on innovation indicators in energy enterprises, aiming to accurately implements to support enterprises and drive regional development. Traditional manual evaluation methods are inefficient and prone to human interference. By using a gradient boosting regression tree model to construct a scoring prediction model, instead of manual evaluation methods, both accuracy and efficiency can be ensured. Experimental results show that this prediction model outperforms other models such as random forest regression and can guarantee prediction accuracy.

Keywords: Innovation Capability; Gradient Boosting; Ensemble Learning; Machine Learning

1. Introduction

In today's society, innovation drive should be regarded as a new power source for economic development, and the roles of scientific and technological capabilities and labor quality in social development should not be viewed in isolation, but should be linked to promote social progress.

The strength of government policy support and financial investment are related to the innovation ability of enterprises. Shao & Wang et al [1] pointed out that government subsidies have a significant incentive effect on the technological innovation activities of enterprises based on the study of unlisted companies in China's Shanghai and Shenzhen A-share markets from 2012 to 2020; Li [2] set up a static and dynamic linear model and a threshold model to analyze the data of China's new energy listed companies, and found that government innovation and non-innovation subsidies can play a role in promoting corporate innovation; Sun et al [3] used the PSM-DID model as a tool, and found that the implementation of R&D expense-related policies has a positive impact on corporate innovation behavior. Therefore, the auditing department pays attention to the development of enterprise innovation capacity, and arranges relevant auditing experts to design different innovation indicators to analyze the changes of enterprise innovation capacity based on the data of enterprise-related business information, environmental conditions, innovation inputs, innovation outputs, financial growth, etc., so as to provide the government departments with the basis for decision-making and optimize the structure of capital inputs, etc. Other scholars in China have proposed to use the method based on support vector machine to construct the innovation ability score prediction model, but this method can eliminate the influence of human factors, but there is room for improvement of the prediction effect. In addition, other scholars have proposed to use random forests and other algorithms to construct a model to predict the impact of innovation policy on the innovation capacity of enterprises, which can effectively help the government to make decisions on enterprise subsidies and other issues. Therefore, relying on machine learning methods to score the prediction of enterprise innovation capacity is conducive to the audit department to analyze the changes in enterprise innovation capacity. At the micro level, it is conducive to discovering the shortcomings of its innovation ability, providing a basis for decision makers to strengthen innovation management, improve innovation mechanisms and enhance

competitive advantages. At the macro level, eligible innovative enterprises can benefit from government policy support, promote regional prosperity and form regional advantages.

The problem of predicting enterprise innovation ability score is actually a regression problem. Friedman [4] pointed out that the key to regression is to optimize the function, the purpose is to find out the function of the dependent variable on the independent variable, so that the loss function expectation is minimized. In recent years, regression prediction is widely used in various fields, such as Lu et al [5] used the total least squares (TLS) method to construct a regression model to predict the life of the battery; Gu et al [6] proposed a wind speed prediction method based on the dynamic spectral regression generalized learning system and multimodal information (DSR-BLS), which makes a certain contribution to the accurate prediction of the wind speed in order to ensure the reliability of the electric power grid and the economical and efficient operation; Mohammad et al [7] proposed a wind speed prediction method based on the dynamic spectral regression generalized learning system and multimodal information, they trained seven regression models for enhanced sediment transport prediction.

Regarding the prediction of corporate innovation capability, the authors constructed different prediction models for experiments, and the experimental results found that the traditional single regression model has problems such as low accuracy and insufficient generalization. For example, Support Vector Regression (SVR), Linear Regression (LR) and other models are not satisfactory in the problem of enterprise innovation capacity prediction. Integrated learning [8] can merge multiple individual learners in order to reduce the generalization error, get more reasonable boundaries, reduce the overall error rate and improve the model performance [9]. In this paper, we propose to use Gradient Boosting Regression Tree (GBRT) algorithm for regression prediction of enterprise innovation ability, GBRT belongs to one of the integrated learning algorithms, and the results show that the model can fit the results similar to the expert score of the auditing team, which is better than other models.

2. Integrated Learning Algorithms

2.1 Integrated Learning

Integration learning has spread to various industries and fields, and it can be seen in problems such as feature selection and regression prediction. As shown in Figure 1, integrated learning relies on a certain strategy to organically combine multiple learners. Among them, learning algorithms such as BP algorithm [10] and SVM algorithm [11] are often used to build individual learners. Integrated learning, on the other hand, integrates the results produced by all individual learners through a certain strategy, such as averaging, voting, and learning. Therefore, constructing a model through integrated learning algorithms will be more stable in terms of results and more capable of generalization than a single model.

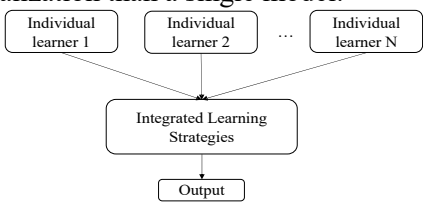

Figure 1. Schematic Diagram of Integrated Learning Principle

Mainstream integrated learning methods are divided into two categories according to the degree of independence between base learners. That is, Boosting boosting algorithm [12], in which base learners must be generated sequentially, and bagging algorithm [13], in which base learners can be generated in parallel. In addition to the difference in the strength of dependency between base learners, the two also differ in sample selection and weight adjustment; Boosting can adjust the weights according to the error rate so that Boosting accuracy is often higher than Bagging.

2.2 Gradient Boosted Regression Tree

The calculation of residuals in the boosting tree algorithm is more complicated, which leads to lower training speed, Friedman first proposed to use gradient boosting regression tree GBRT, which represents the negative gradient value of the loss function as residuals to improve the training speed. GBRT belongs to a kind of generalization of Boosting algorithms, which has been widely used in various fields in recent years. Wang G et al[14] investigated the

relationship between coating friction wear and test parameters by using a ML algorithm based on a gradient boosting regression tree (GBRT) ML algorithm to predict the coefficient of friction (COF) and wear rate, and investigated the relationship between coating friction and wear and test parameters; Abdelbasset et al [15] established three models to estimate the optimal solubility of an anticancer drug, and found that the GBRT was the most effective; Jiang S et al [16] human and used the gradient boosting regression tree (GBRT) algorithm model for slope By comparing the prediction results with those of different algorithms, it is shown that the GBRT model has the highest prediction accuracy.

Algorithm 1 Gradient boosted regression tree GBRT algorithm

Input: training dataset $\{(x_i, y_i)\}_{i=1}^{n}$, loss function $\mathrm{L}(x, f(x))$

Output: strong learner $f_M(x)$

Initialize the model:

$$f_0(x) = arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \quad (1)$$

For m = 1 to M:

M regression trees are constructed iteratively, noting m as the mth tree:

Noting N as the number of samples, find the residual values r_{mi} ,i=1,2,…,N:

$$r_{mi} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \quad (2)$$

The residual value r_{mi} is fitted using the training set $\{(x_i, y_i)\}_{i=1}^{n}$;

Calculate the $h_m(x)$-weight coefficient γ_m.

$$\gamma_m = arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, f_{m-1}(x_i) + \gamma h_m(x_i)) \quad (3)$$

Update the model, where v represents the learning rate:

$$f_m(x) = f_{m-1}(x) + v\gamma_m h_m(x) \quad (4)$$

End for

Output model $f_M(x)$

From equation (4), it can be seen that the number of trees M and the learning rate v affect the prediction accuracy of the model. The number of regression trees M is also called the maximum number of iterations, which is easy to overfitting or underfitting when set improperly; the learning rate v is also called the step size, and the learning rate set appropriately will help prevent overfitting. The number of regression trees M and the learning rate v often need to be adjusted in combination, and their optimization process will be discussed later.

3. Experiment

3.1 Experimental environment and Parameter Settings

The experimental environment used for the experiments in this paper is shown in Table 1.

Table 1. Experimental Environment

Type of experimental environment	Parameters
CPU	AMD Ryzen 7 3700x
computer memory	162GB
Programming language	Python 3.7
Operating system	Windows 10

3.2 Experimental Data Set

The experiments use the data of 500 Sichuan energy enterprises in 2017 provided by the Audit Office of Sichuan Province as the training data, which includes 67 characteristic data of the enterprises and the corresponding audit expert group scores. The expert group score is the mean value of each audit expert after scoring the enterprise's innovation capability based on the enterprise data, and takes the value of floating point value from 0 to 100. Some enterprise characteristics are shown in Table 2.

Table 2. Characteristics of Selected Enterprises

Feature Type	Feature
Basic information	Registered capital, size, etc.
Environmental information	Affiliated administrative regions, industries, etc.
Financial and tax data	Sales revenue growth, profit growth, etc.
Innovation input data	Number of researchers, R&D expenses, etc.
Innovation output data	Number of patents, revenue from high-tech products, etc.

3.3 Data Preprocessing

The preprocessing of enterprise data in this paper mainly consists of (1) data de-duplication. If there is enterprise data with the same enterprise number, it is deleted. (2) Data standardization. Scaling enterprise data by ratio, mapping all types of enterprise data uniformly to the same interval helps to improve the training efficiency. (3) Null value filling. There is a small amount of missing enterprise data in the sample, and the average of the

values in the same column that are not empty is used to fill the missing values.

3.4 Experimental Design

In order to make the GBRT model more effective, Experiment 1 uses Mean Absolute Percentage Error (MAPE) as the evaluation index, adjusts and optimizes the hyperparameters through grid search and cross-validation method, and studies and analyzes the effect of hyperparameters on the error. In order to verify the prediction effect of gradient boosting tree due to other models, Experiment 2 trains Adaboost, Bagging, GBRT, SVM, RF, LR and other models, records their error values and makes comparisons. Among them, this experiment refers to the current widely used error evaluation indexes and selects Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) as the evaluation indexes.

3.5 Results and Analysis of Experiment 1

The maximum number of iterations M and the learning rate v of the hyperparameters of the GBRT model have a large impact on the model prediction accuracy. Experiment I designs different M and v values, uses MAPE as the model evaluation index, and combines grid search and cross-validation to find the optimal M and v combination. The prediction results under different combinations are shown in Table 3.

Table 3 Mean MAPE Averages for Models with Different Maximum Number of Iterations and Learning Rates

M	v				
	0.006	0.010	0.014	0.018	0.020
20	7.5044%	7.3773%	7.3189%	7.2150%	7.1242%
50	7.3207%	7.0845%	6.9875%	6.8264%	6.7017%
100	7.0847%	6.7598%	6.6542%	6.5013%	6.4022%
200	6.7613%	6.4457%	6.3660%	6.2631%	6.2075%
300	6.5683%	6.3067%	6.2453%	6.1664%	6.1447%
400	6.4457%	6.2296%	6.1763%	6.1352%	6.1303%
500	6.3663%	6.1758%	6.1415%	6.1227%	6.1273%
600	6.3081%	6.1474%	6.1276%	6.1181%	6.1201%
700	6.2645%	6.1349%	6.1169%	6.1138%	6.1191%
800	6.2315%	6.1271%	6.1152%	6.1114%	6.1233%

In Table 3, the M values in the first column from 20-800 are the maximum number of iterations in Eq. (2), i.e., M regression trees created. The v value from 0.006-0.020 in the first row is the learning rate in Eq. (4), i.e., the step size for updating the model. The experiment utilizes the K-fold cross-validation method to train the model for K times, and averages the K average absolute percent error values as an evaluation index for hyperparameter tuning, where K takes the value of 10. When M takes a value greater than 600 and v takes a value greater than 0.014, the model effect improvement is less obvious as the values of M and v increase. When the maximum number of iterations and the learning rate are 800 and 0.14, respectively, the model effect is the best, and the average value of MAPE is 6.1233%, which is better than other parameter combinations. The experiment predicts the innovation capability scores of 100 firms after constructing the model using the optimal hyperparameter combinations described above. Only a small number of predicted values differed greatly from the actual values, such as the 58th sample value and the model predictions were unsatisfactory. Most of the predicted values are close to the actual values, such as the 8th and 9th sample values, which can be fitted to approximate the results with the actual values. Therefore, the GBRT model can fit the auditing experts' scores of corporate innovation capability better.

3.6 Results and Analysis of Experiment 2

In order to compare and analyze the effect of different models on the problem of predicting the scoring of enterprise innovation ability, Experiment II uses AdaBoost, Bagging and other algorithms to train the model, and takes MAPE and RMSE as the model evaluation indexes, and the ratio of the training set to the test set division is 3:1.

Table 4. Comparison of MAPE and RMSE of Different Models

Model	MAPE	RMSE
RF	5.8149%	5.7457
GBRT	5.7671%	5.6783
AdaBoost	6.0144%	5.9157
Bagging	6.2738%	6.1737
SVM	7.0588%	6.9870
LR	6.9731%	6.8719

As shown in Table 4, the model obtained from GBRT training is better than the other models in the problem of predicting firms' innovation capability, with a MAPE of 5.7671% and an RMSE of 5.6783. Among them, the model effects of RF and GBRT are close to each other,

but GBRT is slightly superior in this regression problem. The reason for this is that RF is not sensitive to outliers, while GBRT will be more sensitive to outliers based on error rate sampling. Therefore, on this regression problem, GBRT can better fit the audit panel's rating of firms' innovation capability compared to other prediction models.

4. Conclusion

With the introduction of innovation-driven development strategy, government auditing departments pay more and more attention to the assessment of enterprise innovation ability, hoping to quantify the innovation ability, discover the shortcomings of enterprises, assist the decision-making departments to adjust the strength and direction of policy support, and accurately improve the competitiveness of a certain region or a certain industry. However, the workload of the audit expert group in assessing the innovation ability of enterprises based on enterprise-related data is large, and at the same time, it is easy to make misjudgments. In order to improve the efficiency and ability of the Audit Office in assessing the innovation ability of enterprises, this paper proposes to construct a prediction model of the innovation ability score of enterprises using the GBRT algorithm, and with the help of the GBRT's advantages of the robustness of the outliers in the output space, it can be fitted to similar scoring effects of the audit expert group. After verification, the model effect is better than the model constructed by five types of algorithms such as Adaboost.

References

[1] Kaichao Shao, Xiaohua Wang. Do government subsidies promote enterprise innovation? —Evidence from Chinese listed companies. Journal of Innovation & Knowledge, 2023, Vol. 8(4): 100436.

[2] Li Qing, Wang Maoqiong, Liuxu Xiang Li. Do government subsidies promote new-energy firms' innovation? Evidence from dynamic and threshold models. Journal of Cleaner Production, 2020, Vol.286: 124992.

[3] Sun J, Long J. Will R&D Expenses and Deduction Policies Promote Company Innovation? World Scientific Research Journal, 2019, 5(9):147-152.

[4] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of Statistics, 2000, 28(2): 337-407.

[5] Ting Lu, Xiaoang Zhai, Sihui Chen, et al. Robust battery lifetime prediction with noisy measurements via total-least-squares regression, Integration, 2024, 96: 102136.

[6] Ziwen Gu, Yatao Shen, Zijian Wang et al. Wind speed prediction utilizing dynamic spectral regression broad learning system coupled with multimodal information. Engineering Applications of Artificial Intelligence, 2024, Vol.131: 107856.

[7] Mohammad Abdullah Abid Almubaidin. Enhancing sediment transport predictions through machine learning-based multi-scenario regression models. Results in Engineering, 2023, Vol.20: 101585.

[8] Dietterich T G. Ensemble learning. The Handbook of Brain Theory and Neural Networks, 2002, 2:110-125.

[9] Klaus Nordhausen. Ensemble Methods: Foundations and Algorithms. International Statistical Review, 2013, Vol. 81(3): 470.

[10] Whittington J C R, Bogacz R. Theories of error back-propagation in the brain. Trends in Cognitive Sciences, 2019, 23(3): 235-250.

[11] Cortes C, Vapnik V. Support vector machine. Machine Learning, 1995, 20(3): 273-297.

[12] Freund Y. Boosting a weak learning algorithm by majority. Information and Computation, 1995, 121(2): 256-285.

[13] Breiman L. Bagging predictors. Machine Learning, 1996, 24(2): 123-140.

[14] Wang Guoqing, Ruan Yuling, Wang Hongxing et al. Tribological performance study and prediction of copper coated by MoS2 based on GBRT method. Tribology International, 2023, Vol.179.

[15] Kamal W A , H. S E , Sameer A , et al. Development of GBRT Model as a Novel and Robust Mathematical Model to Predict and Optimize the Solubility of Decitabine as an Anti-Cancer Drug. Molecules, 2022, 27(17): 5676-5676.

[16] Song J, Jinyuan L, Sai Z, et al. Landslide risk prediction by using GBRT algorithm: Application of artificial intelligence in disaster prevention of energy mining. Process Safety and Environmental Protection, 2022, 166: 384-392.

