2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)

Cost prediction algorithm for mining equipment based on genetic algorithm and machine learning

Ximing Jian
Zijin Mining Group Company Limited
Longyan, China
ximing.jian@zijinmining.com

Bo Yu
School of Informatics
Xiamen University
Xiamen, China
yu$_b$@zjky.cn

Houjin Chen
School of Informatics
Xiamen University
Xiamen, China
chenhoujin1248@163.com

Lvqing Yang*
School of Informatics
Xiamen University
Xiamen, China
lqyang@xmu.edu.cn

Yifan Liu
University of California San Diego
California, United States
y0rkl1u@outlook.com

Qingkai Wang
State Key Laboratory of Process Automation in Mining & Metallurgy
Beijing, China
56740773@qq.com

Kuangren Tang
Serbia Zijin Copper Co. Ltd.
Bor region, Serbia
kuangren.tang@zijinmining.com

Jiangtao Li
Serbia Zijin Copper Co. Ltd.
Bor region, Serbia
vincent.li@zijinmining.com

***Abstract*—In recent years, China has a large number of large and medium-sized open pit coal mines, and the annual production has increased. However, due to the long mining production cycle, mining process is complex, if not controlled will increase production costs. Some enterprises in order to better cost control and efficiency of open pit mining, started to develop mining cost analysis platform, although effective, due to the lack of more accurate and intelligent cost prediction, enhancement is not high. In this paper, based on the cost analysis platform of mobile equipment for open-pit mining of Zijin Serbian copper mine, a cost prediction model based on XGBoost is proposed, while using data enhancement to solve the problem of too small a dataset and combining with the genetic algorithm (GA) to adjust the hyper-parameters to improve the model performance, and the proposed model achieves good results in the dataset of Zijin Serbian copper mine.**

Keywords-Cost Forecast; Machine Learning; Multi-objective Optimization

I. INTRODUCTION

Open-pit coal mining has the characteristics of large production scale, production safety and so on, and the mining recovery rate is high. For modern mining enterprises, intelligent and mechanized equipment has gradually replaced manual labor as the main productivity.

With the development of machine learning, artificial algorithms and models play an important role in various industries XGBoost is used to analyze and enhance cardiovascular disease diagnosis and prediction[1], early plant disease detection and classification[2] and other biomedical fields, financial statement fraud detection[3], bank customer loss prediction[4] and other financial risk control fields. Good results have been achieved in recommendation systems such as movie recommendation[5]. In the field of mining, some research has also begun to use machine learning algorithms to predict costs, For example, Guo Hongquan et al.[6] used artificial neural networks(ANN), random forests (RF), support vector machines (SVM), and classification and regression trees (CART) to predict mining capital cost (MCC) of open-pit copper mine projects. [7] proposed a mining capital cost estimation model based on regression tree algorithm. Hamidreza Nourali et al.[8] developed a model based on support vector regression (SVR) to estimate the capital cost of mining projects. [9] studied the usability of long short-term memory (LSTM) method in solving the uncertainty problem of unit cost.

However, there are usually a large number of hyperparameters in the machine learning model, and the setting of these hyperparameters will affect the effect of the model. The traditional manual parameter adjustment method not only consumes a lot of time, but also gets no good effect. In order to solve this problem, we use genetic algorithm to solve the problem of model tuning. Genetic algorithm is a kind of optimization algorithm, in the case of modern machine learning and deep learning model parameters of tens of millions, Optimization algorithms are playing an increasingly important role, with genetic algorithms used in [10] to optimize the hyperparameters and architecture of deep learning networks. [11] In order to overcome the limitations of traditional single-objective optimization algorithms in terms of model stability, an improved multi-objective arithmetic optimization algorithm (MOAOA) is proposed to optimize the parameter selection of the Deep Extreme Learning Machine (DELM) model. The problem of low efficiency and poor stability of parameter selection is effectively solved.

On the basis of previous studies, this paper uses injection noise and other methods to solve the problem of less data sets, uses XGBoost for modeling, and uses multi-objective optimization algorithm to adjust hyperparameters, and proposes a cost prediction model in the field of mining cost.

II. MATHOD

The costs incurred in the mining process consist of a series of complex factors, and this study aims to predict the costs incurred by mining equipment such as mining cards and drilling rigs during the working process. This chapter will be presented in three parts: data processing, model prediction and parameter optimization. The overall architecture of the algorithm is shown in Figure 1 below.

979-8-3315-0976-7/25/$31.00 ©2025 IEEE

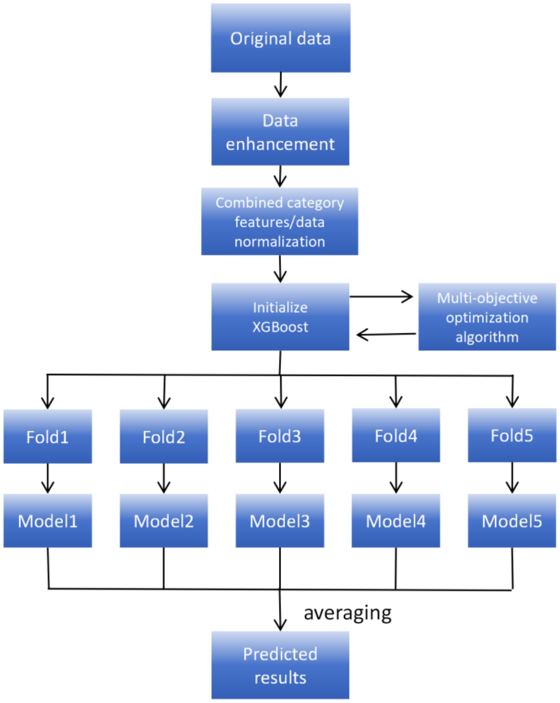

Figure 1. Algorithm General Flowchart

A. Data Processing

The dataset used in this study is the real production process data collected from Zijin Mining MS Mine since January 2019, including equipment type, working time, workload, energy consumption, etc., totaling 6019 data. Due to the small amount of data, this study uses noise injection as well as feature combination to expand the dataset while enhancing the generalization ability of the model.

Noise injection uses Gaussian noise with a mean of 0 and standard deviation of 0.1 as shown in equation (1) below.

$$X^{*} = X + N(0，0.1) \tag{1}$$

The category features are subjected to feature crossover and subsequently the category features are encoded using CatBoostEncoder. Catboost is a goal based categorization encoder. It is a supervised encoder that encodes the categorized columns based on the objective values. The formula is shown in (2) below.

$$\text{result} = \frac{\text{TargetSum+Prior}}{\text{FeatureCount+1}} \tag{2}$$

where TargetSum is the sum of the tagged values of the specified category features, FeatureCount is the number of occurrences of the specified category features, and Prior is the sum of tagged values/total number of category features

B. Cost Prediction Model

XGBoost is an efficient gradient boosting decision tree algorithm, the core of which is to use the integration idea-Boosting idea to integrate multiple weak learners into one strong learner through a certain method, that is, the result of each tree is to fit the residuals between the previous tree and the target value, and all the results are accumulated to get the final result, so as to achieve the enhancement of the effect of the whole model. The formula is shown in (3) below. In this paper, XGBoost model is used for prediction.

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3}$$

where $\hat{y}_i^{(t)}$ is the prediction of sample i, $\hat{y}_i^{(t-1)}$ is the prediction of the first t-1 trees, and$f_t(x_i)$ is the prediction of the t-th tree.

C. Parameter Optimization

In this study, a genetic algorithm based on population entropy is used for parameter optimization, which simultaneously saves the best individuals in each generation of the population as a population pool for cross mutation.The flow chart of the algorithm is shown below. The genetic algorithm will encounter the problem of large population diversity in the early stage and small population diversity in the late stage in the iteration process, thus falling into the local optimum, so in order to ensure the effectiveness of the model, we use the adaptive control strategy based on population entropy. By adjusting the probability of cross mutation and heredity in the late stage of the population, the diversity of individuals in the population is increased in the late stage of the iteration to solve the problem of local optimization. The flowchart is shown in Figure 2.

In this paper, population entropy [12] is used to control the proportion of cross-mutations and genetically generated offspring, and the similarity of an individual P_i to every other individual P_j in the population is calculated by summing their Euclidean distances as shown in Equation (4).

$$\mathrm{P_i} = \sum_{\mathrm{j=1}}^{\mathrm{len(P)}} (|\mathrm{P_i} - \mathrm{P_j}|^2) \tag{4}$$

where len(P) is the population size, $P_{i,}$ P_j are the individuals in the population, expressed as two-dimensional coordinates.

Then, this paper utilizes Gaussian distribution to transform the similarity into probability, as shown in Equation (5).

$$P_i = exp\left\{\frac{-P_i}{2 \cdot \sigma(P_i)}\right\} \tag{5}$$

where $\sigma(P_i)$ is the variance of individuals in the population.

After normalizing the probabilities, the entropy of the aggregate is calculated by the following equation (6).

$$E = -\sum_{i=1}^{len(P)} (P_i * \log_2 P_i) \tag{6}$$

After obtaining the population entropy, it is compared with the historical average population entropy, and the ratio α of cross-variation and heredity is narrowed down according to the comparison results, when the population entropy is smaller and α is larger, the probability of cross-variation is larger. The formula is shown in Equation (7).

$$\alpha = \begin{cases} max(\frac{\widetilde{\alpha}}{\gamma}, 0), E < avg\_E \\ min(\widetilde{\alpha} \ \cdot \gamma, 1), E > avg\_E \end{cases} \tag{7}$$

where $\widetilde{\alpha}$ is the alpha value of the previous iteration, γ is the scaling factor, avg_E is the historical average population entropy.

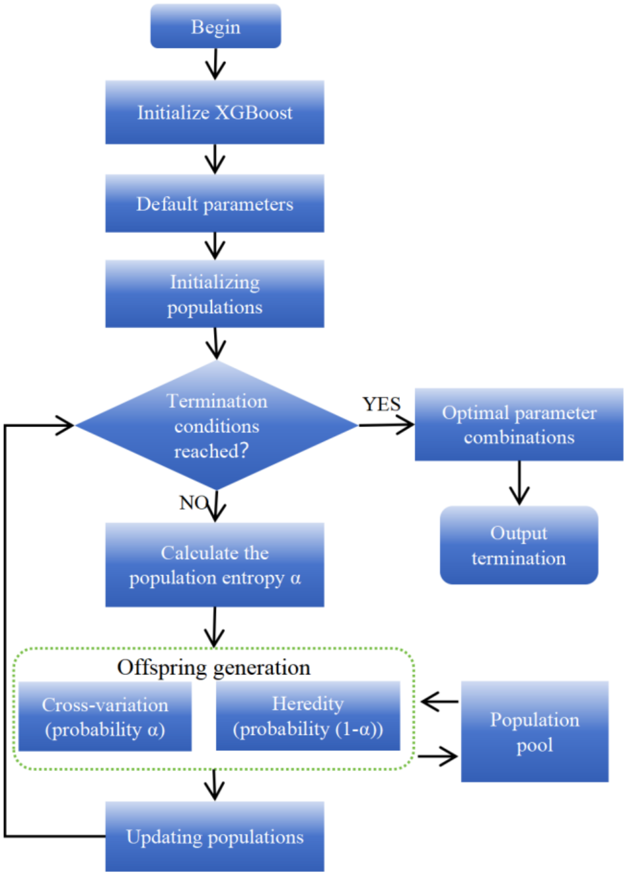
Begin
Initialize XGBoost
Default parameters
Initializing populations
Termination conditions reached?
YES
Optimal parameter combinations
Output termination
NO
Calculate the population entropy α
Offspring generation
Cross-variation (probability α)
Heredity (probability (1-α))
Population pool
Updating populations

Figure 2. Parameter Optimization Flowchart

III. Experiment

This section first describes the assessment metrics. It then discusses the results of the comparison and ablation experiments.

A. *Performance metric*

In this study, we use the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Resolvable Coefficient(R2), which are commonly used in regression models. These three indicators are defined mathematically as follows:

1) MAE:

MAE is a measure of the distance between predicted and observed values. A smaller MAE value indicates a better fitting effect of the model. The mathematical formula for this is as follows:

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \qquad (8)$$

2) RMSE:

RMSE measures the gap between predicted and observed values, emphasizing error amplification. A smaller RMSE signifies more accurate predictions by the model. Its formula is:

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|} \qquad (9)$$

3) R^2:

R2 quantifies the proximity of observed values to the fitted regression line, indicating how well the model elucidates dependent variable variability. Ranging from 0 to 1, higher values denote superior fit. The formula is:$\bar{y}_i$

$$R^2(y,\hat{y}) = 1-\frac{\sum_{i-1}^{n}(y_i-\widehat{y_i})^2}{\sum_{i-1}^{n}(y_i-:\bar{y}_i)^2} \qquad (10)$$

where n is the number of samples, y is the true value, and $\hat{y}$ is the observed value.$\bar{y}$ is the average value of y.

B. *Experimental results*

We first compared the effect of XGBoost model with machine learning models such as LightGBM, AdaBoost, linear regression, random forest, etc., in which the parameters in XGBoost model and LightGBM model are kept the same, and the experimental results are shown in the Table 1 below, which shows that, among the above five models, XGBoost has the smallest MAE and the fitting effect R^2 is the best.

Table 1 Comparative experimental results of five models

Model	MAE	RMSE	R^2
XGBoost	**3.40**	9.42	**0.98**
LightGBM	3.67	9.35	0.97
AdaBoost	6.69	9.97	0.97
Linear Regression	6.98	11.5	0.95
Random Forest	4.43	**7.70**	0.97

Since we used noise injection operation in the data processing stage, in order to prevent the model from overfitting, we used five-fold cross-validation, and the results are shown in Figure 3, where XGBoost-Optimal is the result after using optimization means.

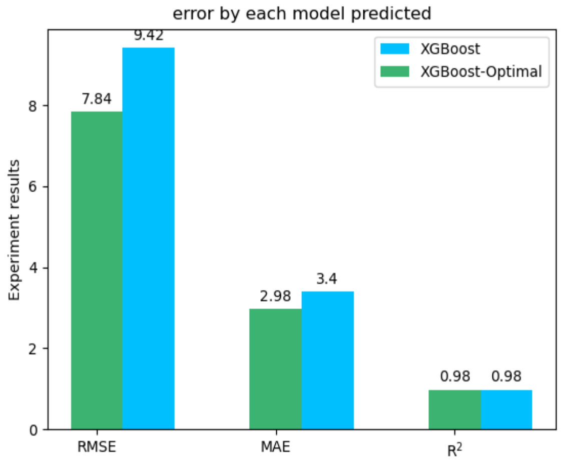

Figure 3 Comparison of experimental results

We use genetic algorithms to tune and optimize the hyperparameters for XGBoost, which contain the depth of the model tree, the number of evaluators, the learning rate, the sampling ratio, and the sample weights, and the combination of hyperparameters given by using genetic algorithms gives a further improvement in the effectiveness of the model as compared to the original parameters that we used. The parameter comparison is shown in Table 2 and the experimental results are shown in Figure 4, where XGBoost-Tuning_Parameters is the result of optimizing the hyperparameters using the improved genetic algorithm. It is worth noting that although the results of parameter optimization will increase the complexity of the model, the training and prediction time of the model are basically the same, and the accuracy is greatly improved, which is an acceptable improvement.

IV. Conclusions

This paper proposes a method based on XGBoost and multi-objective optimization for predicting the mining cost of open pit mining equipment, which firstly uses methods such as noise injection for data enhancement and XGBoost and five-fold cross-validation for prediction, and due to the large number of hyper-parameters of the model, this paper uses genetic algorithms to adjust and optimize the hyper-parameters, and the experimental results prove the method's The experimental results proved the effectiveness of the method.

Table 2 Parameter comparison

Parameter name	Original	Optimization results
max_depth	6	15
n_estimators	500	1445
learning_rate	0.08	0.0471
subsample	0.5	0.0702

Note: max_depth represents the depth of the tree, n_estimators represents the number of decision trees, learning_rate is the learning rate, subsample is the proportion of samples.

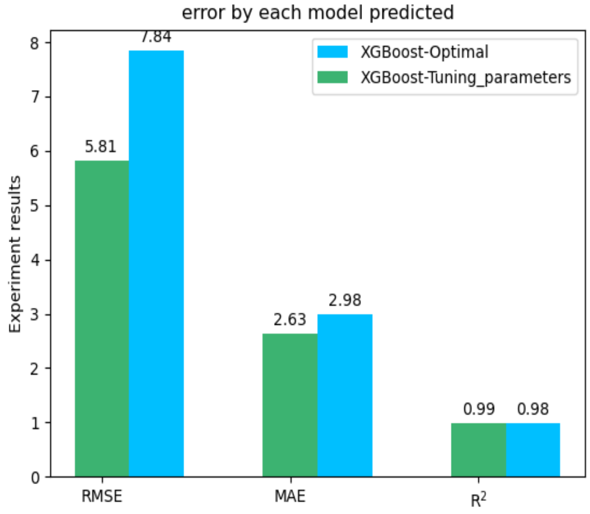

Figure 4 Experimental results after optimization of parameters

Acknowledgment

This work was supported in part by the project “Development of a Cost Intelligent Analysis System for Zijin Mining Equipment” under Project No. 20243160A0723, in part by the project “Research and Application of a Mining Production Control Platform Based on IoT and Big Data” under Serial No.20240866, funded by the 2024 Industry-University-Research Collaboration Program of the Longyan Xiamen University Integration Research Institute, and in part by the project “Research and Application of Mine Workers’ Behavior Recognition Based on RFID and Deep Learning,” funded by the Open Research Fund of the State Key Laboratory of Intelligent Manufacturing in Metallurgical Processes (2025).

References

[1] Sandeep Tomar, Deepak Dembla and Yogesh Chaba, “Analysis and Enhancement of Prediction of Cardiovascular Disease Diagnosis using Machine Learning Models SVM, SGD, and XGBoost” International Journal of Advanced Computer Science and Applications(IJACSA), 15(4), 2024. http://dx.doi.org/10.14569/IJACSA.2024.0150449.

[2] Shukla, P. et al. (2024). Detection and Classification of Diseases in Coffee Plant Using CNN-XGBoost Composite Model. In: Kole, D.K., Roy Chowdhury, S., Basu, S., Plewczynski, D., Bhattacharjee, D. (eds) Proceedings of 4th International Conference on Frontiers in Computing and Systems. COMSYS 2023. Lecture Notes in Networks and Systems, vol 975. Springer, Singapore. https://doi.org/10.1007/978-981-97-2614-1_43.

[3] Xinfeng Dou, Rong Liu, Shengpeng Yin, "WOA-XGBoost-based financial statement fraud detection, " Proc. SPIE 13184, Third International Conference on Electronic Information Engineering and Data Processing (EIEDP 2024), 131846L (5 July 2024); https://doi.org/10.1117/12.3032976.

[4] Hu, Z., Dong, F., Wu, J., Misir, M. (2024). Prediction of Banking Customer Churn Based on XGBoost with Feature Fusion. In: Tu, Y.P., Chi, M. (eds) E-Business. New Challenges and Opportunities for Digital-Enabled Intelligent Future. WHICEB 2024. Lecture Notes in Business Information Processing, vol 517. Springer, Cham. https://doi.org/10.1007/978-3-031-60324-2_13.

[5] Gopal Behera, Sanjaya Kumar Panda, Meng-Yen Hsieh, Kuan-Ching Li, “Hybrid collaborative filtering using matrix factorization and XGBoost for movie recommendation, ” Computer Standards & Interfaces, Volume 90, 2024, 103847, ISSN 0920-5489, https://doi.org/10.1016/j.csi.2024.103847.

[6] Hongquan Guo, Hoang Nguyen, Diep-Anh Vu, Xuan-Nam Bui, ”Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach, ”Resources Policy, Volume 74, 2021, 101474, ISSN 0301-4207, https://doi.org/10.1016/j.resourpol.2019.101474.

[7] Chengkai Fan, Na Zhang, Bei Jiang & Wei Victor Liu. (2023) Prediction of truck productivity at mine sites using tree-based ensemble models combined with Gaussian mixture modelling. International Journal of Mining, Reclamation and Environment 37:1, pages 66-86.

[8] Hamidreza Nourali, Morteza Osanloo, “Mining capital cost estimation using Support Vector Regression (SVR), Resources Policy, ”Volume 62, 2019, Pages 527-540, ISSN 0301-4207, https://doi.org/10.1016/j.resourpol.2018.10.008.

[9] Özdemir, A. C. (2022). PREDICTION OF UNIT COST OF AN OPEN-PIT CHROME MINE USING LONG SHORT-TERM MEMORY (LSTM) METHOD. Proceedings of the 27th International Mining Congress and Exhibition of Turkey, IMCET 2022, 116–126.

[10] Christian Kazadi Mbamba, Damien J. Batstone, “Optimization of deep learning models for forecasting performance in the water industry using genetic algorithms, ”Computers & Chemical Engineering, Volume 175, 2023, 108276, ISSN, 0098-1354, https://doi.org/10.1016/j.compchemeng.2023.108276.

[11] L. Li, Z. Huang and G. Ding, "Indirect Prediction for Lithium-Ion Batteries RUL Using Multi-Objective Arithmetic Optimization Algorithm-Based Deep Extreme Learning Machine, " in IEEE Access, vol. 11, pp. 110400-110416, 2023, doi: 10.1109/ACCESS.2023.3320058.

[12] Qian, W.; Xu, H.; Chen, H.; Yang, L.; Lin, Y.; Xu, R.; Yang, M.; Liao, M. A Synergistic MOEA Algorithm with GANs for Complex Data Analysis. Mathematics 2024, 12, 175. https://doi.org/10.3390/math12020175