Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Research on Training Performance Prediction Model Based on LightGBM

Jingzhe Peng [1, *], Feiyang Pan [1], Jiaxing Li [1], Yixin Duan [2]

[1] School of Physical and Electronic Sciences, Changsha University of Science and Technology, Changsha, China, 410114

[2] School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, China, 410114

* Corresponding author: 3460334283peng@gmail.com

Abstract. This study aims to build an accurate performance prediction model to assist the training and development of skilled workers. This article uses the advanced LightGBM algorithm and combines it with rich technical worker training data in a certain industry to predict trainees' training results. In the data processing stage, this article conducted detailed data cleaning to eliminate outliers and ensure the accuracy and reliability of the data. At the same time, by introducing new variables, this paper further enhances the feature expression ability of the model, thereby improving the accuracy of prediction. To further optimize the model performance, this article innovatively uses the particle swarm algorithm to fine-tune the parameters of LightGBM. Experimental results show that the optimized model performs well in predicting students' training performance, with not only high prediction accuracy but also good stability. This model is expected to become a powerful tool for the training and development of skilled workers, helping various industries to evaluate and improve training effects more efficiently, thereby better meeting the needs of modern production and promoting the improvement of the overall quality of the skilled worker team.

Keywords: LightGBM Model; Training Performance Prediction; Particle Swarm Algorithm; Gradient Boosting Decision Tree.

1. Introduction

With the deep integration of the global economy and the rapid development of science and technology, the training and development of skilled workers has gradually become a key factor in promoting the continued progress of various industries. Especially in the fields of manufacturing, engineering construction and high-tech industries, the demand for highly skilled labor is increasing day by day, which has promoted the widespread popularization and in-depth development of professional skills training. Every year, many skilled workers devote themselves to various professional skills improvement courses to better meet the needs of modern production through systematic learning and practice. To evaluate and improve training effects more efficiently, it is particularly important to establish a scientific and accurate performance prediction model [1].

Establish a performance prediction model using a huge data set involving 6 different types in an industry, covering 160 training schools, and a total of 32,165 students [2]. This data set records in detail the assessment results of each student's five core skills when entering and leaving school, as well as their overall score performance. This article hopes that by building a suitable prediction model based on machine learning [3], it can accurately predict the training results of students, and then provide more personalized training programs for educational institutions and enterprises, optimize resource allocation, and ultimately promote the team of skilled workers. Improvement of overall quality. This will not only help improve training efficiency and quality, but also inject new vitality into the long-term development of related industries.

2. Data preprocessing and feature engineering

2.1. Data cleaning

For the rigor and accuracy of subsequent data analysis, the data needs to be cleaned; given that the data is quantitative data and the data volume is greater than 5000, which is a large sample data, the Kolmogorov–Smirnov test is first used to test the normal distribution of the data. The KS test statistic is D the maximum vertical distance between the empirical distribution function $F_n(x)$ and the theoretical distribution function F(x). Among them, the empirical distribution function $F_n(x)$ is defined as $x$the proportion of data points in the sample that are less than or equal to. For na sample containing n data points, the empirical distribution function can be expressed as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\,(X_i \le x) \tag{1}$$

Where $I(X_i \le x)$is an indicator function that takes a value of $X_i \le x$ at that time and a value of 0 otherwise. This article uses the maximum value among $\sup_x$ all possible $x$values, and the KS test statistic Dcan be expressed as:

$$D = \sup_x |F_n(x) - F(x)| \tag{2}$$

This statistic, D which is an empirical distribution function, with the critical value, the results show that all variables approximately satisfy the normal distribution, so 3Sigma processing is used. Outliers outside the range of the mean plus or minus 3 times the standard deviation are excluded.

2.2. New variable derived.

After cleaning the data, to find suitable characteristic variables to establish a model, this article conducts Pearson correlation analysis on each possible existing variable and the total score of leaving school [4]. Let x_i sum be the y_i abscissa $i$and ordinate of the observation value, $\overline{x}$ and sum be the mean of $\overline{y}$ sum $y$respectively x, then the degree of linear correlation between the two variables r is:

$$r = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{3}$$

The results show that among all $|r| \ge 0.7$variables, the school number has the highest correlation, followed by the scores of each student's school entry skills assessment. To summarize this related analysis, this article believes that the entrance examination scores of each student's skills represent the individual's level before admission, and the school number can reflect the education level of each school. However, the school number cannot quantify the education level, so new variables will be introduced to explain the education level of each school [5]. This article first subtracts the entrance examination scores from the school leaving examination scores of each student to obtain the performance changes before and after training:

$$\begin{aligned} &Skill\_Variations \\ &\quad = School\_Leaving\_Skills\_Assessment\_Results \\ &\quad - Entrance\_Skills\_Assessment\_Results \end{aligned} \tag{4}$$

Then take each school as the basic unit, if each school contains n student, and calculate the average change in each skill of each student in the college:

$$Skill\_Change\_Average = \sum_{i=1}^{n} Skill\_Variations_i \quad (5)$$

Pearson correlation analysis was performed again, and the new variables all had good correlations [4].

3. Establishment of LightGBM model

3.1. Model building

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework based on Gradient Boosting Decision Tree (GBDT), which is an iterative decision tree algorithm that fits the negative gradient of the loss function. To construct each decision tree, and add the results of these trees to obtain the final prediction result [6]. This article uses LightGBM for model training [7, 8]. The core functions of LightGBM and GBDT are the same and can be expressed as:

$$F_m(x) = \sum_{i=1}^{m} f_i(x) \quad (6)$$

This article selects the features most relevant to students' performance from the data set: each student's entrance scores for each skill and each student's school skill training level as the input of the model; the school leaving scores of each college are used as the output of the model for model training. Therefore, $F_m(x)$ it is $m$the prediction function of students' school leaving assessment scores after the first iteration, and $f_i(x)$ it is $i$the prediction function of the school leaving assessment scores of the first decision tree. During the training process, each new decision tree $f_i(x)$ is obtained by fitting the current prediction error (the negative gradient of the loss function).

3.2. Model solving

During the model solving process, to improve the prediction accuracy, the parameters of LightGBM need to be optimized. This article uses particle swarm algorithm for parameter optimization. Each particle represents a set of LightGBM parameter configurations, including learning rate, maximum depth of the tree, etc. Through the iterative search of the particle swarm algorithm, this article can find an optimal set of parameter configurations to achieve the best prediction performance of the model [9].

In the initial stage of the algorithm, a set of particles needs to be randomly generated, and the position and velocity of each particle are randomly initialized. Each dimension of the position vector corresponds to a LightGBM parameter, and each dimension of the velocity vector determines the moving direction and step size of the parameter in the search space. In each iteration, each particle updates its speed and position based on its own historical optimal position (individual optimal) and the historical optimal position of the entire group (global optimal) [10]. The update formulas for speed and position are as follows:

$$v_{id}^{k+1} = w \cdot v_{id}^{k} + c_1 \cdot r_1 \cdot (p_{id}^{k} - x_{id}^{k}) + c_2 \cdot r_2 \cdot (g_d^{k} - x_{id}^{k}) \quad (7)$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \quad (8)$$

Among them, v_{id}^{k} and x_{id}^{k} represent $i$the speed and position of the th particle in the th dimension respectively; $w$are the $k$inertial $d$weights; c_1 and c_2 are the learning factors; r_1 and r_2 are random numbers; p_{id}^{k} are the individual optimal positions; x_{id}^{k} and are the global optimal positions.

When the validation error (average error of cross-validation) no longer decreases significantly over consecutive iterations, the algorithm will stop iterating and output the currently found optimal parameter combination. After the parameter optimization is completed, the optimized parameters are

used to train the LightGBM model. During the training process, the model will continuously fit the characteristics of the data and reduce the prediction error through the gradient boosting method. The training process of LightGBM can be expressed as the process of minimizing the loss function:

$$\min_{\Theta} \sum_{i=1}^{N} L(y_i, F(x_i; \Theta)) \tag{9}$$

Where $N$is the sample size and y_i is the iThe true value of $F(x_i; \Theta)$ the sample is the predicted value of the Θ LightGBM model for the sample, $i$the parameter set of the model, $L$and the loss function. After the training is completed, this article can obtain a LightGBM model that can accurately predict student performance.

3.3. Model evaluation

Model evaluation is an important step in determining the performance of the model. In the evaluation of the LightGBM model, this article uses the root mean square error (RMSE), the mean absolute error (MAE) and the coefficient of determination (R^2) as the main evaluation indicators. In addition, we also go deeper by plotting feature importance histograms and loss function plots for each iteration to further understand which features have the greatest impact on grade prediction [11]. For evaluation indicators, take the root mean square error (RMSE) as an example. The root means square error measures the deviation between the model's predicted value and the actual value. Its mathematical formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

Among them, $y_i$is the actual value, $\hat{y}_i$ is the model predicted value, $n$and is the sample number. The smaller the RMSE value, the more accurate the prediction of the model is. The evaluation indicators of the trained model are shown in Table 1:

Table 1. Indicators for the assessment of the model

Data Set	RMSE	MAE	R^2
Training Set	0.515141	0.401828	0.937308
Test Set	0.528836	0.409949	0.897131

According to the table, after model training and verification, the deviation between the model prediction value and the actual value is small, and the model prediction is relatively accurate. The model fits the data well, and the model can explain the variation in the data well. This article selects RMSE as the loss function to draw the loss function graph of each iteration Figure 1:

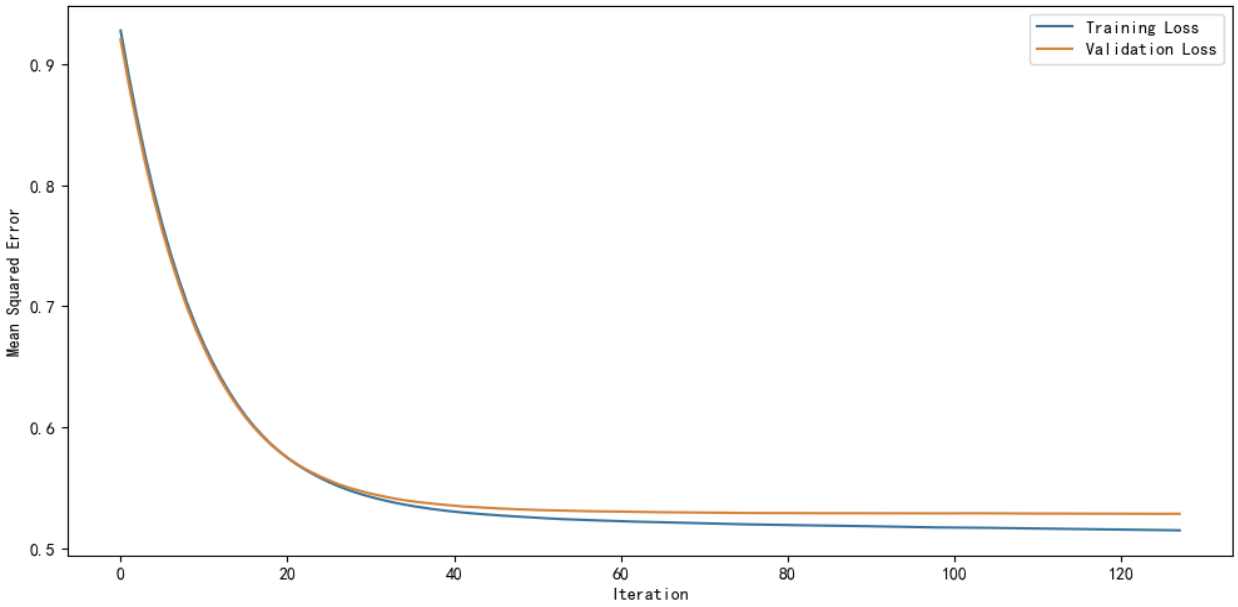

Figure 1. Loss function plot for each iteration

From Figure 1, this article can see that as the number of iterations increases, the loss value of the model gradually decreases and becomes stable. This shows that there is no obvious over-fitting or under-fitting phenomenon in the model during the training process, and the performance of the model is stable.

4. Conclusion

Deeply explores the problem of predicting the training performance of technical workers by using the LightGBM algorithm. Based on a rich training data set in an industry, this article conducted detailed data processing and feature engineering, effectively improving the model's predictive ability by eliminating outliers and introducing new variables. In the process of model construction, this paper adopts the LightGBM algorithm based on gradient boosting decision tree and uses the particle swarm algorithm to optimize the model parameters, thereby obtaining the optimal model configuration. This strategy not only improves the prediction accuracy of the model, but also enhances the stability and generalization ability of the model.

LightGBM prediction model constructed in this article has achieved significant results in predicting trainee training performance. This model can accurately capture the performance characteristics of students during the training process and effectively predict their future performance trends. This result has important practical significance for improving the quality and efficiency of technical worker training. In addition, this study also revealed some key factors that affect training performance, such as the educational level of the school, the entry skill level of the trainees, and the improvement of skills during the training period. These findings help this paper gain a deeper understanding of the nature of skilled worker training and provide a useful reference for the development of future training strategies.

In summary, this study constructed an effective training performance prediction model based on the LightGBM algorithm, which provides strong support for the training and development of skilled workers. This article hopes that this result can be widely used in related industries and make a positive contribution to improving the overall quality and skill level of technical workers.

References

[1] BREJC M, ŠIROK K, KOREN A. Training as strategy for school self-evaluation capacity building [J/OL]. Management, 2019, 24: 73-87. http://dx.doi.org/10.30924/ mjcmi.24.si.5. DOI: 10.30924/mjcmi. 24. si. 5.

[2] ZHANG Y, AN R, CUI J, et al. Undergraduate Grade Prediction in Chinese Higher Education Using Convolutional Neural Networks[C/OL]//LAK21: 11th International Learning Analytics and Knowledge Conference. 2021. http://dx.doi.org/10.1145/3448139.3448184. DOI: 10.1145/3448139. 3448184.

[3] MAURYA L S, HUSSAIN M S, SINGH S. Developing Classifiers through Machine Learning Algorithms for Student Placement Prediction Based on Academic Performance[J/OL]. Applied Artificial Intelligence, 2021, 35 (6): 403 - 420. http://dx.doi.org/10.1080/08839514.2021.1901032. DOI:10.1080/08839514.2021.1901032.

[4] BAAK M A, KOOPMAN R F, SNOEK H, et al. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics[J]. Cornell University - arXiv, 2018.

[5] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods[J/OL]. Computers & Electrical Engineering, 2014: 16-28. http://dx.doi.org/10.1016/j.compeleceng.2013.11.024. DOI: 10.1016/j.compeleceng.2013. 11. 024.

[6] LIANG Y, WU J, WANG W, et al. Product marketing prediction based on XGboost and LightGBM algorithm[C/OL]//Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition. 2019. http://dx.doi.org/10.1145/3357254.3357290. DOI: 10. 1145/3357254. 3357290.

[7] BENTÉJAC C, CSÖRGŐ A, MARTÍNEZ-MUÑOZ G. A comparative analysis of gradient boosting algorithms[J/OL]. Artificial Intelligence Review, 2021: 1937 - 1967. http://dx.doi.org/10.1007/s10462-020-09896-5. DOI: 10.1007/s10462 - 020 - 09896 - 5.

[8] SUN X, LIU M, SIMA Z. A novel cryptocurrency price trend forecasting model based on Light [J/OL]. Finance Research Letters, 2020: 101084. http: //dx.doi.org/10.1016/j.frl.2018. 12. 032. DOI: 10.1016/j.frl.2018. 12. 032.

[9] WANG D, TAN D, LIU L. Particle swarm optimization algorithm: an overview[J/OL]. Soft Computing, 2018: 387 - 408. http: //dx.doi.org/10.1007/s00500-016-2474-6. DOI: 10.1007/s00500-016-2474 - 6.

[10] ASTERIS P, JAIN M, SAIHJPAL V, et al. An Overview of Variants and Advancements of PSO Algorithm [J]. APPLIED SCIENCES-BASEL, 2022.

[11] DEGENHARDT F, SEIFERT S, SZYMCZAK S. Evaluation of variable selection methods for random forests and omics data sets[J/OL]. Briefings in Bioinformatics, 2019: 492 - 503. http: //dx.doi.org/10.1093/bib/bbx124. DOI:10.1093/bib/bbx124.