Employee Performance Prediction based on the Second-order Stacking Algorithm

Yanming Chen [1,a*], Xinyu Lin [2,b] and Kunye Zhan [3,c]

EMAIL: [a]21ymchen@stu.edu.cn; [b]xinyul2002@163.com; [c]2021040486@email.szu.edu.cn

[1,*] Shantou University, Shantou, Guangdong, China

[2] South China Normal University, Guangzhou, Guangdong, China

[3] Shenzhen University, Shenzhen, Guangdong, China

Abstract: This paper attempts to establish an performance prediction model for employees in the field of human resource management based on the second-order stacking algorithm which is an improvement of stacking algorithm. Firstly, the Adaboosting feature importance ranking method is used for feature selection, and then bagging and stacking algorithms are used to establish regression models as control experiments. Finally, a second-order stacking algorithm is used to establish a performance prediction model for employees, achieving minimal error.

Keywords: Employee Performance, Second-order Stacking Algorithm, Machine Learning, Ensemble Learning, Adaboosting

1 Introduction

Employee performance management is a critical issue in human resource management, and employee performance prediction can better assist managers in performance management. Many scholars have conducted relevant research, such as “Predictive power of training design on employee performance: an empirical approach in Pakistan's health sector” [1], “Performance Appraisal System: A Predictor for Performance of Employees in Engineering Sector” [2], and “Application of Data Mining Classification in Employee Performance Prediction” [3].

However, the previous studies have not established a more accurate and less error-prone employee performance prediction model. Therefore, this paper employs a second-order stacking algorithm to establish an employee performance prediction model, which can effectively reduce the prediction error.

2 Theoretical foundation

Machine learning is a branch of artificial intelligence that leverages data and algorithms to enable computers to learn and improve their performance automatically. By utilizing machine learning, we can discover patterns and correlations in data, thereby enhancing prediction accuracy and efficiency, and facilitating our comprehension and resolution of practical issues

FFIT 2023, July 07-09, Chongqing, People's Republic of China

DOI 10.4108/eai.7-7-2023.2338055

[4].The regression prediction model is an important model in machine learning, which uses historical data to make predictions.

Ensemble learning is used to improve the stability and accuracy of the regression prediction model by integrating multiple models, such as bagging, boosting, blending, and stacking. The second-order stacking algorithm employed in this paper is an improved version of stacking. The stacking algorithm is a non-linear ensemble method that employs K-fold cross-validation on each base learner (first-layer model) and utilizes the resulting features to train a meta-learner (second-layer model) [5].The flowchart of the stacking algorithm is illustrated in **Figure 1.**

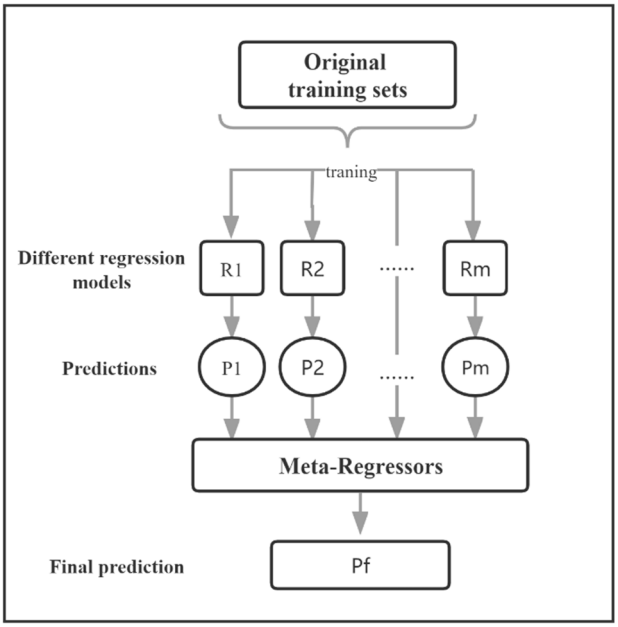

Figure 1: The flowchart of the stacking algorithm

3 Materials and Methods

3.1 Dataset used in the study

The employee performance dataset selected in this paper is obtained from a public database called Kaggle.com. The dataset consists of 1017 observations, including the performance and other related information of employees. The distribution of employee performance in the dataset ranges from 0.23 to 1.11, as illustrated in the performance distribution chart in **Figure 2.**

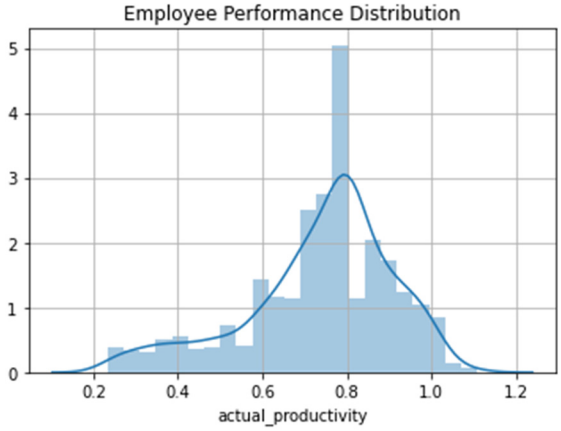

Figure 2: The distribution of the employee performance

The dataset under consideration consists of a total of 25 variables, among which the variable "Employee performance" is taken as the dependent variable and the remaining 24 variables are treated as independent variables. The dataset has already transformed some categorical variables into dummy variables which are 0-1 variables, so there are no categorical variables in the dataset, and all variables are numerical. Rudimentary information about some numerical variables is presented in **Table 1.**

Table 1 Rudimentary information about some numerical variables

	Count	Min	Max	Mean
smv	1017	2.90	54.56	15.15
over_time	1017	0.00	15120.00	4532.94
incentive	1017	0.00	3600.00	40.69
no_of_workers	1017	2.00	89.00	34.85
month	1017	1.00	3.00	1.72
department_finishing	1017	0.00	1.00	0.20
wip	594	7.00	23122.00	1183.18

The variable "smv" refers to Standard Minute Value, which represents the assigned time for a task. "over_time" represents the amount of overtime in minutes by each team and the variable "no_of_workers" represents the number of workers in each team. The distribution of these variables are illustrated in **Figure 3.**

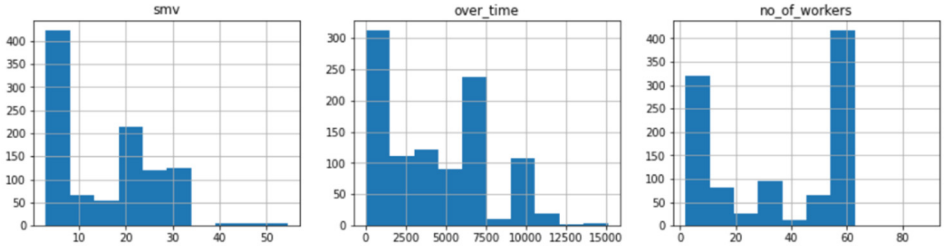

Figure 3: The distribution of three typical variables

3.2 Methods

In this paper, missing values and outliers were handled initially. In order to alleviate the problem of overfitting, the Adaboosting feature importance ranking method was employed for feature selection. Subsequently, a bagging algorithm and a stacking algorithm were selected to construct the regression model. Finally, a second-order stacking algorithm was employed to establish the ultimate employee performance prediction model.

3.2.1 Data cleaning and preprocessing

In this dataset, only the variable "wip" has missing values, with 423 missing values out of 1017 observations. Given that the missing values account for a large proportion of the variable, and it is not of significant importance to this study, the variable was directly deleted.

Moreover, the variable "Employee performance" contains outliers, as 34 observations exceed the normal range of 0-1. Therefore, these 34 observations were replaced by the mean value to address this issue.

In order to improve the performance of the algorithm, this paper uses the minimum-maximum scaling method to scale all numerical data, scaling the data between 0 and 1. The formula for the minimum-maximum scaling method is as follows (1) :

$$X' = \frac{X - Xmin}{Xmax - Xmin} \tag{1}$$

where X is the original value, $Xmax$ and $Xmin$ are the maximum and minimum values, respectively, and X' is the transformed value.

3.2.2 Adaboosting feature importance ranking for feature selection

As ensemble learning algorithms require high precision data, this paper utilizes Adaboosting feature importance ranking for feature selection [6]. Firstly, the dataset is divided into a training set and a test set with a ratio of 8:2. Then, an Adaboosting model with a decision tree as the base model is established. Finally, features are filtered based on the feature importance generated by the model, removing features with an importance value less than 0.01. The six features with the highest importance are shown in **Table 2.**

Table 2 Top 6 features of the Adaboosting feature importance ranking

Variables	Importance
incentive	0.23
no_of_workers	0.19
smv	0.12
over_time	0.11
quarter_Quarter4	0.08
month	0.04

3.2.3 Model building

Bagging is another ensemble learning algorithm, which generates multiple subsets by repeatedly sampling the training set with replacement, trains a base learner on each subset, and then aggregates their results by averaging or voting to obtain the final prediction. Randomforest is a bagging algorithm based on decision trees [7]. Therefore, in this paper, randomforest regression and stacking algorithms are used as control experiments, where the first-layer models of the stacking algorithm include ridge regression, decision tree, Lasso regression, and support vector machine regression, and the second-level model uses decision tree.

Finally, a second-order stacking algorithm is employed in this paper to construct the regression model. Essentially, this involves building multiple different first-order stacking models and then integrating these models using the stacking algorithm, treating them as the first-layer models. The output values of these models are then used as the features of the second-layer model, which is used to make the final prediction. The process of the second-order stacking algorithm is shown in **Figure 4.**

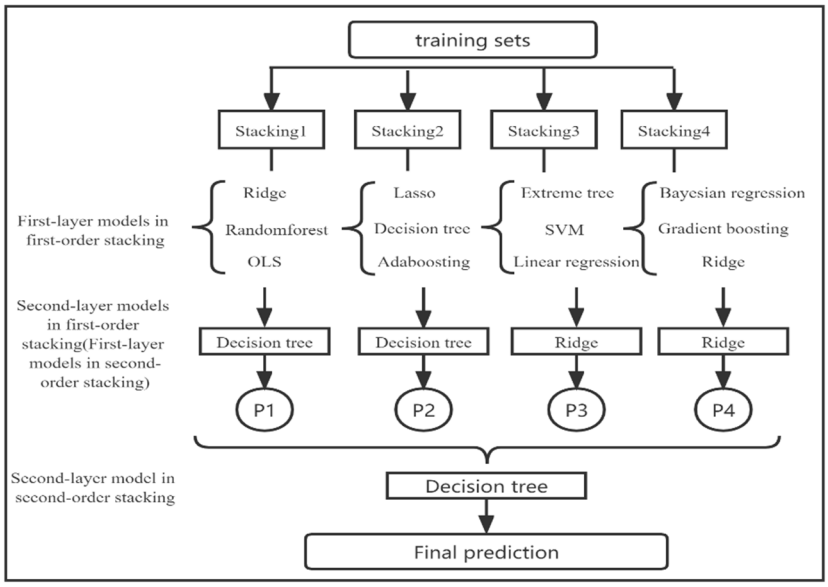

Figure 4: The process of the second-order stacking algorithm

4 Experiments & Results

4.1 Experiment environment

The dataset comes from kaggle.com. The experiment was done in python 3.7.0, and the configuration of the computer is shown in **Table 3.**

Table 3 The configuration of the computer

Hardware	Hardware model
CPU	Intel core i7 CPU 2.90 GHZ
RAM	60.0 GB

4.2 Experiments and results

This paper initially conducts a comparative experiment and the results are demonstrated in **Table 4.**

As the target predicted values are within the range of 0 to 1, the commonly used metric MSE is not quite suitable for evaluating this experiment, as squaring the errors would result in extremely small values. Therefore, we adopt MAE as the metric for assessing the performance of the model [8].

Table 4 The results of the comparative experiment

Model	MAE (training)	MAE (testing)
Randomforest	0.03	0.07
Stacking	0.02	0.05

Secondly, we conducted experiments using the second-order stacking algorithm, and the result is shown in **Table 5.**

Table 5 The result of the second-order stacking algorithm

MAE (training)	MAE (testing)
0.018	0.029

From the experimental results, it is evident that the performance of the second-order stacking algorithm is superior to the control model. Therefore, it is chosen as the final employee performance prediction model.

5 Conclusions

In this paper, we employ a second-order stacking algorithm to establish an employee performance prediction model after data processing, effectively reducing the prediction errors and the risk of overfitting. However, due to the complexity of the second-order stacking algorithm, its running time is relatively long, and further improvements are still required.

References

[1] B.M. Khan, S.B. Ali, and S. Naimatullah. " Predictive power of training design on employee performance: an empirical approach in Pakistan's health sector." International Journal of Productivity and Performance Management. (2022): 3792-3808.

[2] B. Agrawal, and Y. Mandhanya. " Performance Appraisal System: A Predictor for Performance of Employees in Engineering Sector." Training & Development Journal. (2019): 49-54.

[3] M. John, and A. Christopher. " Application of Data Mining Classification in Employee Performance Prediction." International Journal of Computer Applications. (2016): 28-35.

[4] N.A. Jalil, H.J. Hwang , and N.M. Dawi. " Machines Learning Trends, Perspectives and Prospects in Education Sector, " In: Education and Multimedia Technology. (2019): 201-205.

[5] P. Ni, K. Tang, and Z.Y. Wang. " Research on Sentinel-1 sea ice classification based on Stacking integrated machine learning method." Mine Surveying. (2022): 70-77.

[6] D. Wang, H.L. Cheng, W.N. Ding, D.J. Li, and H.N. Liu. " The application of SVR and AdaBoosting algorithms in the interpretation of porosity in fractured carbonate reservoirs." Progress in Geophysics. (2022): 1-13.

[7] C. Qiu, C.H. Zhan, and Q.L. Li. " Development and application of random forest regression soft sensor model for treating domestic wastewater in a sequencing batch reactor, " Scientific Reports. (2023): 9149-9149.

[8] S.M. Robeson, and C.J. Willmott. " Decomposition of the mean absolute error (MAE) into systematic and unsystematic components." PloS one. (2023): e0279774-e0279774. doi: 10.1371/JOURNAL.PONE.0279774