Psychology Research, June 2023, Vol. 13, No. 6, 279-285
doi:10.17265/2159-5542/2023.06.006

Employee Attrition Classification Model Based on Stacking Algorithm

CHEN Yanming
Shantou University
LIN Xinyu
South China Normal University
ZHAN Kunye
Shenzhen University

This paper aims to build an employee attrition classification model based on the Stacking algorithm. Oversampling algorithm is applied to address the issue of data imbalance and the Randomforest feature importance ranking method is used to resolve the overfitting problem after data cleaning and preprocessing. Then, different algorithms are used to establish classification models as control experiments, and R-squared indicators are used to compare. Finally, the Stacking algorithm is used to establish the final classification model. This model has practical and significant implications for both human resource management and employee attrition analysis.

Keywords: employee attrition, classification model, machine learning, ensemble learning, oversampling algorithm, Randomforest, stacking algorithm

Introduction

Employee attrition is an important research topic in human resource management, and many scholars have attempted to establish employee attrition classification models using machine learning methods, such as application research of decision tree algorithm in talent attrition of logistics enterprises (Yang & Li, 2017), employee attrition prediction based on database knowledge discovery (Wu, 2019), research on countermeasures for the problem of core employee attrition in state-owned manufacturing enterprises (Qian, 2021), and the application of machine learning in the field of human resource management (Huang, 2022). However, these widely used algorithms have not been able to improve the accuracy of classification models.

This paper uses the stacking algorithm to integrate multiple models and establish the final employee attrition classification model, which achieves an accuracy close to 1.0 and performs better than other ensemble learning algorithms in the control experiment. This model can be applied to the field of human resource management to help managers better classify employee attrition.

CHEN Yanming, Bachelor degree in Economics, Finance, Department of Applied Economics, Shantou University, Shantou, China.
LIN Xinyu, Bachelor degree in Management, Human Resource Management, Department of Management, South China Normal University, Guangzhou, China.
ZHAN Kunye, Bachelor of Management, Management Science and Engineering, Department of Management Science, Shenzhen University, Shenzhen, China.

Theoretical Foundation

Machine learning achieves specific tasks by allowing computers to learn from data and automatically adjust algorithms (Jalil, Hwang, & Dawi, 2019). Machine learning algorithms can classify, predict, cluster, or optimize based on the features and goals of the data.

Ensemble learning is an algorithm that combines multiple basic machine learning models into a more powerful model. It can improve the accuracy and robustness of the model by weighting or voting the output of multiple models. Ensemble learning can improve the generalization ability of the model by reducing variance and bias.

Materials and Methods

Dataset Used in the Study

To investigate and identify the significant factors responsible for employee attrition, and develop a model for classifying employee attrition, this paper employs a dataset created by IBM data scientists available on Kaggle.com. This dataset comprises 1,470 observations, each of which represents whether an employee has resigned or not, and also includes various information about the employee. However, this dataset is considered an imbalanced dataset, because it only consists of 237 resignations and 1,233 non-resignations.

Based on the information of the employee, an employee attrition classification model can be established. There are 32 variables in this dataset. We eliminated some variables that were not pertinent to our research, leaving us with 30 variables that were ultimately used. Afterward, the variable named “Attrition” serves as the dependent variable, and the other 29 variables are used as independent variables. The independent variables in this dataset contain both numerical and categorical variables, and the distribution of the numerical and categorical variables is shown in Table 1.

Table 1

The Proportion of Two Categories of Variables

Type	Number	Proportion
Categorical	6	20.69%
Numerical	23	79.31%

The partial information of numerical and categorical variables is shown in Tables 2 and 3, respectively.

Table 2

Partial Information About Numerical Variables

	Count	Min	Max	Mean
Age	1,470	18.00	60.00	36.92
Daily rate	1,470	102.00	1499.00	802.49
Distance from home	1,470	1.00	29.00	9.19
Education	1,470	1.00	5.00	2.91
Satisfaction	1,470	1.00	4.00	2.72
Performance rating	1,470	3.00	4.00	3.15
Total working years	1,470	0.00	40.00	11.28
Job involvement	1,470	1.00	4.00	2.73
Job level	1,470	1.00	5.00	2.06

Table 3

Rudimentary Information About Categorical Variables

	Count	Unique	Top	Freq
Business travel	1,470	3	Travel rarely	1043
Department	1,470	3	Research & development	961
Education field	1,470	6	Life sciences	606
Gender	1,470	2	Male	882
Job role	1,470	9	Sales executive	326
Marital status	1,470	3	Married	673

Methods

In this paper, variables are first divided into numerical variables and categorical variables. After scaling the numerical variables and transforming the categorical variables into dummy variables, an oversampling algorithm was employed to address the issue of data imbalance. Subsequently, feature selection was performed using the Point-biserial algorithm and Randomforest feature importance ranking method. Finally, a Stacking algorithm was utilized to integrate multiple models and establish the ultimate employee attrition classification model.

Data cleaning and preprocessing. The dataset under consideration is devoid of missing values and outliers. In order to mitigate the issue of overfitting, this paper utilizes the "min-max rescaling" algorithm to scale numerical variables between 0 and 1. The formula for "min-max rescaling" is as follows Equation (1):

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

where X is the original value, $X_{\max}$ and $X_{\min}$ are the maximum and minimum values, and X' is the transformed feature value.

In this paper, all categorical variables are converted into dummy variables using one-hot encoding, which are 0-1 variables.

Oversampling algorithm. In this dataset, the dependent variable "Attrition" is represented by "1" for those who have left resigned and "0" for those who have not. However, the proportion of "1" and "0" is severely imbalanced, which may lead to significant errors if a classification model is directly established. In order to address this issue, this paper has adopted the random oversampling algorithm, which is the quickest and simplest method (Wang & Liu, 2020). The data before and after processing is illustrated in Figure 1.

Another commonly used algorithm is SMOTE oversampling. However, through comparison, we found that the random oversampling method yields better results for this dataset. This is due to the fact that the SMOTE algorithm can not effectively address the data distribution issue in imbalanced datasets, which may result in marginalization of the data distribution and increase the difficulty of classification algorithms to accurately classify the data.

Point-biserial algorithm for feature analysis. This paper conducts Point-biserial correlation analysis on all variables and the target variable "attrition". In Point-biserial correlation analysis, the correlation represents the relationship between variables, while the *p*-value indicates the significance level. Variables with a *p*-value more than 0.05 are usually considered to have no significant correlation with the target variable. The variables with a *p*-value more than 0.05 are shown in Table 4.

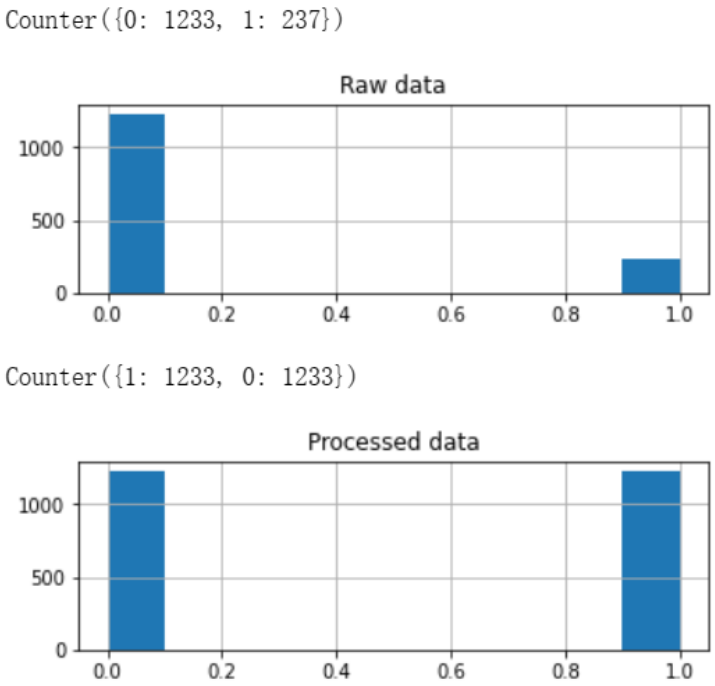

Figure 1. The data before and after oversampling.

Table 4

The Variables With a p-Value More Than 0.05 in Point-Biserial Analysis

Variables	Correlation	*p*-value
Manufacturing director	0.030	0.137
Performance rating	0.016	0.409
Percent salary hike	-0.014	0.472
Research scientist	0.010	0.617
Sales executive	0.006	0.734
Hourly rate	-0.003	0.843

Randomforest feature importance ranking for feature filtering. The point-biserial correlation algorithm can only be employed for preliminary analysis of the correlation between each feature and employee attrition. However, when building a classification model, it is imperative to also consider the intercorrelation among features. Therefore, this paper employs the Randomforest feature importance ranking technique to select the required features for modeling purposes (Wu & Zhang, 2021).

The specific process is as follows: Initially, the dataset is divided into training and testing sets in a 7:3 ratio. Subsequently, all processed variables are inputted into a Randomforest classification model. Then, a Randomforest feature importance ranking can be generated, where the sum of the importance of all features is equal to 1. Finally, features with a feature importance of less than 0.005 are filtered out. This process is shown in Figure 2.

The partial feature importance ranking generated by the Randomforest classification model is shown in Table 5.

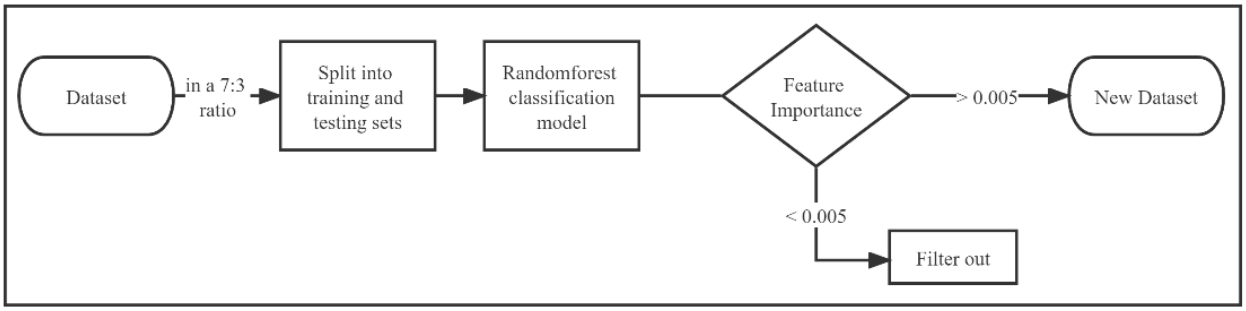

Figure 2. Process of feature filtering.

Table 5

The Top Eight Features of the Adaboosting Feature Importance Ranking

Feature	Importance
Monthly income	0.068
Age	0.063
Monthly rate	0.051
Years at company	0.048
Distance from home	0.047
Total working years	0.045
Percent salary hike	0.042
Satisfaction	0.041

Model building. The stacking algorithm is a non-linear ensemble process, which involves performing cross-validation (K-fold validation) on each base learner (first-layer model), and training a meta-learner (second-layer model) using the results of the base learners as features (Ni, Tang, & Wang, 2022). Typically, a relatively simple model is selected as the second-layer model. The main process of Stacking algorithm is shown in Figure 3.

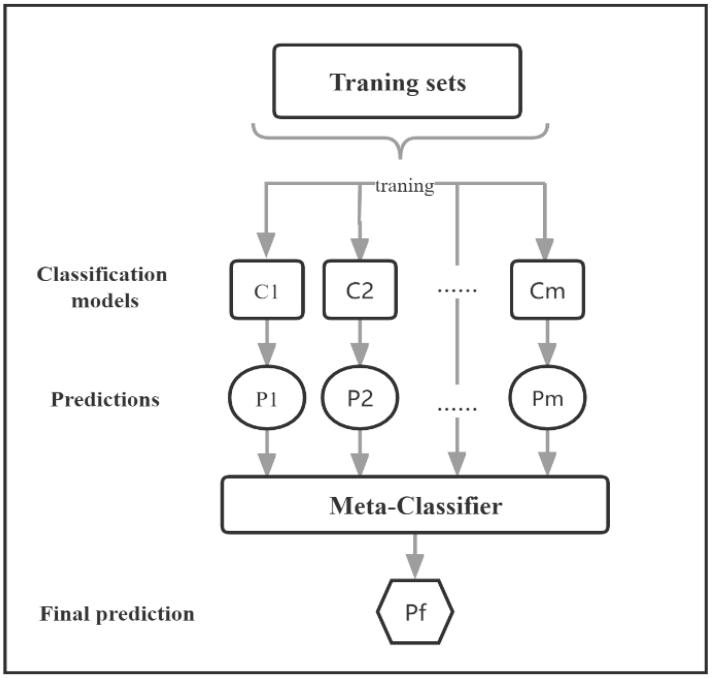

Figure 3. The main process of Stacking algorithm.

This paper employs random forest and Adaboosting as the control experiments, which are two ensemble learning algorithms. Then, we employ a stacking ensemble approach, with random forest classification model, KNN model, Adaboosting classification model, decision tree model, extreme decision tree model, and logistic regression model as the first-layer models. For the second-layer model, we select a decision tree model.

Experiments & Results

Experiment Environment

The dataset comes from a public database named kaggle.com. This experiment was done in python 3.8.0, and the configuration of the computer is shown in Table 6.

Table 6

The Configuration of the Computer

Hardware	Hardware model
CPU	Intel core i7 CPU 2.90 GHZ
RAM	40.0 GB

Experiments and Results

Firstly, comparative experiments are conducted using random forest classification model and Adaboosting model. The random forest classification model is based on decision tree as the base model, and we select the support vector machine (SVM) model as the base model for the Adaboosting model. The results are shown in Table 7.

Table 7

Experimental Results of Two Classification Models

	Training set (accuracy)	Testing set (accuracy)	Testing set (F1-score)
Randomforest	1.0	0.985	0.984
Adaboosting	1.0	0.982	0.982

Secondly, we conducted experiments using the Stacking algorithm, and the result is shown in Table 8.

Table 8

Experimental Result of the Stacking Algorithm

	Training set (accuracy)	Testing set (accuracy)	Testing set (F1-score)
Stacking	1.0	0.998	0.998

The ROC curve of the stacking model is shown in Figure 4.

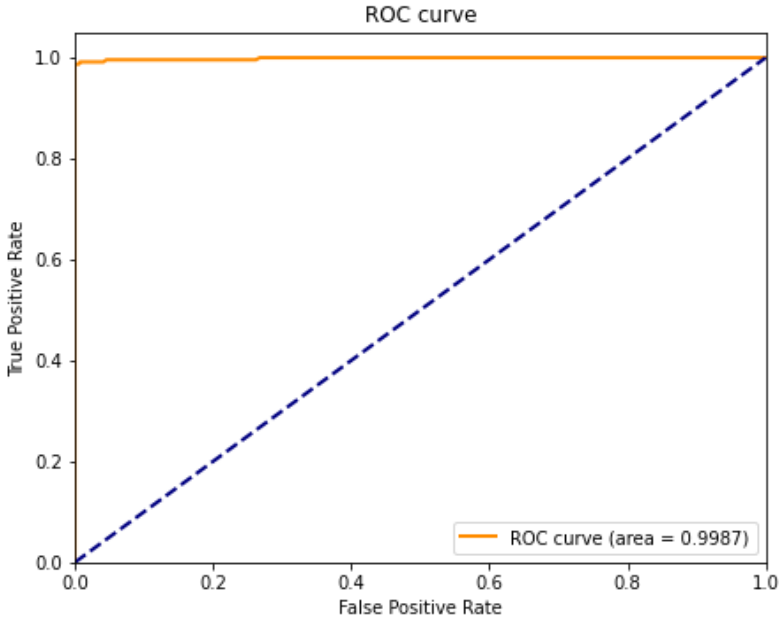

Figure 4. The ROC curve of the stacking model.

From the experimental results, the stacking algorithm can improve the accuracy and F1 score and reduce the risk of overfitting to a certain extent on this dataset.

Conclusions

In this paper, we use oversampling algorithm to address the issue of data imbalance, and employe Stacking algorithm to establish the final employee attrition classification model. Through experimental comparisons, we find that it demonstrated better performance on both the training set and the testing set. However, there are still some flaws. When using the Stacking algorithm, due to the numerous models involved, the process of tuning the model parameters may become more challenging when dealing with more complex datasets.

References

Huang, H. L. (2022). The application of machine learning in the field of human resource management. *Human Resources Development, 23*, 92-93.

Jalil, N. A., Hwang, H. J., & Dawi, N. M. (2019). Machines learning trends, perspectives and prospects in education sector. In *Proceedings of the 2019 3rd international conference on education and multimedia technology* (pp. 201-205). doi:10.1145/3345120.3345147

Ni, P., Tang, K., & Wang, Z. Y. (2022). Research on sentinel-1 sea ice classification based on stacking integrated machine learning method. *Mine Surveying, 1*, 70-77.

Qian, J. (2021). Research on countermeasures for the problem of core employee attrition in state-owned manufacturing enterprises. *China Market, 32*, 19-21.

Wang, D., & Liu, Y. (2020). Denoise-based over-sampling for imbalanced data classification. In *Proceedings of 2020 19th international symposium on distributed computing and applications for business engineering and science* (DCABES 2020).

Wu, D. (2019). Employee attrition prediction based on database knowledge discovery. *Science and Technology & Innovation, 14*, 16-19.

Wu, W. J., & Zhang, J. X. (2021). Feature selection algorithm of random forest based on fusion of classification information and its application. *Computer Engineering and Applications, 57*, 147-156.

Yang, J., & Li, Y. H. (2017). Application research of decision tree algorithm in talent attrition of logistics enterprises. *Logistics Engineering and Management, 8*, 154-156.