Application of Stacking Ensemble Learning Models in Student Academic Performance Prediction and Classification

Authors: Ye Zihao, He Shuangtai, Ye Zihao

Date: 2024-08-06T00:00:00+00:00

Abstract

For the binary classification problem of student academic early warning, this paper proposes an ensemble learning model based on Stacking. The model incorporates students' family background, academic performance, and environmental factors. This study employs BernoulliNB and HistGradientBoost to construct base learners, with Logistic regression as the meta-learner for ensemble classification, and compares it with other baseline models. Experimental results demonstrate that the model achieves an accuracy of 93.9%, which is 7.99%, 6.61%, 2.20%, and 2.47% higher than the K-nearest neighbors algorithm, decision tree model, random forest, and LightGBM model, respectively; and the F1-score reaches 0.95, indicating the model's high application value in student academic performance prediction.

Full Text

Application of Stacking Ensemble Learning Model in Student Academic Prediction and Classification

Ye Zihao[1],*, He Shuangtai[2] ([1]School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai 200101, China)

Abstract

Addressing the binary classification problem of student academic early warning, this paper proposes a Stacking-based ensemble learning model that incorporates students' family background, academic performance, and environmental factors. The model employs Bernoulli Naive Bayes and HistGradientBoosting as base learners, with Logistic regression serving as the meta-learner for integrated classification, and compares its performance against other benchmark

models. Experimental results demonstrate that the model achieves an accuracy of 93.9%, surpassing K-nearest neighbors, decision tree, random forest, and LightGBM models by 7.99%, 6.61%, 2.20%, and 2.47%, respectively. Furthermore, the F1 score reaches 0.95, indicating high application value for student academic prediction.

Keywords: Stacking Ensemble Learning; Educational Data Mining; Academic Prediction

1. Introduction

The primary objective of educational institutions is to equip students with comprehensive knowledge and the ability to apply that knowledge to solve real-world problems. With the rapid development of internet technology, educational institutions at all levels have established information platforms that enable teachers to assign and grade homework, administer examinations, and compile statistics; allow students to conveniently submit assignments, organize error logs, and consolidate their learning; and empower administrators to quantify the workload and progress of both teachers and students. Through data exchange with other institutions and mining of platform data, schools can develop more suitable teaching plans for their faculty and student body. As each cohort of students graduates and new cohorts enroll, these platforms accumulate vast amounts of data related to both teachers and students. The rational utilization of this data's latent information can effectively improve student academic performance and teaching quality.

2.1 Stacking Principle

Stacking (Stacked Generalization) is an ensemble learning technique that combines multiple models. First, the original dataset is partitioned through five-fold cross-validation to train several diverse base learner models. Subsequently, the outputs from these base learners serve as a new training set for training a meta-learner model, with the final prediction result being the output of this meta-model. The five-fold cross-validation process is illustrated below.

[Figure 1: see original paper]

The Stacking model can incorporate various types of models, including linear, non-linear, and tree-based models, thereby leveraging the strengths and characteristics of different models to enhance prediction performance. Additionally, the predictions from multiple models can provide more explanatory information in the final stage, resulting in better interpretability.

2.2 Construction of Base Learners and Meta-Learner

In Stacking ensemble models, base learners and meta-learners are organized in multiple layers, where the input of each subsequent layer is determined by the

output of the previous layer. However, an excessive number of layers increases model complexity, reduces interpretability, and leads to prohibitively long training times. Therefore, this study employs a single layer of base learners and meta-learners.

To enhance model generalization capability, we utilize two base models: Bernoulli Naive Bayes and Histogram-based Gradient Boosting. Bernoulli Naive Bayes is a binomial distribution classifier based on Bayes' theorem, applicable when input data follows a binomial distribution and suitable exclusively for binary classification problems. In practical applications, Bernoulli Naive Bayes classification is also widely employed in neural network research. Histogram-based Gradient Boosting is a novel gradient boosting tree algorithm proposed by the scikit-learn team in 2020. Compared with traditional gradient boosting trees, it demonstrates superior performance when handling high-dimensional, large-scale discrete data, supports adaptive processing of missing values, and exhibits robust characteristics. For the meta-learner, we adopt a relatively simple Logistic regression model to prevent overfitting. During the training process of the Stacking ensemble learning model, we employ the GridSearchCV algorithm to select optimal parameters for both base learners and the meta-learner. The overall framework of the model is shown below.

[Figure 2: see original paper]

3. Data Processing and Feature Engineering

Feature engineering refers to the process of extracting experimental features from raw datasets, where these features can effectively represent the data content. The primary steps include data preprocessing, feature analysis, and feature transformation.

3.1 Dataset Introduction

The dataset utilized in this study is sourced from the open data platform Kaggle, titled "Predict students' dropout and academic success." This dataset provides comprehensive information about undergraduate students in higher education institutions, encompassing 35 features related to student basic information, campus learning behavior, and local economic characteristics. These data facilitate in-depth understanding of factors influencing student academic completion and exhibit high authenticity and reliability.

3.2 Data Preprocessing

Data preprocessing primarily includes data cleaning, data integration, and data transformation. Data cleaning involves handling missing values, outliers, irrelevant values, noise, and duplicate data. Data integration consolidates data from different sources into a unified database table. Data transformation compresses or expands extremely large or small numerical data to enable models to

better learn data characteristics. This study first performs normalization and standardization scaling on the data.

3.3 Pearson Correlation Calculation and Feature Transformation

Before model training, it is common practice to screen and select dataset features by removing low-correlation and redundant features, thereby reducing model training complexity and enhancing interpretability. Therefore, this study employs the Pearson correlation coefficient to calculate feature associations. The Pearson correlation coefficient is a statistical method for measuring linear correlation between two variables, with a value range of [-1, 1]. A positive coefficient indicates positive correlation, a negative coefficient indicates negative correlation, and a coefficient of zero indicates no linear correlation. Evidently, certain student features in the raw dataset exhibit high correlation and require merging. The correlation heatmaps of dataset features are shown in the figures.

[Figure 3: see original paper]
[Figure 4: see original paper]

As shown in Figure 3, the similarity between parents' qualification and occupation information (Father's/Mother's qualification/occupation) is relatively high. Figures 3 and 4 reveal redundancy among various attributes of students' two-year learning behavior (Curricular 1st/2nd sem units). Consequently, this study merges these feature information through summation and proposes new feature columns, `family_{{{edu}}_{{background}}}` and `Curricular units`, to preserve this information.

3.4 Categorical Feature Description

Based on the aforementioned feature transformations, this study categorizes data features into three major categories: student basic information, campus learning behavior, and local economic characteristics. Specific feature information is detailed in the tables below.

4. Experiments and Results Analysis

The experiments were conducted on a computer with an Intel i7-10750H processor and 40GB RAM, running a 64-bit Windows 10 operating system. The model was constructed using Jupyter Notebook with Python 3.10, employing libraries including pandas, numpy, matplotlib, seaborn, and scikit-learn. The dataset contains 76,518 records, randomly split with 90% allocated as the training set and the remaining 10% as the test set for model validation.

4.1 Model Generalization Ability Evaluation Metrics

The research objective is to predict whether students can successfully graduate based on their relevant features, classifying predictions into two categories: graduation (P) and dropout (N), constituting a standard binary classification

problem. The confusion matrix for this binary classification prediction problem is presented below.

This study employs five commonly used evaluation metrics to assess model performance: Accuracy, Recall, Precision, F1 score, and AUC score. Accuracy represents the proportion of correctly predicted samples to total samples. Recall, also known as sensitivity, geometrically represents the proportion of actual positive cases correctly identified by the model. Precision refers to the proportion of correctly predicted positive cases among all predicted positive cases. The F1 score is the harmonic mean of recall and precision, frequently applied in imbalanced data scenarios as a comprehensive performance indicator. AUC (Area Under the Curve) refers to the area under the ROC (Receiver Operating Characteristic Curve), with values ranging from 0 to 1. When AUC equals 0.5, the model performs equivalently to random guessing; values closer to 1 indicate stronger ability to distinguish between positive and negative classes. Although AUC is unaffected by class imbalance, it is only applicable for evaluating binary classification problems.

4.2 Model Hyperparameter Settings

Model hyperparameters are parameters set before the learning process begins. Researchers assign specific values to these parameters to enhance model learning and performance. Common parameter tuning methods include GridSearchCV and RandomizedSearchCV. In recent years, population optimization algorithms with shorter runtime and better convergence, such as the Sparrow Search Algorithm (SSA) and Mantis Optimization Algorithm, have gained significant attention in deep learning research. This study utilizes the GridSearchCV method for parameter tuning. The optimal hyperparameters for each learner are detailed in the table below.

4.3 Model Generalization Ability Analysis

Ensemble learning combines the strengths of multiple learners through specific strategies, leveraging their respective advantages to improve classifier performance. Therefore, the integrated classifier should outperform individual classifiers. To verify the prediction performance of the proposed Stacking ensemble model, this study compares it against KNN, SVC, Decision Tree, Gradient Boosting, and LightGBM on the dataset to demonstrate the superiority of our design. Finally, based on the metrics defined in Section 4.1, we compare the accuracy, recall, precision, F1 score, and AUC score of each classifier, as illustrated in the figure.

[Figure 5: see original paper]

5. Conclusion

As an emerging field in recent years, educational data mining aims to deeply explore student academic performance and capabilities. This study leverages student educational data to propose a Stacking ensemble learning model. The model employs Bernoulli Naive Bayes and Histogram-based Gradient Boosting as base learners, with Logistic regression as the meta-learner. Parameter tuning via grid search yields satisfactory results. The findings demonstrate that the proposed model achieves an accuracy of 93.9% and an F1 score of 0.95, surpassing other advanced methods. This research not only contributes to the further development of educational data mining but also provides valuable reference for university faculty in improving teaching quality and students' comprehensive literacy.

References

[1] Ye Zihao, Zhang Xiaojiao, Xu Guanglin. Research on teaching reform of "Dynamic Website Technology" course[J]. Curriculum Education Research, 2021(34): 26-27.
[2] Li Yufan, Zhang Huifu, Liu Shangli, et al. Research progress on educational data mining[J]. Computer Engineering and Applications, 2019, 55(14): 15-23.
[3] Su Yingfang. Application of educational data mining technology in university curriculum construction[J]. Office Automation, 2024, 29(4): 55-58.
[4] Chen Quanji, Niu Yanmin. Research on learning situation data analysis based on educational data mining[J]. Journal of Bingtuan Education Institute, 2023, 33(3): 62-67, 80.
[5] You Pu, Liu Xingfu. Application of Stacking ensemble learning in sales forecasting[J]. Software Guide, 2022, 21(4): 103-108.
[6] Lin Ping, Lü Jianchao. Research on identification of information adoption in online health community Q&A based on Stacking ensemble learning[J]. Information Science, 2023, 41(2): 135-142.
[7] Zou Zelin, Cheng Xia, Liu Ziwei, et al. Application of Stacking ensemble learning model in forest stock volume estimation[J]. Hubei Forestry Science and Technology, 2023, 52(4): 52-57.
[8] Wang Hui, Li Changgang. Application of Stacking ensemble learning method in sales forecasting[J]. Computer Applications and Software, 2020, 37(8): 85-90.
[9] Zhang Liu, Chen Yifei, Yuan Jiawei, et al. Application of Stacking ensemble learning model in hybrid grade classification prediction[J]. Computer Systems & Applications, 2022, 31(7): 325-332.
[10] Sun Hao, Pan Hong. Bank customer product subscription prediction based on Stacking ensemble model[J]. Artificial Intelligence, 2023(6): 74-81.

Author Contributions

Ye Zihao, He Shuangtai: Conceived the research idea and revised the final manuscript.

Ye Zihao: Designed the research protocol, collected and cleaned data, performed analysis and experiments, and drafted the manuscript.

Corresponding Author: Ye Zihao (E-mail: zihao_{ye}@qq.com)

Note: Figure translations are in progress. See original paper for figures.