ISSN 2616-5775 Vol. 7, Issue 8: 74-78, DOI: 10.25236/AJCIS.2024.070811

Prediction and Analysis of University Student Performance Based on Machine Learning

Xiaobo Zhu*

School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China
**Corresponding author: zhuxb@cqupt.edu.cn*

Abstract: *The prediction of student performance is an important research direction of educational data mining. However, current predictive models often fail to sufficiently reflect the learning process and lack a quantitative analysis of the impact of different predictors. In this study, we collected the learning data of 395 students enrolled in the Python Programming course offered by Chongqing University of Posts and Telecommunications, and developed a model based on the random forest algorithm to predict and analyze student performance. Evaluation results indicate that the model can accurately predict student performance (R^2 = 0.55, RMSE = 4.68 points and MAE = 3.61 points). Moreover, an importance analysis of the predictors revealed that the score of the 2nd unit test had the greatest impact on student performance, followed by the score of the 4th unit test and the average score of homework. This study provides guidance for early warning of students' learning difficulties and for continuous curriculum enhancement, thereby serving as a reference for enhancing teaching quality.*

Keywords: *Student Performance Prediction; Random Forest; Feature Importance*

1. Introduction

Under the background of the big data era, education and information technology are increasingly integrated, giving rise to a data-driven education model centered on big data processing and artificial intelligence. In this context, educational data mining aims to analyze and mine large volumes of educational data to understand students' learning conditions, teaching quality, and the teaching environment, thereby formulating effective educational policies and plans [1]. Compared to traditional experiential teaching management methods, educational data mining can quickly uncover hidden patterns within data, providing university teachers with more efficient and accurate references for their teaching activities.

Educational data mining has been widely used in areas such as student learning behavior analysis, student performance prediction, educational resource optimization analysis, and teacher-student interaction behavior analysis. Among these, student performance prediction is an important research direction. It aims to predict student performance by mining relevant data, thereby providing decision support for students and educational administrators [2]. Currently, most studies employ two main methods for predicting student performance. One method is based on the recommendation system approach. For example, Jembere et al. proposed that the performance prediction problem is essentially a matrix completion problem and utilized matrix factorization to solve the prediction problem [3]; Ren et al. employed a neural collaborative filtering model to predict grades [4]. The second method involves machine learning techniques. Zhang et al. systematically reviewed non-process data features and process data features used in performance prediction, outlining the construction of student performance prediction models based on machine learning [5]. Li et al. introduced a method for predicting student performance using student behavior data and optimized their model through Stacking integration [6]. Liu et al. utilized a multilayer feature fusion method to enhance the feature representation of student performance information [7].

In summary, numerous data mining studies have been conducted on student performance prediction, which holds significant value in assisting teaching management. However, current research still exhibits two main shortcomings. On one hand, many aspects of student learning that significantly impact performance are inadequately represented in predictive models. On the other hand, there is a lack of quantitative analysis regarding the impact of predictors, leading to limited model interpretability. Therefore, this study aims to integrate multiple indicators of student learning processes, utilize machine

ISSN 2616-5775 Vol. 7, Issue 8: 74-78, DOI: 10.25236/AJCIS.2024.070811

learning methods to predict university students' course performance, and perform feature importance analysis on predictors to enhance the interpretability of the prediction model.

2. Model Development and Results Analysis

2.1. Data Sources

The data used in this study were obtained from the Python Programming course offered by Chongqing University of Posts and Telecommunications. The course utilizes several information platforms for long-term course construction. From these platforms, we collected learning data from 395 students, including course grades, review and preview activities, class attendance records, homework submissions, and unit test results. Subsequently, these data were preprocessed, which included format standardization, checking for missing values, and filling them as necessary.

2.2. Selection of Predictors

Eight predictors from four learning sessions were selected in this study (Table 1). The correctness of review and preview exercises was chosen for the review and preview session, while attendance rate was selected for the classroom learning session. For the homework session, the predictors included the average score and submission rate of homework. For the classroom test session, the predictors comprised the scores of four unit tests. The content of these unit tests includes basic data types and programming control structures, comprehensive applications of data types and programming control structures, function design and calling, and file operations.

Table 1: Predictors for student performance prediction

Learning sessions	Predictors
Review and preview	Correctness of review and preview exercises
Classroom learning	Attendance rate
Classroom test	Score of the 1st unit test
	Score of the 2nd unit test
	Score of the 3rd unit test
	Score of the 4th unit test
Homework	Average score of homework
	Submission rate of homework

2.3. Model Structure

In this study, the random forest (RF) algorithm was selected for constructing the prediction model. RF is an ensemble learning algorithm based on decision trees and the bagging method [8]. For regression problems, the fundamental component of RF is the regression tree, which is a decision tree designed for continuous target variables. In the bagging ensemble, RF employs the bootstrap sampling method to draw multiple samples from the input data with replacement. Subsequently, it constructs regression trees on each of these samples. These regression trees are independent of each other, and at each node, a subset of features is randomly selected to determine the optimal feature set for branching. The predictions from all trees are averaged to produce final predictions.

2.4. Model Training and Evaluation

The most critical hyperparameters of RF are the number of regression trees and the maximum depth of each regression tree. In this study, we employed the grid-search method to determine these hyperparameters. The grid-search method is an exhaustive method that iterates over all combinations within a given range of hyperparameters and determines the final parameter set based on comparing prediction accuracies across different combinations. We adopted a 10-fold cross-validation strategy for model validation. In the 10-fold cross-validation, the dataset was randomly divided into 10 equal-sized folds. During each training iteration, one fold was used as the validation set, and the remaining nine folds served as training data. This process was repeated 10 times to ensure comprehensive validation. Finally, all the predicted values from all 10 folds were compared with the observed values. The coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE) were selected as evaluation metrics. R^2 measures the overall prediction ability of the model, RMSE measures the overall

ISSN 2616-5775 Vol. 7, Issue 8: 74-78, DOI: 10.25236/AJCIS.2024.070811

quality of the prediction, and MAE measures the average bias of the prediction [9]. These metrics were calculated as follows:

$$R^2 = \left(\frac{\sum_{i=1}^{n} \left(O_i - \bar{O}\right)\left(P_i - \bar{P}\right)}{\sqrt{\sum_{i=1}^{n} \left(O_i - \bar{O}\right)^2} \sqrt{\sum_{i=1}^{n} \left(P_i - \bar{P}\right)^2}} \right)^2 \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(O_i - P_i\right)^2} \tag{2}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left|O_i - P_i\right| \tag{3}$$

where O_i is the ith observed value, P_i is the ith predicted value, $\bar{O}$ and $\bar{P}$ are the means of the observed and predicted scores respectively; and n is the number of observation samples.

Finally, the optimal number of regression trees was determined to be 55, with a maximum depth of 6 for each regression tree. The results of the cross-validation show that most predicted and observed scores are distributed around the 1:1 solid line, with R^2 = 0.55, RMSE = 4.68 points and MAE = 3.61 points, indicating that the RF model developed in this study can accurately predict student performance (Figure 1).

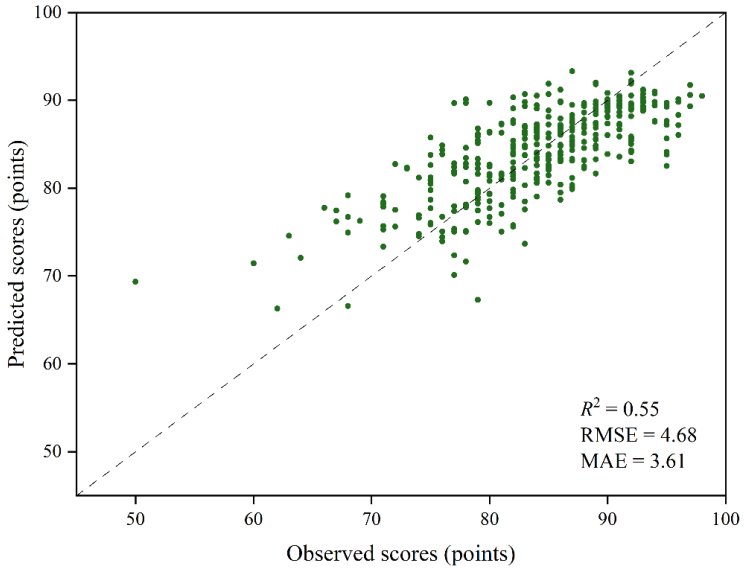

Figure 1: Comparison of predicted and observed scores

2.5. Predictor Importance Analysis

Through quantitative analysis of predictor effects on student performance, this study provides insights for student learning and teacher guidance. Feature importance in random forests relies on the Gini index to measure the contribution of each features to the improvement of prediction accuracy, as utilized here to assess predictors' impact on student performance. The results show that the most important predictor for student performance prediction is the score of the 2nd unit test (31.78%), followed by the score of the 4th unit test (16.95%), average score of homework (15.85%), and the score of the 1st unit test (14.17%). The correctness of review and preview exercises (10.02%) and the score of the 3rd unit test (9.74%) also contribute significantly. Attendance rate (0.80%) and submission rate of homework (0.69%) exhibit comparatively lower importance (Figure 2).

The content of the 2nd unit test covers the comprehensive application of data types and programming control structures, foundational to Python language mastery and subsequent learning, thereby significantly impacting student performance. The content of the 4th unit test is file operation, which is one of the difficult parts of the Python language content. The average score of homework reflects the

ISSN 2616-5775 Vol. 7, Issue 8: 74-78, DOI: 10.25236/AJCIS.2024.070811

students' ability to master the knowledge on a daily basis. Notably, the correctness of review and preview exercises proves influential, underscoring the importance of these activities. The limited impact of attendance rate and homework submission rate may stem from their consistently high levels, which fail to distinguish individual student performance differences.

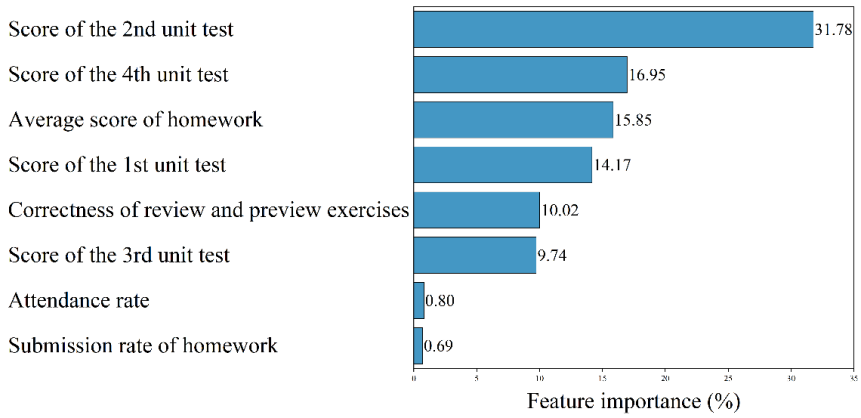

Figure 2: Feature importance of predictors for student performance prediction

3. Limitations and Prospects

In this study, we developed a student performance prediction and analysis model based on RF. Using this model, we accurately predicted student performance and analyzed the impact of each predictor. However, several limitations exist. Firstly, the course has only been offered since 2020, resulting in a relatively short data collection period encompassing only 395 students. This may not fully capture the overall learning dynamics, rendering the data less representative. Secondly, certain indicators reflecting student learning processes, such as classroom interaction frequency and homework completion time, were not incorporated into the predictors. Future studies will address these shortcomings to enhance comprehensiveness.

The model developed here holds promising applications. For instance, traditional teaching methods typically analyze and alert students after the results are announced, which introduces delays [10]. Leveraging our model, student performance can be forecasted before final examinations, enabling targeted and timely interventions to encourage focused preparation. Additionally, continuous curriculum enhancement is pivotal for advancing educational standards and student training quality. The predictor analysis in our study offers vital insights to guide ongoing curriculum improvements.

4. Conclusion

Student performance prediction and analysis represent pivotal research in the convergence of artificial intelligence and education. Applying machine learning to predict student performance holds substantial implications for optimizing educational management. This study focuses on the Python Programming course at Chongqing University of Posts and Telecommunications, where we collected data from 395 students to develop a student performance prediction and analysis model using random forest. The evaluation of our model demonstrates its robust ability to predict university student performance accurately. Furthermore, our analysis of predictor importance reveals that the score of the 2nd unit test exerts the strongest influence on student performance, followed by the score of the 4th unit test and the average homework score, among others. This research offers insights for early student learning warnings and course enhancements, thereby contributing significantly to practical and societal advancements.

Acknowledgements

This work was supported by the Education and Teaching Reform Project of Chongqing University of Posts and Telecommunications (No. XJG22110).

ISSN 2616-5775 Vol. 7, Issue 8: 74-78, DOI: 10.25236/AJCIS.2024.070811

References

[1] Zhang Z, Chu Z. Student achievement prediction model based on XGBoost and SHAP feature analysis [J]. Informatization Research, 2024, 50(03): 34-40.
[2] Cao H J, Xie J. LSTM-based learning achievement prediction and its influencing factors [J]. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 2020, 22(06): 90-100.
[3] Jembere E, Rawatlal R, Pillay A W. Matrix factorisation for predicting student performance[C]. In 7th World Engineering Education Forum, 2017: 513-518.
[4] Ren Z, Ning X, Lan A S, et al. Grade Prediction with Neural Collaborative Filtering[C]. In 2019 IEEE International Conference on Data Science and Advanced Analytics, 2019: 1-10.
[5] Zhang F, Chen J J. Overview of learning data features application to student performance prediction [J]. Software Engineering, 2023, 26(10): 1-4.
[6] Li K W. Grade prediction of university student based on machine learning[J]. Computer Era, 2023, 12: 220-223.
[7] Liu T, Qi H R, Ni W J. Multilayer feature fusion based model for prediction student academic performance[J]. Computer Engineering and Design, 2023, 44(10): 2973-2978.
[8] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.
[9] Yang R M, Zhang G L, Liu F, et al. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem[J]. Ecological Indicators, 2016, 60: 870-878.
[10] Liu B P, Fan T C, Yang H. Research on application of early warning of student's achievement based on data mining [J]. Journal of Sichuan University (Natural Science Edition), 2019, 56(02): 267-272.