Prediction of student online Learning performance based on Machine learning

Yu Chen, Jun Peng

Abstract
This essay uses machine learning methods to predict students’ online learning performance. The characteristics of student’s characteristics, learning behavior characteristics, and environmental factors are extracted by feature engineering, and the random forest model is used for prediction. The experimental results show that the prediction accuracy of the random forest model reaches 0.82, which provides an effective means for the prediction of online learning performance.

Keywords: Machine learning, online learning, student performance prediction, feature engineering, random forest

1 Introduction

1.1 Background and research significance
In recent years, machine learning technology has made remarkable achievements in many fields, and prediction technology has been widely concerned. This study aims to explore the application of machine learning methods to predict students’ online learning performance. The prediction accuracy can be improved by constructing fine feature engineering, selecting a suitable prediction model, and optimizing the model. This research will help promote the development of machine learning technology in the prediction field and provide guidance for various prediction tasks. At the same time, predicting students’ online learning performance will also provide valuable references for online learning platforms and other application scenarios and expand the application range of machine learning technology.

1.2 Overview of relevant work
Many researchers have explored various machine learning methods in student online learning performance prediction. For example, Kotsiantis et al. (2004) used decision trees, naive Bayes, and other methods to analyze students’ learning behaviors and achieved good prediction results. Through linear regression, Macfadyen and Dawson (2010) analyzed the relationship between students’ learning behavior and academic performance. They found that characteristics such as the number of discussions significantly impacted academic performance. Marquez-Vera et al. (2013) predicted students’ learning performance based on a support vector machine (SVM), and the accuracy reached 0.82. However, Pardos and Heffernan (2011) used the random forest method to predict students’ academic performance, which was superior to other models. These studies provide an important reference for this paper’s experimental design and method selection.

At the same time, feature engineering is key in predicting students’ online learning performance. For example, Kovanovi et al. (2015) studied the prediction model based on the characteristics of learning behavior and found that the characteristics such as curriculum participation have higher importance. Tempelaar et al. (2015) discussed the influence of learning environment factors on students’ learning performance, such as network environment and equipment type. Based on these relevant studies, this study will further optimize the feature engineering and model selection to improve the prediction accuracy.

2 Data collection and preprocessing

2.1 Introduction to Online Learning Platform
The online learning platform is an Internet-based educational resource-sharing and learning management system which provides convenient learning and communication environment for students and teachers. The platform provides various course resources covering various subject areas, realizing unlimited learning time and place. In addition, the large amount of data generated during the learning process provides strong support for teachers and researchers to analyze students’ learning behaviors, optimize teaching methods and make personalized recommendations. Using machine learning technology, students’ online learning performance can be predicted, bringing more innovation and improvement to the online education sector.
2.2 Data Collection and preprocessing methods

Data collection: Obtain student data from online learning platforms, including personal information, learning behavior, and interaction records. The details include the following:

- Personal information: age, gender, location, etc.
- Learning behavior: login frequency, learning duration, number of tasks completed, etc.
- Interactive records: number of posts, likes, replies, etc.

Data preprocessing is to clean and organize the collected data to meet the input requirements of the machine learning model, which mainly includes:

- Missing value processing: interpolation or deletion method is used to process missing data;
- Outlier processing: identify and eliminate outliers by boxplot and other methods;
- Data standardization: standardizing data processing to eliminate dimensional influence;
- Category feature coding: The category feature is used for unique heat or label coding.

Example after data preprocessing:

<table>
<thead>
<tr>
<th>No.</th>
<th>Age</th>
<th>Gender</th>
<th>Region</th>
<th>Login Frequency</th>
<th>Learning Duration</th>
<th>No. of tasks completed</th>
<th>No. of posts</th>
<th>No. of likes</th>
<th>No. of replies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>Male</td>
<td>Beijing</td>
<td>5</td>
<td>120</td>
<td>10</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>Female</td>
<td>Shanghai</td>
<td>3</td>
<td>100</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

3 Feature engineering

When predicting students’ online learning performance, we need to extract features from multiple perspectives:

3.1 Personal characteristics of students

Personal characteristics of students are important sources of information for predicting students’ online learning performance. These characteristics include but are not limited to gender, age, family background, geographical location, educational background, learning style, subject interests, and expertise. These characteristics can help us to have a deeper understanding of students’ basic information and learning characteristics. By reasonably introducing students’ personal characteristics into the machine learning model, more useful information about students can be provided to the model, thus improving the accuracy and reliability of the prediction of students’ online learning performance. Meanwhile, by analyzing the relationship between different individual characteristics and learning performance, educators can better understand the needs of students and provide them with personalized learning support.

3.2 Learning behavior characteristics

Characteristics of learning behavior are the core factors in predicting students’ online learning performance. By capturing students’ specific behaviors in online learning platforms, students’ learning performance can be more directly reflected. The characteristics of learning behavior mainly include course visits, learning duration, homework submission, participation in discussion boards, resource downloads, course completion, learning progress, etc. These features help reveal students’ interaction on the platform and reflect students’ learning enthusiasm, participation, and mastery. It is helpful to improve the accuracy of predicting students’ online learning performance by incorporating these characteristics into the analysis of the machine learning model. At the same time, it can provide educators with detailed information about students’ learning behavior to optimize teaching methods and strategies.

3.3 Characteristics of environmental factors

Environmental characteristics are another important aspect affecting students’ online learning performance, which involves the external conditions and background factors students face in the learning process. These characteristics include curriculum design and difficulty, quality of teaching resources, level of teacher-student interaction, cooperation and competition among classmates, quality of network environment, etc. These environmental factors can help us understand the overall teaching quality of online learning platforms and the influence of external conditions on students’ learning performance. In the machine learning model, considering the characteristics of environmental factors can help improve the comprehensiveness of the prediction of students’ online learning performance and provide a reference for educators to improve the teaching environment and improve students’ satisfaction.

3.4 Feature selection method

Feature selection method plays an important role in feature engineering, which can help screen out the most influential features in predicting students’ online learning performance, reduce the complexity of the model and improve the accuracy of the prediction. Commonly used feature selection methods include the filtering method (e.g., Chi-square test, Pearson correlation coefficient),
wrapping method (e.g., recursive feature elimination, forward selection method), and embedding method (e.g., LASSO, ridge regression). The filtering method is selected mainly according to the degree of correlation between the feature and the target variable. The optimal feature subset is determined by subset search and model evaluation. The embedding method automatically selects features in the process of model training. According to the specific problems and data sets, appropriate feature selection methods can be selected to improve the model’s performance for predicting students’ online learning performance.

4 Machine learning model

4.1 Model Selection

In this study, three representative machine learning models, including linear regression, decision tree, and random forest, are adopted to predict students’ online learning performance:

4.1.1 Linear regression:

Linear regression is a simple and easy-to-explain model suitable for the linear relationship between the feature and the target. Linear regression models try to find a linear equation that minimizes the sum of squares of error between the predicted value and the actual value. We can use a linear regression model to explore the degree of influence of different characteristics on students’ online learning performance to provide a reference for optimizing the online learning environment.

4.1.2 Decision Tree:

The decision tree is a model which is easy to understand and can deal with nonlinear relations. The decision tree divides the data set into different subsets through recursive partitioning of features, thus forming a tree-like structure. The decision is made along the tree structure according to the eigenvalue in the prediction, and the final prediction result is obtained. A decision tree in predicting students’ online learning performance can help us determine the relationship between the features and provide educators with targeted suggestions.

4.1.3 Random Forest:

Random forest is an ensemble learning method that reduces overfitting risk and improves generalization ability by constructing multiple decision trees and combining their prediction results. Random forests can capture complex interactions between features to achieve better results in predicting students’ online learning performance.

4.2 Model Training and Optimization

Cross-validation, grid search, and other techniques will be used to train and optimize these models for hyperparameter adjustment. By dividing the data set into a training set and a validation set, the model is trained and evaluated for multiple rounds to evaluate the model’s performance on unknown data. Grid search finds the parameter setting that optimizes model performance by traversing the given parameter combination.

4.3 Model Evaluation and Comparison

By comparing the predictions of each model on the validation set, we can evaluate their performance and select the best-performing model. After selecting the best model, we will use it to predict students’ online learning performance to help educators and online learning platforms better meet students’ needs.

5 Experimental Results and Analysis

5.1 Description of the data set

Based on a data set of 5,000 students, this study conducted an in-depth analysis of students’ characteristics, learning behavior characteristics, and environmental factors characteristics. Here is a description of some of the features in the data set:

Age of students: The age range is 15-19 years old, the mean age is 17.23 years old, and the standard deviation is 1.38 years old.

Gender: In the data set, female students accounted for 48.26%, and male students accounted for 51.74%.

Online learning duration: Students’ online learning duration ranges from 32 to 1217 hours, with an average online learning duration of 519.61 hours and a standard deviation of 291.53 hours.

Number of clicks on learning resources: The number of clicks on learning resources of students ranges from 53 to 3523, the average number of clicks is 1357.42, and the standard deviation is 835.16.

Students’ self-learning ability: The data set contains students’ self-learning ability index on a scale of 1-5 (5 being the strongest). The mean self-learning ability was 3.17 points, and the standard deviation was 1.29 points.

Student participation in class: Student participation in class is reflected in the number of speeches and questions in class. The number of lectures ranged from 1 to 207, the average number of lectures was 46.12, and the standard deviation was 37.98. The number of questions in the class ranged from 0 to 149, the average number of questions was 28.34, and the standard deviation was 21.11.

Peer interaction: The data set also considers interactions between students and their peers, including the number of responses to online discussion boards and the number of offline group studies. The number of replies in online discussion forums ranges from 0 to 413, with an average of 65.73 and a standard deviation of 49.61. Offline group learning times ranged from 0 to 104 times, the average
group learning times were 23.68, and the standard deviation was 15.32.

Test scores: Test score data ranges from 0 to 100, with an average test score of 64.52 and a standard deviation 15.72.

Family economic status: The family economic status index in the data set ranges from 1 to 5 (5 being the most affluent). The mean household economic status was 2.98 points, and the standard deviation was 1.16 points.

Type of school: 30.47% of students in the data set were from public schools, and 69.53% were from private schools.

Learning styles: The data set contains students’ learning styles, classified as visual, auditory, literate, and kinesthetic. Visual-type students accounted for 32.12%, auditory-type students accounted for 28.36%, literacy-type students accounted for 21.84%, and kinesthetic-type students accounted for 17.88%.

Course difficulty: Course difficulty data ranges from 1 to 5 (5 being the most difficult). The mean course difficulty is 3.25, and the standard deviation is 1.03.

Teacher teaching level: The teacher teaching level index in the data set ranges from 1 to 5 points (5 points indicates the best). The average teacher teaching level was 3.48 points, and the standard deviation was 1.21 points.

Network environment: Students’ network environment data can be divided into four grades: excellent, good, medium, and poor. Among them, excellent accounts for 31.62%, good accounts for 42.73%, medium accounts for 18.84%, and poor accounts for 8.31%.

Device usage: The data set records the types of devices students use, including PCS, tablets, and mobile phones. PC usage was 55.38%, tablet usage 26.16%, and mobile usage 18.46%.

Daily learning time: Students’ daily learning time ranges from 0.5 to 6 hours, the average daily learning time is 2.87 hours, and the standard deviation is 1.23 hours.

Study days per week: Students’ study days per week range from 1 to 7 days, the average study days per week is 4.61 days, and the standard deviation is 1.94 days.

By analyzing these characteristics, we can better understand students’ learning status to provide strong support for the subsequent prediction model. At the same time, the detailed description of these features is helpful for more accurate feature selection and processing to improve the model’s prediction accuracy.

5.2 Feature significance analysis

In this study, filtering methods (e.g., Chi-square test, Pearson correlation coefficient), wrapping methods (e.g., recursive feature elimination, forward selection), and embedding methods (e.g., LASSO, Ridge regression) were used to analyze the importance of various features. The specific results are as follows:

Filtering method: Through the chi-square test and Pearson correlation coefficient calculation, the feature importance ranking is as follows: Daily learning duration (0.45), weekly learning days (0.42), course difficulty (0.36), learning behavior data (0.33), equipment usage (0.27), network environment (0.25), learning style (0.23), teacher’s teaching level (0.21), student’s characteristics (0.18).

Wrapping method: Using recursive feature elimination and forward selection method, the feature importance ranking is as follows: Daily learning duration (0.48), course difficulty (0.41), learning behavior data (0.37), weekly learning days (0.34), equipment usage (0.29), the network environment (0.25), learning style (0.22), teacher’s teaching level (0.20), and student’s characteristics (0.17).

Embedding method: The feature importance ranking obtained by LASSO and Ridge regression analysis is as follows: Daily learning duration (0.47), course difficulty (0.40), learning behavior data (0.35), weekly learning days (0.33), equipment usage (0.28), the network environment (0.24), learning style (0.22), teacher’s teaching level (0.19), student’s characteristics (0.16).

Combining the results of the three methods, we get the final feature importance ranking: Daily learning duration (0.47), course difficulty (0.39), learning behavior data (0.35), weekly learning days (0.33), equipment usage (0.28), the network environment (0.24), learning style (0.22), teacher’s teaching level (0.20), student’s characteristics (0.17).

Through these analyses, we can find that characteristics such as daily learning time, course difficulty, and learning behavior data are highly important in predicting students’ online learning performance. In contrast, factors such as students’ personal characteristics, learning styles, and teachers’ teaching level are relatively low. This helps us focus on these important features and improve the model’s accuracy when building the prediction model.

5.3 Comparison of prediction results of each model

This study used three representative machine learning models, namely linear regression, decision tree, and random forest, to model important features (such as daily learning time, course difficulty and learning behavior data, etc.) from the previous analysis. The specific modeling process is as follows:

Linear regression: The least square method fits the linear regression model. First, the data set was divided into the training set and test set, with the training set accounting for 70% and the test set accounting for 30%. For linear regression, we need to calculate the weight of each feature to predict students’ online learning performance. The regression coefficients obtained from the training set are as follows: daily learning time (0.342), course difficulty (-0.218), and learning behavior data (0.457). The test
set is predicted after training on the training set, and the model performance is calculated.

Decision tree: According to the characteristics of students’ online learning, results are recursively divided, and finally, a decision tree is generated. First, we use information gain as a feature selection criterion. To avoid overfitting, we set the maximum depth of the tree as six and the minimum number of split samples as 10. The optimal parameter combination was selected through five-fold cross-validation, and the parameters finally selected were: maximum depth four and minimum split sample number 7.

After the decision tree model is trained on the training set, the test set is predicted, and the model performance is calculated.

Random Forest: Random Forest is an integrated learning model with multiple decision trees. We set the number of trees in the forest as 100 and sampled the training data with the bootstrap sampling method. Again, to avoid overfitting, the maximum depth of the tree is set at 6. The optimal parameter combination was selected through the five-fold cross-validation, and the parameters finally selected were: the number of trees was 50, and the maximum depth was 5. After the random forest model is trained on the training set, the test set is predicted, and the model performance is calculated.

The performance of each model on the test set is shown in the following table:

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Recall</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>0.78</td>
<td>0.75</td>
<td>0.76</td>
<td>0.8</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.81</td>
<td>0.78</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.85</td>
<td>0.84</td>
<td>0.84</td>
<td>0.87</td>
</tr>
</tbody>
</table>

As seen from the above table, the performance of the three models on the test set differs greatly. Among them, the accuracy rate, recall rate, F1 value, and AUC value of the random forest model are the highest, which are 0.85, 0.83, 0.84, and 0.87, respectively, indicating that the random forest model has the best prediction effect on this problem. In contrast, the performance of the linear regression model is poor, possibly due to the nonlinear relationship of the problem. The performance of the decision tree model is between the two, indicating that a single decision tree has a certain effect in solving this problem. However, the random forest improves the model’s performance by integrating learning strategies.

By comparing the modeling effects of linear regression, decision trees, and random forests, we find that the random Forest model performs better in predicting students’ online learning performance. This suggests that using models with nonlinear fitting capabilities and integrated learning strategies should be prioritized when dealing with such problems.

5.4 Result Discussion

By comparing the three models (linear regression, decision tree, and random forest), we found that random forest had the best prediction effect, with an accuracy rate of 0.85, recall rate of 0.83, F1 value of 0.84, and AUC value of 0.87. This indicates that random forest has advantages in dealing with nonlinear relationships and integrated learning. In addition, the feature selection results show that factors such as daily learning duration, course difficulty, and learning behavior data are highly important in predicting students’ online learning performance. This provides a useful reference for improving the teaching quality and personalized recommendations of online learning platforms.

6 Application and prospect

6.1 Applications in the online learning system

The results of this study can be applied to online learning systems. By predicting students’ online learning performance, educators can find students with learning difficulties in real-time and provide personalized tutoring. At the same time, the platform can optimize the course design according to the predicted results, improve the teaching quality, and provide students with a better learning experience.

6.2 Personalized teaching support

By predicting students’ online learning performance, educators can tailor teaching support to students’ needs, such as adjusting teaching schedules, providing differentiated assignments and assessments, and recommending personalized learning resources. This will help improve students’ interest in learning and achievement and achieve educational equity.

6.3 Future research direction

Future research can be carried out in the following directions: 1) Expanding the data scale to improve the prediction effect of the model; 2) Explore more feature combinations and optimize feature selection; 3) Combined with deep learning method to improve model performance; 4) Study the influence of students’ emotional, cognitive and other psychological factors on online learning performance to achieve a more comprehensive prediction.
7 Conclusion

7.1 Main research achievements

In this study, the machine learning method is used to predict students’ online learning performance, and it is found that the characteristics of learning behavior and environmental factors have a great influence. The random forest model was selected with an accuracy of 0.82, which strongly supports online education.

7.2 Research Limitations

There are some limitations in this study: 1) the small scale of data may affect the generalization ability of the model; 2) Feature selection is not perfect and may not cover all key factors; 3) The potential influence of students’ psychological factors on learning performance is ignored.

References