Prediction of Learning Performance in Online Course Based on Linear Regression Model

Jing Gou, Yuan Qin, Yi Luo, Pingyin Luo, Wen He*
School of Big Health and Intelligent Engineering, Chengdu Medical College, Chengdu, China
*Corresponding Author: Wen He Email: hewen41731@163.com

2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) | 978-1-6654-2207-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITAIC54216.2022.9836468

***Abstract*—This paper takes the online learning behavior and performance data of online courses as the research object, analyzes the factors that affect the course performance in online learning behavior, and establishes a performance prediction model. First, the influencing factors of course achievement are analyzed by calculating the correlation coefficient between learning behavior features and course achievement; then, a linear regression performance prediction model is constructed by using single or multiple learning behavior features, and the regression coefficients are solved by the least squares method or gradient descent method ; Finally, the mean square error and coefficient of determination are used to evaluate the model performance. The experimental results show that the top three learning behavior features that have the greatest impact on course performance are the audio and video learning time, number of chapter study times and the number of task points completed, while the multiple linear regression model established using these three learning behavior characteristics and assignment scores has the highest prediction accuracy. The research results can provide reference for online course teachers and learners, help to promote online course learning early warning and performance prediction, and improve the quality of online course teaching.**

Keywords—Online Course; Learning Performance Prediction; Linear Regression Model

I. INTRODUCTION

With the rapid development of the Internet and information technology, more and more educators have networked teaching resources, forming many high-quality online courses, providing learners with online learning opportunities. In the process of online course learning, learners have accumulated a large amount of learning behavior data. It will be a meaningful work to mine and analyze the learning behavior data, find the factors that affect the course performance and establish the performance prediction model. Reference [1] analyzes the relevant course data of a certain senior high school, studies the factors that affect the course performance, and uses the decision tree algorithm to realize the achievement prediction. Reference [2] takes the grades of the first-year course of the 2016 students majoring in information and computing science major in a university as the input, and the average graduation grade as the output, and uses the BP neural network to establish a grade prediction model, which can realize the prediction of graduation grades. Reference [3] uses data mining method to explore the influencing factors of online learners' academic achievement, and uses ensemble learning method to construct a prediction model for academic achievement classification. Reference [4] studied the correlation between college entrance examination information and C programming course scores, and proposed a correlation prediction and analysis model based on random forest algorithm. Reference [5] converts learning behavior data into advanced behavior indicators through literature analysis and in-depth processing of original data, and then uses neural network, decision tree and linear regression algorithms to establish a grade prediction model. Reference [6] predicts course grades based on multiple dimensions such as the professional courses scores, unit tests, and mobile teaching data, and compares the predicted results with the actual grades of the course. Reference [7] combines the data of multiple dimensions such as student behavior, personal attributes and historical grades, and uses support vector machine to predict course grades. Reference [8] established a deep neural network prediction model between MOOC learning behavior and performance through the study of the edx open data set, and put forward personalized teaching feedback and intervention measures according to the experimental results.

From the research of relevant references, at present, researchers' prediction of online course performance is mainly based on different influencing factors, and there is a lack of relatively unified feature indicators of influencing factors; In addition, the prediction of performance is mainly classified prediction, and the prediction result is a discrete data. For online courses, it is more necessary to make continuous value prediction of performances to accurately quantify the learning performance of courses. Therefore, this study will focus on the learning behavior data generated by the online course platform itself, analyze the learning behavior indicators affecting the course performance, and then establish a multiple linear regression model to predict the performance, so as to provide teachers with students' academic early warning, course performance prediction and other references.

II. PERFORMANCE PREDICTION BASED ON LINEAR REGRESSION

A. Data Collection and Preprocessing

This paper takes the data of an online course in our college as the research object, and there are total of 541 students in 12 classes participated in the course. All the data generated in the course learning are collected through the online learning platform, including the students' personal information (student number, name, department, major, class, etc.), learning behavior data (number of task points completed, number of chapter study times, audio and video study time, assignment scores, etc.) and course examination results. In order to protect the privacy of learners, the personal information of students such as student numbers, names, departments, majors and classes is deleted, and then the learning behavior data and course examination results are combined to construct a data set. In order to facilitate subsequent performance analysis and prediction, we encode each feature in the dataset, as shown in Table 1.

TABLE 1. LEARNING BEHAVIORAL FEATURE ENCODING

Learning behavioral feature	Code
Number of task points completed	F1
Number of discussions	F2
Number of chapter study times	F3
Audio and video learning time	F4
Assignment scores	F5
Course examination results	P

B. *Correlation Analysis of Learning Behavior and Achievement*

Online learning behavior mainly includes five dimensions of data, including the number of task points completed, the number of chapter study times, the audio and video learning time, the number of discussions, and assignment scores. By calculating the Pearson correlation coefficient between the features of each learning behavior and course performance [9] It can intuitively reflect whether there is a linear correlation between each feature and course achievement, and its calculation formula is as follows:

$$r=\frac{\sum_{i=1}^{m}\left[\left(x^{(i)}-\bar{x}\right)\left(y^{(i)}-\bar{y}\right)\right]}{\sqrt{\sum_{i=1}^{m}\left(x^{(i)}-\bar{x}\right)^{2}\sum_{i=1}^{m}\left(y^{(i)}-\bar{y}\right)^{2}}} \tag{1}$$

In formula (1), $x^{(i)}$ and $y^{(i)}$ are a certain feature value and target value of the i-th sample, $\bar{x}$ and $\bar{y}$ are the mean values of the samples. The value range of r is [-1,1]. If r=1, it means that x and y are completely positively correlated; if r=-1, it means that x and y are completely negatively correlated; if r=0, it means that x and y are not correlated. Typically, r-values in the following range indicate different correlation strengths between variables: 0.8-1.0, very strong correlation; 0.6-0.8, strong correlation; 0.4-0.6, moderate correlation; 0.2-0.4, weak correlation; 0-0.2, very weak or no correlation. After calculating the correlation coefficient matrix between each learning behavior characteristic and achievement, it is visualized as a heat map, and the result is shown in Fig. 1.

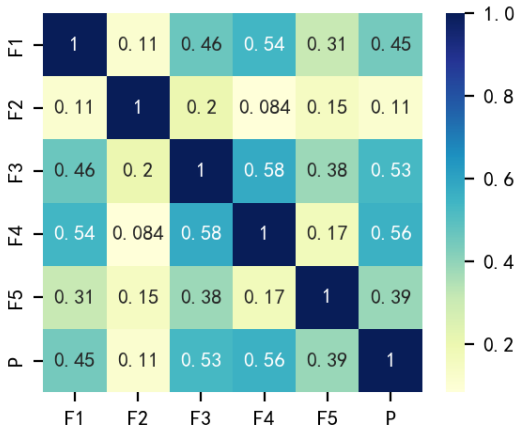

Fig. 1. Correlation Analysis Heatmap

It can be seen from the heat map that the number of task points completed, the number of chapter study times, and the audio and video learning time are moderately correlated with the course achievement; assignment scores are weakly correlated with course achievement; and there was no correlation between the number of discussions and course achievement. There is also a certain correlation between various learning behaviors, which interact and influence each other, and the final achievement of the learner is the result of the joint action of various learning behaviors.

C. *Multiple Linear Regression Prediction Model*

The multiple linear regression model is a powerful tool for regression analysis. Its basic idea is to use regression equation to quantitatively explain the linear dependence between two or more independent variables and dependent variables, and try to find the mathematical expression that can best represent the relationship between independent variables and dependent variables. Through the establishment of multiple linear regression model, the relationship between learning behavior and course achievement can be fitted, so as to realize the prediction of achievement.

Let $\hat{y}$ be the dependent variable and $x_1, x_2, \cdots, x_n$ be the n-dimensional independent variable representing learning behavior features, then the multiple linear regression model fitting the relationship between learning behavior features and course achievements can be expressed as:

$$\hat{y}=\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n \tag{2}$$

where θ_0 is the regression constant and $\theta_1, \theta_2, \cdots, \theta_n$ are the regression coefficients. To solve θ_0~θ_n, we set $x_0=1$, then:

$$\hat{y} = \sum_{j=0}^{n} \theta_j x_j = \theta^T x \tag{3}$$

Define the loss function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2 \tag{4}$$

And the regression parameters can be solved by the ordinary least square method (OLS) or the gradient descent method (GD).

1) Using Ordinary Least Squares (OLS) to Solve Regression Parameters

a) Assuming there are m samples in total, $x^{(i)} = [x_1^{(i)} \quad x_2^{(i)} \quad \cdots \quad x_n^{(i)}]$, $i = 1, 2, \cdots m$, then substituting into formula (2) can form m equations:

$$\begin{cases} y_1 = \theta_0 + \theta_1 x_1^{(1)} + \theta_2 x_2^{(1)} + \cdots \theta_n x_n^{(1)} \\ y_2 = \theta_0 + \theta_1 x_1^{(2)} + \theta_2 x_2^{(2)} + \cdots \theta_n x_n^{(2)} \\ \cdots\cdots \\ y_m = \theta_0 + \theta_1 x_1^{(m)} + \theta_2 x_2^{(m)} + \cdots \theta_n x_n^{(m)} \end{cases} \tag{5}$$

Let:

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \cdots \\ \theta_n \end{pmatrix} \tag{6}$$

Then the loss function $J(\theta)$ can be expressed as:

$$J(\theta) = \frac{1}{2} \| X\theta - Y \|^2 = \frac{1}{2} (X\theta - Y)^T (X\theta - Y) \tag{7}$$

b) Calculate the derivative of the loss function $J(\theta)$ to obtain:

$$\frac{\partial}{\partial \theta} J(\theta) = X^T (X\theta - Y) \tag{8}$$

c) Let $\frac{\partial}{\partial \theta} J(\theta) = 0$ in formula (8), then:

$$\theta = (X^T X)^{-1} X^T Y \tag{9}$$

The calculated θ is the optimal solution. The calculation of the OLS method is simple and efficient, but there are also some limitations: first, the OLS method needs to calculate the inverse matrix of $X^T X$, but the inverse matrix of $X^T X$ may not exist, and the OLS method cannot be used directly to solve it; second , when the sample size is very large, computing the inverse of $X^T X$ is a very time-consuming task, or even infeasible.

2) Using Gradient Descent (GD) to Solve Regression Parameters

In some cases, when the regression parameters cannot be solved by the OLS method, the gradient descent method can be used to solve iteratively. The calculation process is as follows:

a) Calculate the gradient of the loss function at the current position. For θ_j, the gradient calculation formula is as follows:

$$\begin{aligned} \frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2 \\ &= 2 \cdot \frac{1}{2} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) \frac{\partial}{\partial \theta_j} (\hat{y}^{(i)} - y^{(i)}) \\ &= \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) \frac{\partial}{\partial \theta_j} \left(\sum_{j=0}^{n} \theta_j x_j^{(i)} - y^{(i)} \right) \\ &= \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) x_j^{(i)} \end{aligned} \tag{10}$$

b) Multiply the step size by the gradient of the loss function to obtain the descent distance of the current position, that is: $\alpha \frac{\partial}{\partial \theta_j} J(\theta)$.

c) It is determined that for all θ_j, the distance of its gradient descent is less than a small value ε, if it is less than ε, the algorithm terminates, and all the current θ_j is the final result, otherwise, go to step *(d)*.

d) Update all θ_j. For θ_j, the update expression is as follows. After the update is completed, continue to step *(a)*.

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta), \quad j = 1, 2, \cdots, n \tag{11}$$

After solving each θ by using the OLS method or gradient descent method, the multiple linear regression model can be established. At the same time, in the multiple linear regression model, when the number of modeling features is only one, the θ that needs to be solved are only θ_0 and θ_1, and the regression model established at this time is called a simple linear regression model.

III. Experimental Process and Result Analysis

The experimental environment is Windows 10, Python3.8, NumPy and Pandas libraries. Carry out experimental research according to the following process: First, the data set is divided into training set and test set according to the ratio of 7:3, and then according to the correlation coefficient between the learning behavior features and the course performance in the correlation analysis, sort the features relevancy degree, and screen out the features with high degree of relevance to the course performance, and then use these features to establish a linear regression prediction model of course performance, and use ordinary least squares or gradient descent to solve the regression parameters; analyze the prediction accuracy of modeling with each feature, and then compare the accuracy of modeling with all features. Compare and find out the model with the highest

prediction accuracy, so as to achieve performance prediction based on learning behavior data.

The experiment uses MSE and R^2 to evaluate the performance of the model. MSE represents the mean of the residual sum of squares of the true value and the predicted value. The smaller the value, the better. Its expression is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)}) \tag{12}$$

R^2 represents the goodness of fit of the model, and its value range is between 0 and 1 (it can also be negative). The higher the value of R^2, the higher the interpretation degree of the independent variable to the dependent variable, and the more accurate the prediction result. Its calculation formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2} \tag{13}$$

We recorded the performance of prediction models built with different learning behavioral features on training and test sets, as shown in Table 2:

TABLE 2. MODEL EVALUATION

Features	Training set		Testing set		3-Fold cross-validation
	MSE	R^2	MSE	R^2	R^2
F1	134.6671	0.1829	112.8028	0.2468	0.1809
F2	160.0632	0.0289	156.4020	-0.0443	-0.0432
F3	116.2476	0.2947	111.2873	0.2569	0.2831
F4	116.9949	0.2902	95.1468	0.3647	0.3023
F5	135.2376	0.1795	139.4392	0.0689	0.1166
F1, F3, F4, F5	92.4715	0.4390	**89.4035**	**0.4030**	**0.4170**
F1-F5	92.3573	0.4396	90.0808	0.3985	0.4127

According to the correlation coefficient between each learning behavior feature and course performance in the correlation analysis, the most relevant feature is the audio and video learning time, the second is the number of chapter study times, and the third is the number of task points completed. The R^2 of the prediction model based on these three features on the test set is 0.3647, 0.2569 and 0.2468 respectively, and the 3-fold cross-validation scores are 0.3023, 0.2831 and 0.1809, which is higher than the model based on F2 or F5 features. This also shows that the audio and video learning time, the number of chapter learning times and the number of task points completed are the three most critical features of learning behavior that affect the course performance; while the "number of discussions" has a very weak or no correlation with course performance, and the R^2 modeled by it is negative, which indicates that the number of discussions is not a factor that directly affects course performance; when modeling with all learning behavior features, Its MSE on the test set is 90.0808, R^2 is 0.3985, and the 3-fold cross-validation score is 0.4127. The prediction accuracy is higher than the prediction model using any single feature, which also confirms that learners' final achievement is the result of the joint action of various learning behaviors. When modeling with all other learning behavior features excluding the "number of discussions" feature, its MSE on the test set is 89.4035, R^2 is 0.4030, and the 3-fold cross validation score is 0.4170, which is higher than the prediction model established with all features, which shows that in the linear regression prediction model used in this case, removing the extremely weakly correlated or uncorrelated features is helpful to improve the prediction accuracy of the model.

IV. CONCLUSION

This paper analyzes the correlation between online course learners' learning behavior and course performance, and establishes linear regression and multiple linear regression performance prediction models using different learning behavior features. The experimental results show that learning behavior features such as audio and video learning time, number of chapter learning times and number of task point completed are the most key factors in students' online learning, and have the greatest impact on students' performance; At the same time, the multiple linear regression performance prediction model based on the number of task points completed, number of chapter learning times, audio and video learning time and assignment scores performs best, and its 3-fold cross validation score is 0.4170, which exceeds the accuracy of modeling using all five features. In the teaching process of online courses, the results can be predicted through the learning behavior data generated by students' online learning, and the results can be fed back to teachers and students in time, so that teachers can grasp students' learning situation in real time and implement teaching intervention. Students can also adjust their learning state in time through feedback, which plays a good role in promoting the improvement of the teaching quality of online courses. However, this study also has some shortages: from the experimental data, it can be seen that the goodness of fit of the performance

prediction model established through the online learning behavior data is not high. The reason is that there is only a medium or lower correlation between the features of learning behavior and the course performance, which indicates that there may be some factors other than online learning behavior that affect the course performance. We will continue to study relevant problems in the future, find more factors that can affect the course performance, increase the feature number and data volume of the data set, and try to use other regression prediction models to improve the accuracy of online course performance prediction.

ACKNOWLEDGMENT

This project is supported by National Innovation Training Program for College Students (No. 201913705005).

REFERENCES

[1] L.Q. Hu, G. Zhao, "Research on Influencing Factors of Machine Learning Algorithm on Student Achievement Based on Data Mining," Journal of Nanchang Hangkong University: Natural Sciences, vol. 35, issue 3, pp.43-48+97, 2021.

[2] M.H. Yao, J.S. Li, N. Wang, "Prediction of College Students' Performance Based on BP Neural Network," Journal of Jilin University (Information Science Edition), vol.39, issue 4, pp.451-455, 2021.

[3] Z.J. Chen, X.L. Zhu, "Research on Prediction Model of Online Learners' Academic Achievement Based on Educational Data Mining," China Educational Technology, issue 12, pp.75-81+89, 2017.

[4] C. Jin, R.Y. Cui, Y.H. Zhao, "Research on correlation analysis between college entrance examination information and college program design course scores based on machine learning," Journal of Yanbian University (Natural Science), vol.46, issue 4, pp.366-370, 2020.

[5] F.Q. Sun, R. Feng, "A Research on Online Learning Achievement Factors Based on Learning Analysis," China Educational Technology, issue 3, pp.48-54, 2019.

[6] D. Song, D.B. Liu, X. Feng, "Course Performance Prediction and Course Early Warning Research Based on Multi-source Data Analysis," Research in Higher Education of Engineering, issue 1, pp.189-194, 2020.

[7] B.P. Liu, T.C. Fan, H. Yang, "Research on application of early warning of students' achievement based on data mining," Journal of Sichuan University(Natural Science Edition),vol.56, issue 2, pp.267-272, 2019.

[8] P.F. Lin, X.Q. He, T.T. Chen, H.J. Wu, J.H. He, "Prediction of Loss and Teaching Intervention for Learners in MOOC from Perspective of Deep Learning," Computer Engineering and Applications,vol.55, issue 22, pp.258-264, 2019.

[9] Sebastian Raschka & Vahid Mirjalili, Python Machine Learning-Third Edition. Birmingham, Packt Publishing, 2019.