IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 September 2023, accepted 13 November 2023, date of publication 20 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3334915

APPLIED RESEARCH

Analyzing Effective Factors of Online Learning Performance by Interpreting Machine Learning Models

WEN XIAO[1] AND JUAN HU[2]
[1]School of Educational Science, Anhui Normal University, Wuhu 241002, China
[2]School of Computer and Software, Anhui Institute of Information Technology, Wuhu 241002, China

Corresponding author: Wen Xiao (cyees@163.com)

This work was supported by the Anhui Philosophy and Social Sciences Planning Youth Project under Grant AHSKQ2022D097.

ABSTRACT Analyzing the effective factors influencing online learning performance is a research topic that has garnered significant attention. Traditional approaches, such as multiple regression and structural equation models, tend to assume linearity, while non-linear machine learning models lack interpretability. To address this gap, we propose a framework that interprets machine learning models to analyze the effective factors of online learning performance. By applying this framework to four benchmark datasets of online learning, we examine the differential impact of various factors on performance, explore the interactions among these factors, and identify the key factors for representative learners. Our findings indicate that: 1) non-linear machine learning models, particularly Decision Regression, offer better representation of the non-linear relationship between effective factors and online learning performance compared to classical multivariate regression; 2) the factor of online learning behavior exerts a greater influence on performance than demographic features, academic background, or online curriculum design; 3) online learning behavior features exhibit additional interaction effects on performance; and 4) learners with medium performance are influenced by diverse effective factors, with active participation in online learning activities emerging as the most crucial means to improve performance. Interpreting machine learning models presents an innovative approach for analyzing the effective factors of online learning performance, which can be extended to other factor analysis studies. The results of this research provide valuable insights for optimizing machine learning models in predicting online learning performance and enhancing learner outcomes.

INDEX TERMS Explainable AI, effective factors, learning performance, machine learning, online learning.

I. INTRODUCTION

With the support of the rapidly developing information technology infrastructure, online learning is becoming more and more popular and important [1]. At the same time, asynchronous online learning is also the main way for learners to conduct personalized learning and lifelong learning [2]. Different types of educational institutions have set up many online courses which contain digital learning resources and activities on the Learning Management System(LMS) and Massive Online Open Course (MOOC) platform. Because the instructors and learners are separated in time and space in asynchronous online learning, how to analyze the factors that affect learners' online learning performance is the key problem to improve learners' online learning performance and design of online courses.

In recent years, many researchers have investigated the factors that affect learners' online learning performance based on constructivism, cognitivism and other educational and psychological theories. According to these theories, the researchers proposed corresponding hypothesis of effective factors of online learning performance and used correlation analysis, Structural Equation Model (SEM) and other methods to test the proposed hypothesis. Researchers investigated the demographic, cognitive, psychological, behavioral and other features of learners through questionnaire.

The associate editor coordinating the review of this manuscript and approving it for publication was S. Chandrasekaran.

The results of previous studies show that the factors affecting learners' online learning performance include learning motivation [3], cooperative learning [4], autonomous learning ability [5], positive learning emotion [6], infrastructure and hardware [7], self-regulation [8], preparation [9], personality characteristics [10], etc. This method of factor analysis has a solid theoretical basis, but identification and analysis rely heavily on the professional quality and manual work of researchers, and cannot be used for rapid and real-time analysis in large-scale online learning.

On the other hand, digital platforms such as LMS and MOOC that support online learning can automatically record the features and logs of each learner in an incidental manner to generate a big online learning dataset. More and more researchers extract features of learners from these datasets, use nonlinear machine learning models to predict learners' online learning performance, and investigate the effect of different features on online learning performance [3], [11], [12]. Machine learning models commonly used by researchers include Decision Trees [13], Support Vector Machines [14], Naive Bayes [15] and Artificial Neural Networks [16]. These machine learning models can fit online learning datasets very well, and can also achieve high accuracy of prediction for unknown samples [17]. However, nonlinear machine learning models cannot directly generate the diverse impact of features on performance through the weights of features in linear models. They cannot be directly used to analyze the factors that have significant effect on online learning performance.

Therefore, by interpreting the process and results of the machine learning models for predicting online learning performance, the effective factors of online learning performance can be analyzed more accurately, and the impact of different factors can be compared fairly. The Shapley value from the game theory is the average marginal contribution of a feature's value in all possible coalitions [18], which can fairly distribute the difference between the output of the machine learning models and the expected output among features, so Shapley value of features can be used to compare impact among different features. Shapley Additive Explanations (SHAP) value is the best Shapley value based on game theory [19], which is selected as an indicator to represent the impact of different features on the output of machine learning models.

Specifically, the main contributions of this study are as follows: First, the framework of analyzing the effective factors of online learning performance by interpreting the machine learning models has been proposed; Secondly, SHAP values of different features in online learning performance benchmark datasets are calculated, and the impact and interaction of different factors are compared; Thirdly, the representative samples in the benchmark datasets are selected through K-Medoids clustering algorithm, and differences of effective factors of learners with the same performance are analyzed.

The following section of this paper is organized as follows. Related works are listed and discussed in Section II. The framework for analyzing the effective factors of online learning performance by interpreting machine learning models is proposed and explained in Section III. The analysis results of the four benchmark datasets are presented in Section IV. Discussion of analysis results is presented in Section V. The conclusion and limitations of this study are included in Section VI.

II. RELATED WORKS

Previous studies about analyzing the effective factors of online learning performance can be divided into two categories. In the first category, researchers analyze the effective factors through hypothesis and test, while in the second category, researchers use machine learning models and online learning datasets.

In related works based on hypothesis and test, researchers analyzed the possible effective factors on online learning performance based on specific theories in social sciences such as pedagogy and psychology, and put forward corresponding hypotheses. The researchers collected data by issuing questionnaires and conducting interviews, and used correlation analysis or structural equation model to test whether the hypothesis was accepted or rejected.

Peggy et al. applied Online Collaborative Learning (OCL) theory integrating with cognitive development to evaluate the effectiveness of student learning performance through OCL. They investigated 373 respondents through the online survey, and used structural equation model to analyze the impact of seven factors on online learning performance: online collaborative tools, collaboration with peers, student engagement, idea generating, idea organizing, intelligent convergence, and student learning outcome. The research results show that these factors have significant positive effects on online learning performance [4]. Ma et al. constructed a questionnaire on the impact of learner autonomy on online learning performance. The research results show that four factors, preparation of technology and target plan, utilization of materials in learning contents, regulation of learning process, and evaluation of learning effect, have significant effects on online learning performance [5]. Zhu et al. collected 1088 samples of online courses in China using questionnaires, and investigated the impact of learning motivation and positive emotions on college students' online learning performance using correlation analysis and multiple mediator analysis [6]. Rajabalee and Santally coded and analyzed the feedback of 665 students, and the results showed that there was a weak but positive significant impact between participation and online learning performance [20]. Li et al. used the correlation analysis to investigate the impact of 12 factors on social conditional learning and online learning performance. The results show that motivation regulation, trust building, effectiveness management, cognitive strategies, time management, goal setting, task strategies, peer support, team assessment, seeking help, environmental construction and team supervision are significantly related to students' performance, and team supervision is negatively correlated with team performance [21]. Oyelere et al. collected data from

63 students using questionnaires, and analyzed online learning performance by using multiple linear regression on team experience and self-regulated learning related factors [8]. The results of this study show that there is a correlation between autonomous learning, online course level and students' online course achievements. Wei and Chou used questionnaires to investigate 356 students and constructed a structural model to analyze whether online learning perception and online learning preparation affect students' online learning performance [9]. The results show that students' online learning perception and preparation have a direct and positive impact on online learning performance. Ejubovic et al. investigated 375 students with questionnaires. Based on multiple regression analysis, they tested the impact of five factors in self-regulated learning on online learning performance. The research results show that four factors, including environment structuring, computer self-efficiency, social dimension, and metacognitive strategies, have a positive impact on online learning performance [22]. The comparison of these studies is shown in Table 1.

TABLE 1. Comparison of related works based on hypothesis and test.

Study	Data Source	Factors	Model
[4]	Questionnaire	7 factors about online collaborative learning	Structural Equation Modeling (PLS-SEM)
[5]	Questionnaire	4 factors about online collaborative learning	Structural Equation Modeling
[6]	Questionnaire	7 factors about online learning motivation and positive emotions	Structural Equation Modeling
[20]	Questionnaire	student satisfaction and engagement	correlation analysis
[21]	Questionnaire	12 factors about social learning	correlation analysis
[8]	Questionnaire	9 factors about teamwork experience and self-regulated learning	multiple regression analysis
[9]	Questionnaire	online learning perception and preparation	Structural Equation Modeling
[22]	Questionnaire	5 factors about self-regulated learning	multiple regression analysis

Since digital platforms such as LMS and MOOC that support online learning can automatically record a large number of learners' online learning behaviors and other logs in a non-interference manner, more and more researchers begin to extract learners' demographic, online learning behaviors and other features from online learning dataset, use machine learning models to fit these datasets, and analyze the impact of different factors on online learning performance.

Maier et al. extracted 10 evaluation scores and grades in the online learning process from LMS logs, and used unsupervised auto-encoders to extract latent features from these features, further improving the accuracy of performance prediction model [23]. Zsuzsanna et al. also used the unsupervised self-organizing map model to extract latent features from logs, and used these features as the input of the prediction model based on logical regression [24]. Although these extracted latent features can improve the accuracy of online learning performance prediction model, they are semantically ambiguous and have no obvious help in analyzing the effective factors of online performance.

Wang et al. extracted 9 features about learners' demography and online learning behavior from LMS logs, and proposed an prediction model of online learning performance based on recurrent neural networks, with the prediction accuracy reaching 78.85% [25]. High prediction accuracy means that this machine learning model can fit online learning performance dataset well. However, they did not interpret the model and could not identify which of the nine features had a more significant impact on online learning performance. Lee et al. extracted online learning behavior features from LMS logs and established a prediction model of online learning performance using artificial neural networks. Similarly, they did not analyze the impact of different factors on performance [11]. Mi et al. extracted six online learning behavior features from LMS logs, and used multiple linear regression to analyze the impact of different combinations of features on prediction accuracy of prediction model [25]. Their experimental results show that the selected combination of features can further improve the prediction accuracy of the model. This also shows that different features have different impact on online learning performance, and the combination of different features may also have interaction effects. Aydoğdu used the demographic and online learning behavior features extracted from LMS logs to establish a prediction model of online learning performance based on feedforward neural network, which reached 80.47% of the prediction accuracy [12]. In order to analyze the impact of different factors on online learning performance, they analyzed the weight of links between neurons in the artificial neural network, pointing out that login times and time spent have the most significant impact on online learning performance. Meng et al. used the multiple regression model to analyze the relationship between online learning behavior and online learning performance. Through the weight of different factors in the multiple regression model, we can infer the impact of different factors on the final performance [3]. Saqr et al. used an open source social network analysis tool called Gephi to analyze the online discussions of learners from Moodle in four courses, extracted the centrality features and basic attributes of social networks, such as size, density, average, etc., as the effective factors of online learning performance [26]. They used a multiple regression model to analyze the correlation between these factors and online learning performance. The research results show that students' centrality and interaction characteristics in social networks have a strong positive correlation with online learning performance. The comparison of these studies is shown in Table 2.

TABLE 2. Comparison of related work based on machine learning model.

Study	Data Source	Factors	Model
[25]	Logs in LMS	9 factors about demography and online learning behavior	Recurrent Neural Network
[23]	Logs in LMS	10 factors about grades of assignment and practical exam	linear discriminant analysis
[25]	Logs in LMS	6 factors about online learning behavior	Multivariable Linear Regression
[26]	Online discussion of four courses in LMS	three groups of centrality measures obtained by quantitative network analysis，and the properties of networks	Multivariable Linear Regression
[12]	Logs in LMS	10 factors of students' demography and online learning behavior	Feedforward Neural Network
[3]	Logs in LMS	6 factors about online learning behavior	Multivariable Linear Regression

In general, although researchers have made significant achievements in analyzing effective factors of online learning performance, there are still some challenges in these two types of studies. Studies based on hypothesis and test have a solid theoretical basis, but the size of the data collected through questionnaire is small. The proposal of hypothesis relies on the professional knowledge and experience of researchers, and the analysis of effective factors is subjective. Structural equation model, multiple regression and correlation analysis can only identify the linear relationship between factors and online learning performance, and cannot analyze the interaction between different factors. The prediction model established by researchers based on nonlinear machine learning model and online learning dataset overcomes the previous shortcomings, but they did not interpret the machine learning model, analyze the impact of different factors on online learning performance, the interaction between factors and the effective factors of representative learners.

III. METHOD

The purpose of this study is to bridge the gap of studies based on machine learning models and online learning datasets, and analyze the effective factors of online learning performance, the interaction of different factors and the factors of representative learners by generating SHAP values of different features, that is, the impact on the prediction results of machine learning models. Therefore, the research questions in this study are as follows:

RQ1: Which features have more significant impact on learners' online learning performance? which factors are these features belong to?

RQ2: Whether multiple factors have interactive effects on online learning performance?

RQ3: Whether effective factors of learners with the same performance are the same or different?

According to the research objectives and research questions, the research design framework of this study is shown in Figure 1.

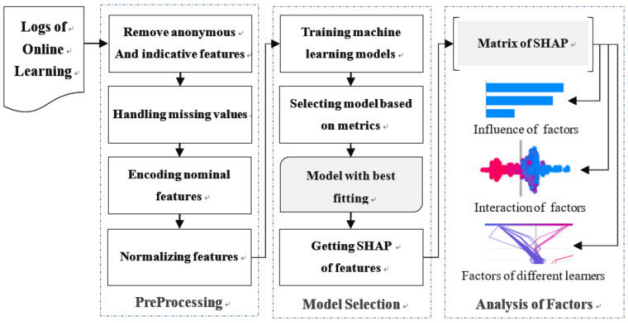


FIGURE 1. Research design framework of this study.

A. DATASETS AND TOOLS

According to the research objectives and research questions of this study, four online learning performance benchmark datasets from two famous machine learning dataset Repository, UCI Machine Learning Repository [27] and Kaggle, were selected. These datasets were selected for the following reasons. Firstly, the four datasets contain actual online learning performance of learners, which can be used to train supervised machine learning models. Secondly, there are few missing and error values in the four datasets. Thirdly, these datasets contain multiple features and factors, which can be used to compare the impact of features and factors. Fourthly, xAPIEdu is collected by K-12 students, while other datasets involve university and adult learners, which makes the result of factor analysis more universal. The important properties of these four benchmark datasets are shown in Table 3. In fact, students' performance is more suitable to be represented by numerical values, with larger values indicating better performance. The students' performance represented by distinct labels have lost this semantic information. Therefore, we convert categorical learner performance in xAPIEdu and OULA into numerical values, used to train machine learning models for regression like Harvard and Canvas respectively.

The Python packages selected in this study to establish and interpret machine learning models include pandas 1.15, scikit-learn 1.1.3, shap 0.40.0, and matplotlib 3.5.3. Among them, pandas is used for loading dataset and preprocessing, scikit-learn provides many machine learning models, shap is used to generate shap values of all features in the datasets, and matplotlib is used for visual analysis.

B. PREPROCESSING

The functions provided by pandas are used to preprocess the selected benchmark datasets. Preprocessing includes

TABLE 3. Important properties of the benchmark datasets in this study.

Dataset	Samples	Features	Learners' performance in Datasets	Results of regression
xAPIEdu [28]	480	16	{L,M,H}	{1,2,3}
Harvard (Person-Course2013) [29]	338223	19	[0,1]	[0,1]
Canvas (Network Person-Course)[30]	325199	25	[0,1]	[0,1]
Open University Learning Analytics dataset (OULA)[31]	25793	16	{withdrawn, Fail, Pass, Distinction}	{1,2,3,4}

removing anonymous and indicative features, handling missing values, encoding nominal features and normalization. Firstly, these benchmark datasets contain anonymous privacy features such as student name, email, nationality, and indicative features such as course name and learner's ID. Obviously, these features will not affect learners' online learning performance. Therefore, we remove anonymous and indicative features from each dataset in preprocessing. Secondly, in addition to xAPIEdu, the other three benchmark datasets contain some missing values. Since there are many samples in these datasets, to further improve the quality of the machine learning models, we use dropna() provided by Pandas to discard the samples containing missing values directly, the structure of the datasets has not changed. The numbers of remaining samples for the four benchmark datasets in Table 3 is 480,25211,1113,25793 respectively. Thirdly, online learning performance of learners in the four benchmark datasets is orderly, whether expressed by grades or numerical values. Therefore, in order to unify the impact of different features on online learning performance, we use the LabelEncoder and map() provided by Pandas to transform all nominal features in these datasets into numeric features so that they can be fitted by the regression machine learning model. Finally, in order to unify the order of magnitude of different features, we use the Max-Min Normalization method to convert all values of features into values in the range of [0,1].

C. MODEL SELECTION

In order to analyze effective factors of online learning performance accurately, we choose the machine learning models with the minimum training error on the benchmark datasets as the best model for analysis. We have selected five regressor provided in scikit-learn 1.1.3 named LinearRegression, LogisticRegression, DecisionTreeRepressor, RandomForestRegression, AdaBoostRegression as the representatives of five types of machine learning models. The values of parameters in these models are default values provided by scikit-learn package.

Obviously, machine learning models with smaller training errors can reflect the impact of different factors on online learning performance more accurately. Therefore, we use MAE(1) and R^2(2) as the evaluation indicators of the above five machine learning models.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{1}$$

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

D. FACTOR ANALYSIS

According to the research objectives and questions of this study, we interpret the selected machine learning models from the following three aspects:

Firstly, based on the selected machine learning models, we generate SHAP for each feature of all samples in the dataset, and compare the impact of different factors on online learning performance. The Shapley value is the average marginal contribution of feature value in the possible coalitions. Specifically, the Shapley value of feature value ***j*** in the ***i-th*** sample φ_{ij} represents the contribution to the output of this sample compared with the average output of all samples in dataset based on machine learning model(3). SHAP value is the best Shapley value based on game theory (5), x' represents a coalition vector where the values of all features are presence. If dataset X contains N samples, each sample contains p features $\{x_1, x_2, \ldots, x_p\}$, X_i is the ***i-th*** sample in X, $X_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$, SHAP values for all features in X_i are expressed as $\{\varphi_{i1}, \varphi_{i2}, \ldots, \varphi_{ip}\}$. The matrix of SHAP values generated by dataset X is expressed as (6). SHAP value of feature x_j denoted by SH_j is the average value of all samples.

$$\varphi_{ij}(val) = \sum_{S\subseteq\{x_{i1},x_{i2},\ldots,x_{ip}\}\setminus\{x_{ij}\}} \frac{|S|!\,(P-|S|-1)!}{p!} \left(val(S\cup\{x_{ij}\}) - val(S)\right) \tag{3}$$

$$val_{xi}(S) = \int \hat{f}\left(x_{i1}, x_{i2}, \ldots, x_{ip}\right) dP_{xi\notin S} - E_X(\hat{f}(X)) \tag{4}$$

$$g\left(x'\right) = \varphi_0 + \sum_{j=1}^{M}\varphi_j \tag{5}$$

$$\begin{bmatrix} \varphi_{11} & \cdots & \varphi_{1p} \\ \vdots & \ddots & \vdots \\ \varphi_{N1} & \cdots & \varphi_{Np} \end{bmatrix} \tag{6}$$

$$SH_j = \frac{1}{N}\sum_{i=1}^{N}|\varphi_{ij}| \tag{7}$$

Secondly, we investigate interaction effect between features, and analyze the additional impact of features combinations on online learning performance. Interaction effect is the additional combined feature effect after considering the contribution of a single feature. In a single sample, the interaction effect of feature x_i and x_j denoted by δ_{ij} equals the total effect minus the main effect of the two features(8). Shapley interaction between two features denoted by φ_{ij} is the average of the values on all possible feature

coalitions S(9). The results of investigation can show that the interaction between different features will have an additional impact on learners' online learning performance.

$$\delta_{ij}(S) = \hat{f}_X\left(S \cup \{x_i, x_j\}\right) - \hat{f}_X\left(S \cup \{x_i\}\right) - \hat{f}_X\left(S \cup \{x_j\}\right) + \hat{f}_X(S) \quad (8)$$

$$\varphi_{ij} = \sum_{S \subseteq \{x_i, x_i, \ldots, x_i\} \setminus \{x_i, x_j\}} \frac{|S|!\,(M - |S| - 2)!}{2\,(M-1)!} \varphi_{ij}(S), \quad M = |S| \quad (9)$$

Thirdly, we investigate SHAP values of representative learners, and analyze effective factors of different learners with the same performance. We use K-Medoids to cluster the benchmark datasets, and take centers of clusters generated by clustering as representative learners, and then analyze the SHAP values of these representative learners. K-Medoids is a clustering algorithm based on similarity [32]. It selects an actual sample as the center of a cluster. The average similarity between the center of cluster and other samples in the cluster is the largest, which can be used as a representative of a group of similar learners.

IV. RESULTS

We used the research framework proposed in Section III to investigate four selected online learning performance benchmark datasets. The results of model selection, impact of factors, interaction of factors and effective factors of representative learners are reported as follows.

A. MODEL SELECTION

We used five regressors to fit the four benchmark datasets after preprocessing, Linear regression is a conventional model used in previous studies to analyze the impact of different factors on students' performance., the other four regressors are universally nonlinear machine learning models. In order to make full use of these datasets and avoid the impact of randomness on the performance of RandomForestRegressor and AdaBoostRegressor, we repeated 100 times to randomly selected 75% of the samples from the benchmark datasets to fit the regressors, and used MAE and R^2 to illustrate the performance of fitting. Table 4 shows the mean and standard deviation of 100 experimental results of five regressors on four benchmark datasets. Because the purpose of this study is to analyze the impact of different factors on online learning performance, that is, the impact of different features on the output of machine learning models, the prediction performance of these models does not need to be significantly concerned.

From Table 4, we can see that all regressors have very stable performance, that is, the standard deviation of all experimental results is very small, which indicates that we can use mean of results to compare the fitting of regressors on datasets. DecisionTreeRegressor can completely fit four benchmark datasets (MAE=0.00, $R^2 = 1.00$), and its results of regression are completely consistent with the actual performance. From the average values of five regressors on four datasets for two indicators, it can be seen that the fitting of LinearRegressor(MAE=0.28, $R^2 = 0.61$) on the four benchmark datasets are worse than LogisticRegressor(MAE=0.3, $R^2 = 0.68$) and RandomForestRegressor(MAE=0.1, $R^2 = 0.95$), which indicates that there are nonlinear correlations between features and online learning performance in benchmark datasets, and the linear regression model cannot reflect the impact of different features on learning performance completely. Performance of AdaBoostRegressor(MAE=0.3, $R^2 = 0.62$) is similar to that of LinearRegressor. Although it is a nonlinear machine learning model based on boosting, it has no obvious advantage for fitting online learning performance datasets. The possible reason is that multiple base_estimators in AdaBoostRegressor are not available for fitting benchmark datasets.

Because DecisionTreeRecressor has the best fit to the benchmark dataset, we chose it as the machine learning model for interpretation.

TABLE 4. Performance of five regressors on four benchmark datasets.

Regressor	Datasets							
	xAPIEdu		Harvard		Canvas		OULA	
	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
LinearRegressor	0.3 3±0 .01	0.6 9±0 .01	0.0 7±0 .00	0.7 3±0 .01	0.1 2±0 .00	0.6 9±0 .02	0.6 ±0. 00	0.3 3±0 .01
LogisticRegressor	0.2 1±0 .01	0.7 1±0 .02	0.1 2±0 .00	0.7 5±0 .01	0.3 ±0. 01	0.7 9±0 .03	0.5 9±0 .01	0.5 ±0. 01
DecisionTreeRegressor	0.0 0±0 .00	1.0 0±0 .00	0.0 0±0 .00	1.0 0±0 .00	0.0 0±0 .00	1.0 0±0 .00	0.0 0±0 .00	1.0 0±0 .00
RandomForestRegressor	0.1 2±0 .01	0.9 5±0 .01	0.0 2±0 .00	0.9 7±0 .00	0.0 3±0 .00	0.9 7±0 .00	0.2 1±0 .00	0.9 1±0 .00
AdaBoostRegressor	0.4 0±0 .02	0.6 7±0 .02	0.0 7±0 .01	0.7 3±0 .01	0.1 1±0 .02	0.7 6±0 .03	0.6 2±0 .01	0.3 3±0 .01

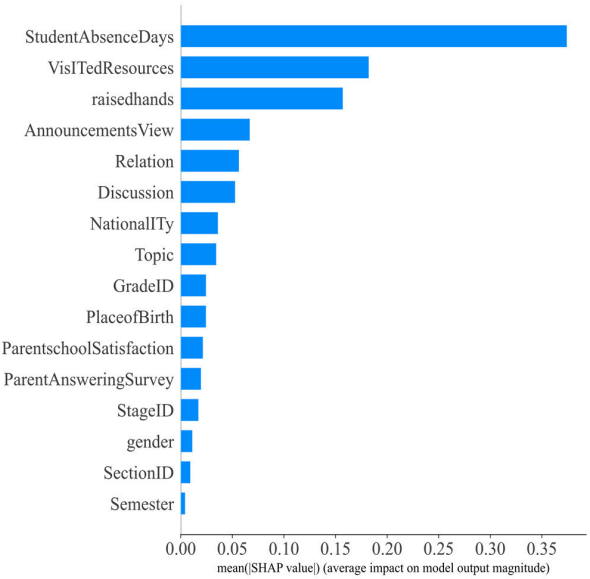
B. INFLUENCES OF FACTORS

In previous studies, researchers generally used Covariance $Cov(X, Y)$ in SEM(10), Person correlation coefficients $\rho_{X,Y}$ (11), or regression coefficients β_i in linear regression(12) to indicate the impact of factors on online learning performance. These coefficients can only represent the linear relationship between factors and performance, and can only be used if all variables follow a normal distribution. The SHAP values of features in nonlinear machine learning models use a fair value allocation method to indicate the contribution of features to the outcomes, which has the advantages of completeness and consistency and can overcome the shortcomings of the aforementioned coefficients in linear models

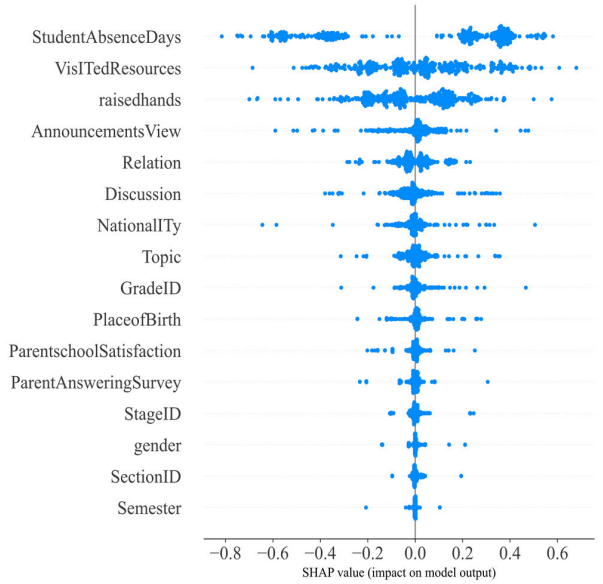
$$Cov(X, Y) = E[(X - E(X))(Y - E(Y)] \quad (10)$$

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad (11)$$

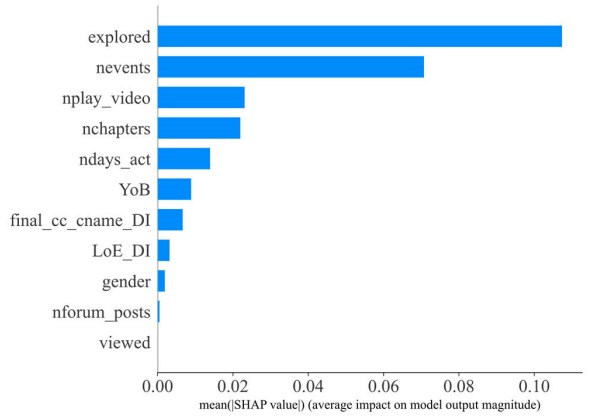
$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon \quad (12)$$



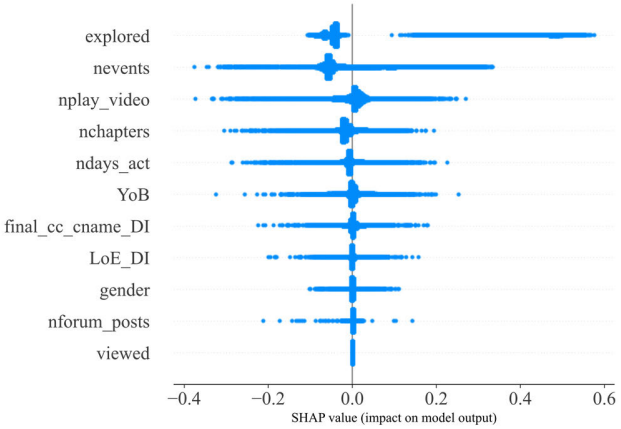
(a)Overall impact of each feature(xAPIEdu)



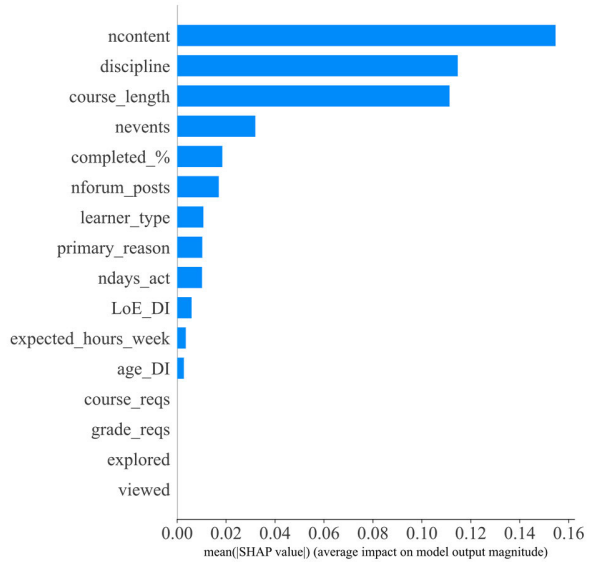
(b) Impact of each value of feature (xAPIEdu)



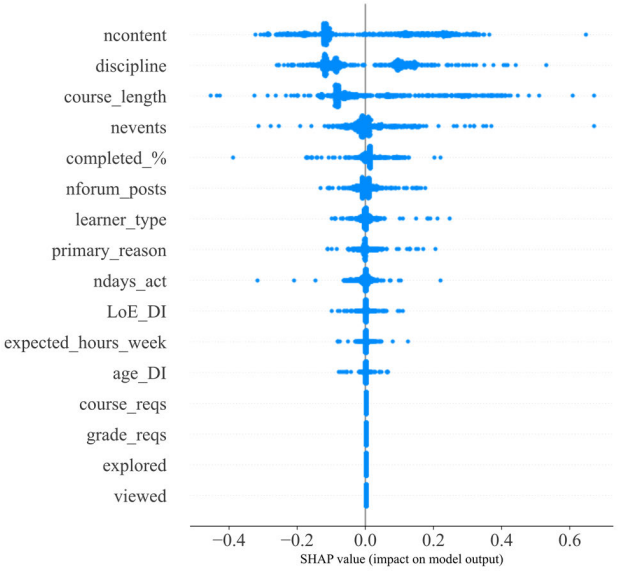
(c) Overall impact of each feature (Harvard)



(d) Impact of each value of feature (Harvard)

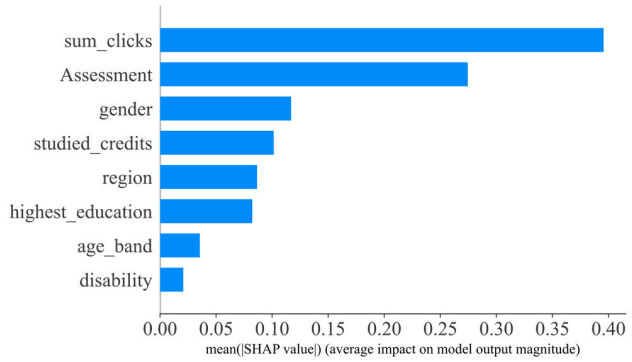


(e) Overall impact of each feature (Canvas)

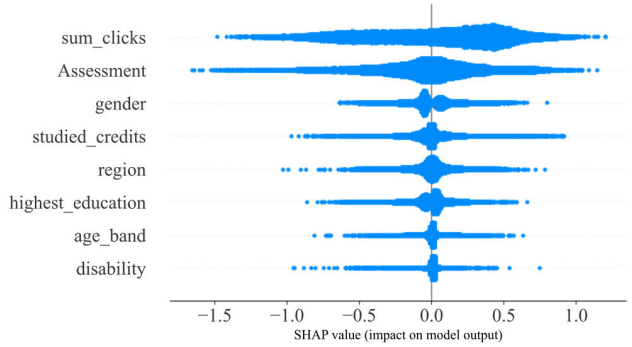


(f) Impact of each value of feature (Canvas)

FIGURE 2. **The impact of features on online learning performance.**



(g) Overall impact of each feature (OULA)



(h) Impact of each value of feature (OULA)

FIGURE 2. ***(Continued.)*** **The impact of features on online learning performance.**

After preprocessing, the students' performance in benchmark datasets was numerical, and the average of all students' performance in each dataset was used as the expectation when generating the SHAP values of features. We use (3) to generate the impact of every value of features in each sample of the benchmark dataset, and use (7) to generate the overall impact of each feature on online learning performance. Compared to expectations, the features in each sample may have a negative or positive impact on learning performance, yet the overall impact of features on performance in each dataset are summations of the absolute values of the impact of features in each sample. Specifically, we use Figure 2 (a)(c)(e)(g) to show the overall impact of each feature and Figure 2 (b)(d)(f)(h) to show the impact of features in each sample, respectively.

Figure 2 shows the following results:

The results in Figure 2 (a) show that the five most effective features in xAPIEdu are StudentAbsenceDays, raisedhands, VislTedResources, Relation, Discussion, which represent the number of students' absent days, the number of questions, the number of visits to digital resources, family relations and the number of online discussions respectively. Four of these five features belong to factor of online learning behavior, and one belongs to learners' demographic factors. The absence of learners means that there is no learning behavior, resulting in the most significant impact on online learning performance. Other features have little impact on online learning performance which belong to demographic, academic background and other factors. In particular, Semester has no impact on online learning performance. The results in Figure 2(b) show that StudentAbsenceDays, raisedhands and VislTedResources have different effects on online learning performance in various samples. The distribution of negative and positive effects of these three features are continuous which show significant positive correlation with online learning performance. However, impact of features are intensive which indicated that these features have no significant impact on most learners.

The results in Figure 2 (c) show that the five most effective features in Harvard are explored, nevents, and nplay_video, nchapters, ndays_act, which represent whether learners have visited more than 50% of course modules, the number of interactions with course, number of play video, the number of interactions course's chapters and number of unique days learners' interacted with course respectively. These five features all belong to factor of online learning behavior. Due to the large number of samples in Harvard, the impact of feature values in different samples on online learning performance are distributed regularly as shown in Figure 2(d), which also shows that learners' online learning behavior have a very stable impact on online learning performance.

The results in Figure 2(e) show that the most effective features in Canvas are completed_%, course_length, discipline, nevents, nforum_posts which represent percent of total required content modules completed, number of days that course officially ran, discipline of the course, count of distinct interactions with course, number of posts total in discussion forums respectively. Three of these features are belong to factor of online learning behavior, course_length and discipline are factors of online course design. It can be seen from Figure 2(f) that, unlike the regular distribution of the impact of online learning behavior features in different samples, the impact of the two features belong to online course design factor in different samples shows obvious aggregation, which means that a specific online course design can have a similar impact on the online learning performance of a group of students.

The result in Figure 2(g) shows sum_clicks and Assessment which represent learners' sum clicks of resource and score of assessment respectively belong to factor of online learning behavior have the most significant impact, and also show that studied_credits which belongs to factor of learners' academic background, has a significant impact on online learning performance. Although gender and region which belong to factor of learners' demography have a considerable impact on performance, the results in Figure 2(h) show that the number of samples affected by these two features is relatively small.

C. INTERACTION OF FACTORS

Covariance(10), correlation coefficient(11) have been used in previous studies to indicate linear correlations between factors, they cannot be used to illustrate the additional impact

of the interaction between the two factors on student performance. Analysis of Variance(ANOVA) indicates whether there is an interaction effect between two factors through the sum of squares of inter group deviations(13), *SSAB* is the sum of squared deviations between the interaction groups of factors *A* and *B, r, c* are the degrees of freedom of A, B respectively, $\overline{y}_{ij}$ is the average of groups, $\overline{y}_{...}$ is the average of total samples, $\overline{y}_{i..}$, $\overline{y}_{.j.}$ are the average of groups of factors A and B respectively, but it can only be used to compare the mean differences between different groups and cannot provide other information. Additionally, it requires the two factors to have natures such as normality, homogeneity, and independence.

$$SSAB = \sum_{i=1}^{r} \sum_{j=1}^{c} (\overline{y_{ij}} - \overline{y_{i..}} - \overline{y_{.j.}} + \overline{y_{...}})^2 \qquad (13)$$

The SHAP value can decompose the interpretation of the model's output into the contributions of each feature, enabling us to analyze the interactions between factors. By calculating the SHAP values of different feature combinations, we can quantify the impact of different combinations on student performance and more clearly indicate the interaction between factors.

We use (9) to generate the interaction effect between two features, that is, the additional impact of the coalition of two feature values on online learning performance. If there is no interaction between the two feature values, the additional effect is 0, otherwise it will have a negative or positive additional impact on online learning performance. The feature interaction effect of the four benchmark datasets is shown in Figure 3. Each point in the figure represents the interaction effect between two feature values in a sample. The farther the point is from the gray vertical line, the stronger the interaction between the two features. Since the diagram of interaction effect on the diagonal represents the effect relationship of the feature itself, we cannot consider it. Because interaction diagrams on both sides of the diagonal are same, we only need to investigate the diagram of interaction effect on one side of the diagonal.

From Figure 3 we can see that:

In xAPIEdu, coalitions of features that have significant additional effects on online learning performance are {StudentAbsenceDays, raisedhands}, {StudentAbsenceDays, VislTedResources}, {raisedhands, VislTedResources}. These coalitions are composed of online learning behavior features, and the distribution of interaction effects is very similar. Combining with Figure 2 (a), we find that there is a strong positive correlation between different online learning behaviors of learners, and the combination of online learning behaviors can bring additional positive effects on online learning performance, that is, multiple active online learning behaviors can induce additional positive effects on online learning performance.

In Harvard, due to the large number of samples, the additional effects of different features combinations on online learning performance are regular, and there is no combination

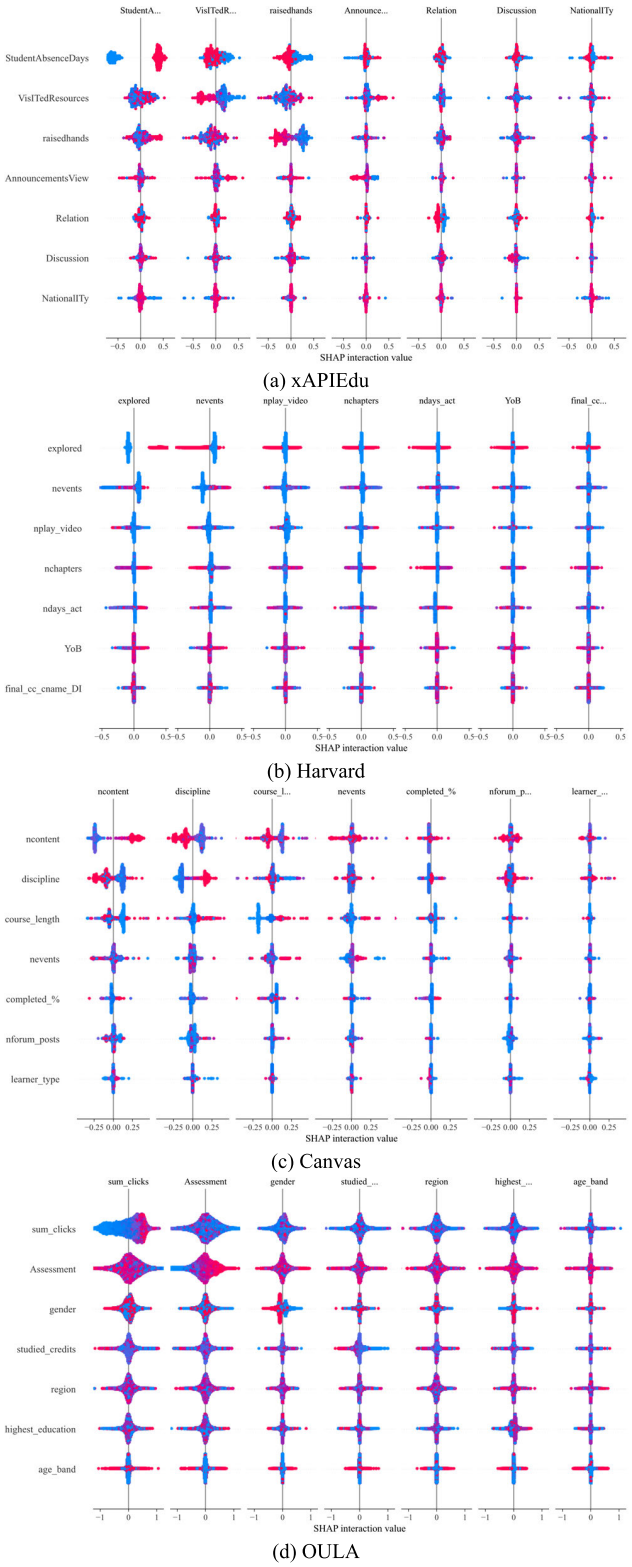


FIGURE 3. Interaction of factors.

of features with significant interaction effects. In contrast, {explored, Nevents} has more negative effects on online learning performance, while {nevents, nplay_video} has more positive effects, which also shows that online learning

behavior can produce significant additional effects on online learning performance. In other words, if learners actively participate in various online learning activities, they are likely to achieve better performance.

In Canvas, coalitions of features that have significant additional effects on online learning performance are {completed_%, course_length}, {completed_%, discipline}, {course_length, discipline}. The first coalition is that completion of online course and length of course will have a significant interaction effect on online learning performance. Combined with Figure 2 (f), we find that with the completion of the course greater, the greater the additional effect on online learning performance. The distributions of the additional effects generated by the coalitions of the other two features are very similar, which may be mainly caused by the additional effects generated by discipline, indicating that the discipline of the online course is closely related to completion of course, and the coalitions of these features will have a significant impact on online learning performance.

In OULA, coalitions of features that have significant additional effects on online learning performance is {sum_clicks, Assessment}, which reflects that multiple online learning behaviors can have additional effects on online learning performance. Combined with Figure 2 (f), we find that sum_clicks and Assessment have a strong positive correlation, and the coalition of them will also generate additional positive effects on online learning performance. If learners actively visit learning resources and get excellent scores in the assessment, they generally have good learning performance.

D. EFFECTIVE FACTORS OF REPRESENTATIVE LEARNERS

In order to investigate the differences of effective factors of learners with the same performance, we use K-Medoids to cluster the four benchmark datasets. Clustering is an effective method for selecting representative samples. In the clustering process, samples are grouped into clusters according to similarity, we can choose the cluster centroids as samples with typical features, which can represent the clusters they belong to [33]. Since K-Medoids is a clustering algorithm based on similarity, the centers centroids generated by K-Medoids can be used as representative samples. For xAPI-Edu and OULA, the number of clusters is set to 10 times the number of performance labels, and for Harvard and Canvas, the number of clusters is set to 50. In summary, we use the cluster centroids obtained from clustering as typical samples and select 30, 50, 50, and 40 representative learners from four benchmark datasets respectively for analysis.

The results of analysis are shown in Figure 4. In Figure 4, each line represents a representative learner. The impact of different features on online learning performance is represented by the twists and turns of the line. The value at the top of the figure represents the learner's performance, and the vertical gray line represents the average performance of all learners.

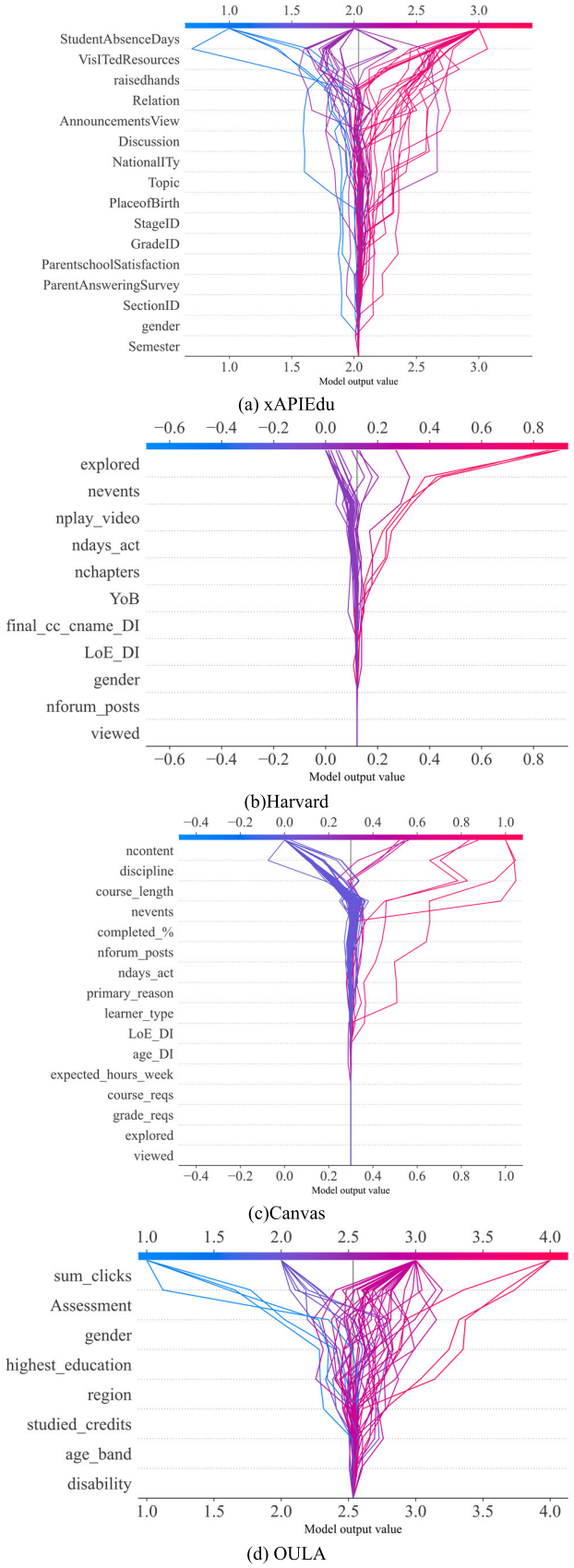


FIGURE 4. Effective factors of representative learners.

The results in Figure 4 show that:

In xAPIEdu, lines which represents learners with low (1.00) and high (3.00) performance mostly overlap in (a), and online learning performance is positively correlated with three features of online learning behavior that are StudentAbsenceDays, raisedhands, and VislTedResources, which indicates that effective factors of students with low and high learning performance are very similar. On the contrary, there are obvious differences in effective factors of different learners with Medium (2.0), which also shows that effective factors of online learners are diverse, especially for learners with normal performance. Most of them are affected by factor of online learning behavior, but a small number of learners are significantly affected by academic background, demography and other factors.

In Harvard and Canvas, most learners' online learning performance is failure or dropout (0.0), and most of the lines which represent these learners are overlap, indicating that effective factors of learners with failure or dropout are factor of online learning behavior which includes features of explored, nevents, completed_%. We can also find that there is a significant positive correlation between online learning behavior factors and online learning performance.

In OULA, the lines represent learners with different performances are separated by features of sum_clicks and Assessment clearly, showing that learners' online learning performance is positively correlated with factor of online learning behavior. A small number of learners with 1. 0 (withdrawn) are greatly affected by academic background and demographic factors, which indicates that the factors leading to online learning dropout are diverse.

V. DISCUSSIONS

In previous studies, researchers used structural equation models and multiple regression analysis to analyze the impact of different factors on online learning performance. These factors include online collaborative learning, online learning motivation, social learning, self-binding, etc., but linear models cannot expose the nonlinear relationship between the features of learners and online learning performance. In other studies, researchers used machine learning models and demographic, online learning behavior, academic background and other factors to predict learners' online learning performance accurately, but there was a lack of interpretation for the machine learning models, so they cannot distinguish the impact of different factors on online learning performance. In order to bridge the gap, we interpreted machine learning models based on four famous online learning performance benchmark datasets in this study.

The key findings of the study are as follows:

Nonlinear Machine learning model can better fit the online learning performance dataset, which is conducive to exploring the nonlinear relationship between different factors and online learning performance. The results show that LinearRegressor has the worst performance compared with four nonlinear models on the four benchmark datasets, especially on OULA, and the linear model can hardly fit this dataset (MAE=0.6, $R^2 = 0.33$). LogisticRegressor has better performance than LinearRegressor, but it is almost invalid on Harvard ($R^2 = 0.21$) and OULA ($R^2 = 0.01$). We may not be able to explore the relationship between factors of learners and online learning performance with an explicit nonlinear function such as sigmoid. RandomForestRegressor has better performance than LinearRegressor and LogisticRegressor on the four benchmark datasets, which conforms to the common sense that ensemble-based models have better performance than single models. In particular, DecisionTreeRegressor has an amazing fitting (MAE=0.00, $R^2 = 1.00$) on the four benchmark datasets, which may mean that the decision tree model based on information-gain is particularly suitable for analyzing online learning performance dataset. From the experimental results in subsection IV-B, it can be seen that learners' online learning performance is often significantly affected by a few characteristics. Accordingly, in the process of constructing decision trees, features with greater differentiation or information entropy will be selected earlier, which should be an important reason why decision trees are especially suitable for online learning performance regression analysis.

Online learning performance is generally significantly affected by a few features which belong to factor of online learning behavior. In the four benchmark datasets, features that have significant impact on learners' online learning performance are xAPIEdu (StudentAbsenceDays, raisedhands, VislTedResources), Harvard (explored, nevents), Canvas (completed_%, course_length), OULA (sum_clicks, Assessment). Eight of the nine features belong to factor of online learning behavior, and only course_length is an online course design factor. The results show that, compared with learners' factors of demography, academic background, and online curriculum design, factor of online learning behavior has the most significant impact on online learning performance. At the same time, factor of online course design will have a concentrated impact on a group of learners. In particular, the aggregation features of learners' online learning behaviors have the most significant impact, such as StudentAbsenceDays in xAPIEdu, explored in Harvard, completed_% in Canvas and sum_clicks in OULA. Because online learning behavior has a definite positive impact on performance, and these aggregation features synthesize various behaviors, so they have more significant impact on performance than single behavior. Based on this finding, we can not only improve learners' performance by urging them to participate in online learning activities more actively, but also predict learners' online learning performance based on online learning behavior features accurately.

Features coalitions with significant impact on online learning performance have a strong interaction effect. Features coalitions with significant interaction effects in benchmark datasets as follow: xAPIEdu {(StudentAbsenceDays,raisedhands), (StudentAbsenceDays, VislTedResources), (raisedhands, VislTedResources)}, Harvard

{(explored,nevents)}, Canvas {(completed_%,course_length)}, OULA {(sum_clicks,Assessment)}. We found that all features coalitions with significant interaction effects belong to online learning behavior factors, which also shows that actively participating in online learning activities can significantly improve learners' online learning performance. On the other hand, there is a strong correlation between learners' different online learning behaviors. For example, learners who actively visit learning resources will also actively participate in the discussion, and it is also positively related to score of assessment.

Learners with similar performance may be affected by different factors. The low performance and excellent performance of learners in the four benchmark datasets are labeled as xAPIEdu (1.0), Harvard (0.0), Canvas (0.0), OULA (1.0) and xAPIEdu (3.0), Harvard (1.0), Canvas (1.0), OULA (4.0) respectively. We find that the effective factors of learners with low and excellent performance are very similar, but learners with medium performance may affected by different factors. For example, in xAPIEdu, most learners with medium performance (2.00) are affected by factors of online learning behavior which include features of AbsenceDays and raisedhands, while some learners are also greatly affected by factors of academic background which include features of GradeID, Relation, etc. In OULA, a small number of learners with medium performance (2.0, 3.0) are significantly affected by demographic factors which include gender, education and other features. In short, for learners with low or high performance, it can be inferred that they have negative and positive online learning behaviors respectively, while learners with medium performance may be affected by various factors.

VI. CONCLUSION AND FUTURE WORK

The main purpose of this research is to interpret the process and results of nonlinear machine learning model's output by using different features' marginal effects on the output, and analyze the impact and interaction of different factors. In this research, we propose a framework of using nonlinear machine learning models to fit online learning datasets and interpret machine learning models based on SHAP values. The results show that nonlinear machine learning models, especially Decision Regression, can fit online learning performance datasets better than traditional linear modes. In other words, these models can better reveal the nonlinear relationship between effective factors and online learning performance. Learners' online learning performance is often significantly affected by a few features, and factors of online learning behavior have a greater impact than demographic, academic background, online curriculum design and other factors. Different online learning behavior features have additional interaction impact on online learning performance. There is a very obvious positive correlation between factors of online learning behavior and online learning performance, while demographic factors of learners and online course design factors have similar impact on a group of learners. The effective factors of learners with medium performance are diverse, and the most important way to improve learners' online learning performance is to make them more actively participate in online learning activities. At the same time, online learning behavior features are more suitable for predicting learners' performance. This research provides a valuable contribution to analyze the effective factors of online learning performance and predict learners' online learning performance using machine learning models.

This research has some limitations. First of all, four benchmark datasets we used are the passively collected by extracting from logs of LMS rather than collected datasets under the guidance of learning theories actively. These datasets contain a small number of features and are not comprehensive enough. Secondly, SHAP-based analysis results lack robustness as it is dependent on the training data. In addition, nonlinear machine learning models are still seen as a black box, the mechanism of effective factors and casual relationship were not thoroughly investigated. In the future, under the guidance of cognitivism and constructivism, we will take the initiative to collect more comprehensive datasets for analysis, and use inferential machine learning to analyze the causal relationships between different factors more intrinsically.

DATA AVAILABILITY STATEMENT

The xAPIEdu datasets analysed during the current study are available in the kaggle repository, [https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data].

The Harvard datasets analysed during the current study are available in the HARVARD Dataverse, [https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147].

The Canvas datasets analysed during the current study are available in the HARVARD Dataverse, [https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/1XORAL].

The OULA datasets analysed during the current study are available in the kaggle repository, [https://www.kaggle.com/datasets/anlgrbz/student-demographics-online-education-dataoulad].

REFERENCES

[1] H. O. Falola, O. O. Ogueyungbo, A. A. Adeniji, and E. Adesina, "Exploring sustainable E-learning platforms for improved universities' faculty engagement in the new world of work," *Sustainability*, vol. 14, no. 7, p. 3850, Mar. 2022.

[2] S. Jacques, A. Ouahabi, and T. Lequeu, "Synchronous E-learning in higher education during the COVID-19 pandemic," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Apr. 2021, pp. 1102–1109.

[3] X. Meng and Z. Hu, "The relationship between student motivation and academic performance: The mediating role of online learning behavior," *Qual. Assurance Educ.*, vol. 31, no. 1, pp. 167–180, Jan. 2023.

[4] P. M. L. Ng, J. K. Y. Chan, and K. K. Lit, "Student learning performance in online collaborative learning," *Educ. Inf. Technol.*, vol. 27, no. 6, pp. 8129–8145, Jul. 2022.

[5] X. Ma, "Influence study of learners' independent learning ability on learning performance in online learning," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 9, pp. 201–213, May 2022.

[6] Y. Zhu, S. Xu, W. Wang, L. Zhang, D. Liu, Z. Liu, and Y. Xu, "The impact of online and offline learning motivation on learning performance: The mediating role of positive academic emotion," *Educ. Inf. Technol.*, vol. 27, no. 7, pp. 8921–8938, Aug. 2022.
[7] C. Chisadza, M. Clance, T. Mthembu, N. Nicholls, and E. Yitbarek, "Online and face-to-face learning: Evidence from students' performance during the COVID-19 pandemic," *Afr. Develop. Rev.*, vol. 33, no. S1, pp. S114–S125, Apr. 2021.
[8] S. S. Oyelere, S. A. Olaleye, O. S. Balogun, and Ł. Tomczyk, "Do teamwork experience and self-regulated learning determine the performance of students in an online educational technology course?" *Educ. Inf. Technol.*, vol. 26, no. 5, pp. 5311–5335, Sep. 2021.
[9] H.-C. Wei and C. Chou, "Online learning performance and satisfaction: Do perceptions and readiness matter?" *Distance Educ.*, vol. 41, no. 1, pp. 48–69, Jan. 2020.
[10] N. Alkış and T. T. Temizel, "The impact of motivation and personality on academic performance in online and blended learning environments," *Educ. Technol. Soc.*, vol. 21, no. 3, pp. 35–47, 2018.
[11] C.-A. Lee, J.-W. Tzeng, N.-F. Huang, and Y.-S. Su, "Prediction of student performance in massive open online courses using deep learning system based on learning behaviors," *Educ. Technol. Soc.*, vol. 24, no. 3, pp. 130–146, 2021.
[12] Ş. Aydoğdu, "Predicting student final performance using artificial neural networks in online learning environments," *Educ. Inf. Technol.*, vol. 25, no. 3, pp. 1913–1927, May 2020.
[13] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *J. Chemometrics*, vol. 18, no. 6, pp. 275–285, Jun. 2004.
[14] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.
[15] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naive Bayes," in *Encyclopedia of Machine Learning*, vol. 15. Switzerland : Springer, 2010, pp. 713–714.
[16] I. N. Da Silva, D. H. Spatti, R. A. Flauzino, L. H. B. Liboni, and S. F. D. R. Alves, *Artificial Neural Networks*, vol. 39. Switzerland : Springer, 2017.
[17] W. Xiao, P. Ji, and J. Hu, "A survey on educational data mining methods used for predicting students' performance," *Eng. Rep.*, vol. 4, no. 5, May 2022, Art. no. e12482.
[18] K. Hausken and M. Mohr, "The value of a player in n-person games," *Social Choice Welfare*, vol. 18, pp. 465–483, 2001.
[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.
[20] Y. B. Rajabalee and M. I. Santally, "Learner satisfaction, engagement and performances in an online module: Implications for institutional e-learning policy," *Educ. Inf. Technol.*, vol. 26, no. 3, pp. 2623–2656, May 2021.
[21] C. Li, Y. Peng, P. Peng, and L. Cao, "A study of fuzzy modeling analysis of factors influencing socially regulation of learning performance in an online environment," *J. Intell. Fuzzy Syst.*, vol. 41, no. 3, pp. 4639–4649, Oct. 2021.
[22] A. Ejubovic and A. Puška, "Impact of self-regulated learning on academic performance and satisfaction of students in the online environment," *Knowl. Manag. E-Learn.*, vol. 11, no. 3, pp. 345–363, 2019.
[23] M.-I. Maier, G. Czibula, and Z.-E. Oneţ-Marian, "Towards using unsupervised learning for comparing traditional and synchronous online learning in assessing students' academic performance," *Mathematics*, vol. 9, no. 22, p. 2870, Nov. 2021.
[24] Z. Oneţ-Marian, G. Czibula, and M. Maier, "Using self-organizing maps for comparing students' academic performance in online and traditional learning environments," *Stud. Informat. Control*, vol. 30, no. 4, pp. 31–42, Dec. 2021.
[25] X. Wang, L. Zhang, and T. He, "Learning performance prediction-based personalized feedback in online learning via machine learning," *Sustainability*, vol. 14, no. 13, p. 7654, Jun. 2022.
[26] M. Saqr, U. Fors, and J. Nouri, "Using social network analysis to understand online problem-based learning and predict performance," *PLoS ONE*, vol. 13, no. 9, Sep. 2018, Art. no. e0203590.
[27] H. Mi, Z. Gao, Q. Zhang, and Y. Zheng, "Research on constructing online learning performance prediction model combining feature selection and neural network," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 7, pp. 94–111, Apr. 2022.
[28] M. Kelly, R. Longjohn, and K. Nottingham, "The UCI machine learning repository," [Online]. Available: https://archive.ics.uci.edu
[29] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods," *Int. J. Database Theory Appl.*, vol. 9, no. 8, pp. 119–136, Aug. 2016.
[30] *HarvardX Person-Course Academic Year 2013 De-Identified Dataset, Version 3.0, Harvard Dataverse*, HarvardX, Cambridge, MA, USA, 2014.
[31] N. Canvas, "Canvas network person-course (1/2014–9/2015) de-identified open dataset," Harvard Dataverse, Canvas Netw., Cambridge, MA, USA, Tech. Rep., 2016, doi: 10.7910/DVN/1XORAL.
[32] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," *Sci. Data*, vol. 4, no. 1, pp. 1–8, Nov. 2017.
[33] P. Arora and S. Varshney, "Analysis of K-means and K-medoids algorithm for big data," *Proc. Comput. Sci.*, vol. 78, pp. 507–512, 2016.

WEN XIAO received the Ph.D. degree from Hohai University, China, in 2020. He is currently with Anhui Normal University. His research interests include educational data mining and distributed computing. He is a member of the China Computer Federation (CCF).

JUAN HU received the master's degree from Nanjing Normal University, China, in 2010. Since 2022, she has been with the Anhui Institute of Information Technology. Her research interests include big data mining and machine learning.