# Research on Constructing Online Learning Performance Prediction Model Combining Feature Selection and Neural Network

https://doi.org/10.3991/ijet.v17i07.25587

Huichao Mi, Zhanghao Gao, Qiaorong Zhang, Yafeng Zheng(✉)
College of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, China
mmhhcc@126.com

**Abstract—**Learning performance prediction can help teachers find students who tend to fail as early as possible so as to give them timely help, which is of great significance for online education. With the availability of online data and the continuous development of machine learning technology, learning performance prediction in large-scale online education is gaining new momentum. Traditional prediction methods include statistical methods, machine learning and neural networks. Among them, statistical methods and machine learning have low prediction efficiency. Although neural network can improve prediction efficiency, it ignores the impact of artificial feature filtering on model performance, and cannot find key factors for performance prediction, making predictions uninterpretable. Therefore, this paper proposes an online academic performance prediction model that integrates feature selection and neural network. Multiple linear regression analysis is used for feature extraction to obtain key influence features, and then deep neural networks is used for prediction. The results show that the F1 score of our model on large-scale data set is 99.25%, which is 1.25% higher than that of other related models.

**Keywords—**learning performance prediction, machine learning, deep neural networks, multiple linear regression, feature selection

## 1 Introduction

With the promotion of educational revolution by Internet plus, Massive online education has become an important form of education and teaching [1]. However, in the development process of online education platforms, due to the openness of online courses, the unsupervised nature of online classes, and the unrestricted participation and withdrawal of courses, the dropout rate and completion rate of MOOCs remain low [2]. Research shows that the vast majority of MOOC completion rates range from 0.7% to 52.1%, with an average of 12.6%, and about 80% of learners fail to pass, dropout and completion rates have become key issues that hinder the development of MOOCs [3]. To solve this problem, applying machine learning algorithm to student achievement prediction is the most commonly used method at present[4]. It is helpful for teachers to

identify high-risk students early and give active intervention and guidance to find out the potential risk of failure and dropout in the future from the existing online interaction behaviors of students. Researchers have carried out a lot of researches on the construction methods of predictive models based on key behavioral characteristics. Compared with demographic information such as gender and age, explicit and implicit behavioral characteristics such as language characteristics and learning behavior can be used as better feature selection factors[5]. The application of Machine Learning technology can effectively improve the accuracy of the prediction model. Although reported literature has shown that different machine learning algorithms can provide satisfactory results and similar performances, however, due to different problem specifications and evaluation indicators usually directly affect the performance of the algorithms, it is a challenging task to compare and analyze the prediction performance of machine learning algorithms[6]. However, before prediction and analysis, these studies did not consider the impact of input data features on the performance of machine learning algorithms, appropriate features can effectively improve the performance of machine learning [7] and achieve interpretability in educational practice[8].

The study divides learners into five categories from high to low according to student performance, and uses a variety of feature selection methods to determine the optimal feature combination. The optimal feature combination is input into five prediction models for comparative research, including Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM) and Deep Neural Network (DNN). Finally, it is determined that the prediction model combining multiple linear regression and DNN has the best prediction effect.

There are three methods in the study of learning prediction of mass online education: statistics-based method, traditional machine learning method and deep learning method. The statistics-based method is the prediction method that is often used in early studies, including correlation analysis, regression analysis, hypothesis testing, etc. For example, Logical regression and chi-square tests were used to analyze the factors leading to dropout and predict whether students would drop out or no. The results shows that the main reasons for students to drop out are the number of course interactions, number of course visits, number of chapters completed, number of videos watched and number of forum posts[9]. [10] used correlation analysis to analyze dropout rates, and to investigate different behavioral patterns that lead students to drop out, the results showed that first semester grade is a key indicator of early dropout, and that dual degree is a significant factor in increasing the risk of dropout. The advantage of statistical methods is that the technique is simple, easy to operate and understand. However, feature selection in statistical methods is mostly acquired through passive observation, which results in low prediction accuracy, and the processing speed and efficiency cannot be applied to large-scale datasets.

Machine learning algorithm is the main technical method used in MOOC academic performance prediction at present. Machine learning algorithms can be divided into two categories: traditional machine learning based and deep learning based. Traditional machine learning algorithms regards achievement prediction as a classification problem, and establishes prediction model by using logic regression, support vector machine, decision tree, etc. For example, [11] analyzed more than 400,000 user behavior data of

XueTangX platform, LIBSVM model was used to predict dropout based on 49 kinds of user behavior data, the result shows that the prediction accuracy reached 76.56%. [12] extracted the important features of MOOC learners, and then used the enhanced decision tree model to predict the learners who could not complete the course. The prediction accuracy reached 80%. [13] explored the effect of the number of behavioral features on the performance of student dropout prediction through Random Forest, Adaptive Boost, XGBoost, and GradientBoost Classifiers, proving that the accuracy of the model can reach 93.2% when selecting 2 optimal features for prediction, outperforming the selection of a large number of redundant features. Similarly, [14] Analyzed 120542 data from XueTangX platform and extracted seven kinds of characteristic behaviors, such as questioning, discussion and browsing. They used the combination of fine-grained feature generation and logistic regression model to predict, and the accuracy rate of dropout prediction reached 86.3%, the accuracy rate of achievement prediction reached 74.8%.

With the development of deep learning technology, it has been widely used in performance prediction. [15] explored the configuration of multiple feedforward neural network architectures. The research optimizes the accuracy of the model by changing the number of hidden layers and the number of the neurons, and the model achieves 97.46% accuracy. [16] proposed a predictive model based on Convolutional Neural Network (CNN) to predict the drop-out rate of MOOCs. They used CNN to extract the best features from the 7 behavioral features of XueTangX platform, which reduced the complexity of the model and maximized the classification effect. The prediction accuracy of the model reached 86.75%. Some researchers regarded drop-out prediction and achievement prediction as a time series problem. [17] proposed a prediction model based on Recursive Neural Network (RNN), which regarded the user's activities as a time series, and the AUC value of the model reached 88.1%. [18] extracted 17 behaviors from course platform, including online time, online days and video playing times, and used a Long Short-Term Memory recurrent neural network (LSTM) to build a model to predict students' future learning trend and performance. The results show that the AUC of dropout prediction is 86.3%, while the AUC of achievement prediction is 74.8%.

In conclusion, extracting user access behavior data from clickstream can effectively improve model prediction performance. But there are still some problems in using machine learning algorithm to predict academic performance in existing studies. First, in the existing studies, the prediction of dropout is mostly a binary classification problem, that is, drop-out is defined as "yes" or "no", while for the more fine-grained classification problem, the related research is less. Second, due to the complexity of MOOC education environment, the data size is also different, and the prediction performance of various machine learning algorithms on different scale datasets is not clear. Third, neural networks, especially deep neural networks, can obtain better prediction results. However, due to the characteristics of its "black box" operation [19], it is difficult for people to understand the impact of features, activation functions and hidden layer settings on prediction results while obtaining high-performance prediction models, which makes it difficult for teachers to use predicted results to carry out accurate teaching reform.

Based on the above analysis, this study compares the prediction performance of mainstream machine learning algorithms under different data scales in the same data scenario, and uses feature selection methods to extract the optimal feature combination to improve the interpretability of the prediction model. Finally, the accuracy rate, recall rate and F1 score are used to evaluate the prediction performance of these models. The main work of this study is as follows:

- Students are divided into more fine-grained categories: according to their scores, learners are divided into five categories: loser (F), near loser (E), low-quality learner (L), medium learner (M) and high-quality learner (L).
- Analyzing the efficiencies of algorithms under different data scales: five machine learning algorithms are used in six data sets of different scales to compare their prediction efficiencies and analyze the causes.
- A prediction model that combines feature selection and DNN is proposed: multiple regression analysis is used for feature selection to improve the interpretability of educational predictions, which helps teachers better understand the factors that most affect academic performance, and then carry out effective teaching.

This paper is organized as follow. Section 2 introduces five machine learning algorithms. In Section 3, we design an experiment, using different machine learning algorithms to analyze different scale datasets from edX datasets, and divide students into five categories according to the predicted scores. In Section 4, the prediction performance of the five models is compared and analyzed, and it is pointed out that the DNN model with Multiple Linear Regression feature selection method can effectively predict academic performance. Finally, the conclusion and the future development direction of machine learning algorithm applied to student achievement prediction is discussed.

## 2 Related work

### 2.1 Logical Regression (LR)

Logistic Regression is a supervised learning classification method based on generalized Linear Regression analysis [20]. In short, Logistic Regression adds a logical model on Linear Regression model, so that the continuous values predicted by the Linear Regression model are changed into discrete values. For example, for a binary classification problem, the result of the target attribute is 0 or 1, and the logistic regression constructs a mapping function to map the continuous data obtained by the linear model to 0 or 1. The formula of the logistic regression model is as follows:

$$y = g(z) \tag{1}$$

where $z$ is a Linear Regression model, $z = W^T X_i + \theta$, $W$ is the regression coefficient matrix of the linear model, $\theta$ is the offset of the model, and y is predicted value.

In logistic regression, the Sigmoid function is generally selected as the mapping function to solve the binary classification problem. The formula of logistic regression model using the Sigmoid function is as follows:

$$y = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(W^T X_i+\theta)}} \tag{2}$$

In order to avoid overfitting, we use the L2 regularization rule.

### 2.2 Decision Tree (DT)

Decision tree is a common classification and regression method of machine learning. Its basic idea is to make predictions through the conditional probability distribution on feature space and class space [21]. It is a classification model or regression model established using a tree structure, which is composed of nodes and directed edges. It divides the dataset by continuously constructing dividing conditions until the stopping condition is met. In the process of dividing nodes, decision tree uses Gini index, information gain or information gain ratio to find the optimal attributes to construct the dividing conditions. At the same time, in order to avoid over-fitting, the method of pre-pruning or post-pruning is used to remove useless rules. The study uses the tree. Decision Tree Classifier algorithm in Python to construct a decision tree, and uses the trained decision tree model to predict learner performance classification.

### 2.3 Naive Bayes (NB)

Naive Bayes is a supervised linear classification algorithm. It is based on Bayes' theorem and the assumption of independence of characteristic conditions. Naive Bayes algorithm calculates the posterior probabilities of given samples in each category according to the prior probability of each category and the conditional probability of each attribute of given samples, and selects the category with the largest probability value as the predicted category of the given sample [22].

This paper uses Gaussian Naive Bayes in the Naive Bayes classification algorithm to conduct experiments. The formula of prior probability is defined as follows:

$$P(Y = C_k) = \frac{|D_k|}{|D|} \tag{3}$$

where $|D|$ represents the data volume of training set $D$, and $|D_k|$ represents the data volume of the k-th type of data in training set $D$. In the experiment, we use the GaussianNB classifier in Python to construct and train a naive Bayes model, and then the trained naive Bayes model is used to classify the test set data.

### 2.4 Support Vector Machines (SVM)

Support vector machines are considered to be one of the most powerful methods in machine learning algorithms. It is especially applicable for small-scale datasets and is not sensitive to the dimensionality of data. It can handle both classification problems

and regression problems. Moreover, it is also applicable to linear and non-linear problems. Therefore, SVM model is one of the most popular models in machine learning. SVM represents feature vectors as points in a multi-dimensional space. Using a mapping function, it maps different types of vectors to different planes as much as possible, that is, map the feature vectors in a low-dimensional to a high-dimensional space. Therefore, SVM can transform a non-linear separable dataset into a linearly separable dataset, which can better solve the non-linear classification problem [23]. SVM takes the geometric interval as the distance between support vector and hyperplane, and constructs the segmentation hyperplane with the largest geometric interval for classification.

The formula of SVM hyperplane is as follows:

$$\mathrm{W^T \cdot X + b = 0} \tag{4}$$

where $X$ is feature vector, $W$ represents the normal vector of hyperplane, and $b$ represents the offset of hyperplane.

For the hyperplane $(W, b)$ obtained in the feature space of training set $D$, the geometric interval formula of a sample point $(x_i, y_i)$ is defined as follows:

$$\gamma_i = y_i\left(\frac{W^T}{||W||} \cdot x_i + \frac{b}{||W||}\right) \tag{5}$$

where $||W||$ represents the 2-norm of W.

### 2.5 Deep Neural Network (DNN)

Deep Neural Network has become the basis of many artificial intelligence applications [24]. DNN model consists of three parts: input layer, hidden layer and output layer. It contains many hidden layers, so DNN is also called multi-layer Neural Network. This paper uses labeled dataset to train DNN model, and uses the batch gradient descent method to optimize the neural network. In the process of model training, Back Propagation (BP) algorithm or Forward Propagation (FP) algorithm can be used to adjust the mapping function of neurons layer by layer, and finally get the optimal DNN model [25].

This paper uses the TensorFlow platform as the experimental platform. The DNN model built contains 3 hidden layers, and the number of neurons in each layer is 10, 20, and 10 layers. The DNN model structure used in the experiment is shown in Figure 1.

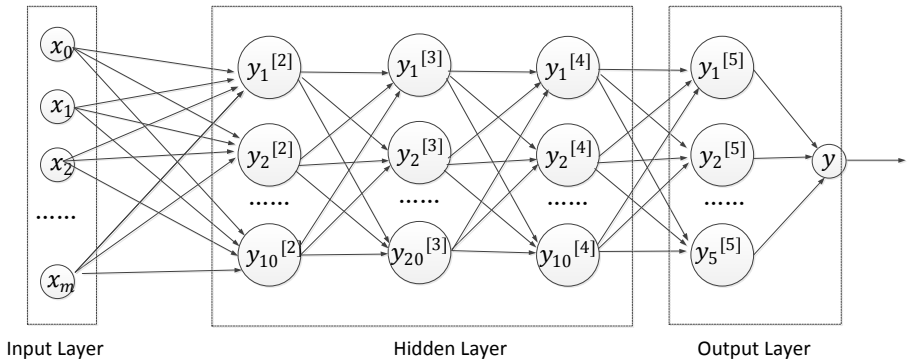


**Fig. 1.** DNN model

where $x_i$ is the *i*-th feature of a data in dataset;$y_i^{[l]}$ is the *i*-th neuron in the *l*-th layer in DNN model; *y* is the predicted classification of DNN model.

In the input layer, the training set after data cleaning is input into the model, and then input into the hidden layer after linear calculation.

In the hidden layers, DNN algorithm uses labeled training set for training to find the model with the best performance. In order to prevent Gradient disappearance problem, we use ReLU function as the activation function of neurons. The formula of ReLU function is as follows:

$$\mathrm{ReLU(x)} = \begin{cases} 0 & \mathrm{x} \le 0 \\ \mathrm{x} & \mathrm{x} > 0 \end{cases} \tag{6}$$

where $x$ is input data.

Since this paper deals with multi-classification problems, we use the SoftMax function as the activation function of the neurons in the output layer to process the input data of the output layer. This function is suitable for multi-classification problems and can convert the prediction result into the probability of each class. Therefore, using the SoftMax function, DNN can process the input value of each neuron in the output layer, and finally select the category with the largest function value as the prediction result. The formula of the SoftMax function is as follows:

$$\mathrm{y_i} = \frac{\mathrm{e^{x_i}}}{\sum_{\mathrm{k}} \mathrm{e^{x_k}}} \tag{7}$$

where $\mathrm{y}_i$ is the output of the *i*-th neuron in the output layer, $x_i$ is the input of the *i*-th neuron. Finally, DNN model selects the class with the largest $\mathrm{y}_i$ value as the prediction result. In order to obtain the optimal model, DNN uses cross entropy as a loss function to calculate the gap between predicted value and true value. The smaller the cross entropy of the model is, the closer the predicted result is to the actual value. The formula for cross entropy is as follows:

$$\mathrm{loss} = -\sum_{\mathrm{i}} \hat{\mathrm{y}}_{\mathrm{i}} \ln \mathrm{y_i} \tag{8}$$

where $\widehat{y_i}$ is the true classification of the *i*-th score in dataset *D*, and $y_i$ is the predicted classification.

# 3 Experiment procedure

## 3.1 Data set description and preprocessing

The edX open dataset used in this study is the MOOC learner dataset launched by Stanford University and MIT jointly, and it is one of the most comprehensive public datasets in terms of the amount and diversity of online learning data currently available. The data set contains 641,138 learner records, including four categories of course information, basic learner information, learner type information and learner behavior information, a total of 18 information fields. This study selects a total of 48,000 pieces of data from Chinese and American learners for analysis. Among them, the learner's behavior information includes the interaction times of the course(nevents), the number of days of course visits(ndays_act), the number of videos played(nplay_video), the number of chapters studied(nchapters), the number of forum posts(nforum_posts), the time of course registration(start_time_DI) and the last login time(last_event_DI). The difference value between the last login time and the course registration time was used as the new learning days variable(ndays), and a total of 6 learning behaviors were used as the input characteristics of the prediction model.

In terms of achievement classification, learners are divided into 5 categories according to their scores.

Loser (F): The score is within the range of 0~0.2.

Near loser (E): The score is within the range of 0.2~0.4.

Low quality learners (L): the score is within the range of 0.4~0.6.

Intermediate learners (M): The score is within the range of 0.6~0.8.

High quality learners (H): The score is within the range of 0.8~1.0.

## 3.2 Research design process

In order to study the prediction effect of each model under different scale data sets, this experiment divides 6 data sets of different scales according to the scale from small to large, and their data volumes are 8000, 16000, 24000, 32000, 40000 and 48000 respectively. The training set data and the test set data are selected according to the data division ratio of 8:2. Under the condition of the same scale data set, each model uses exactly the same number of training set data and test set to facilitate the comparison of performance differences between models. Logical regression algorithm, decision tree algorithm, Naive Bayes algorithm, support vector machine algorithm and deep neural network algorithm are used to build a prediction model to predict score. Use the training set to train these 5 models separately, and input the test set into the trained model for verification. The specific implementation process is shown in Figure 2.

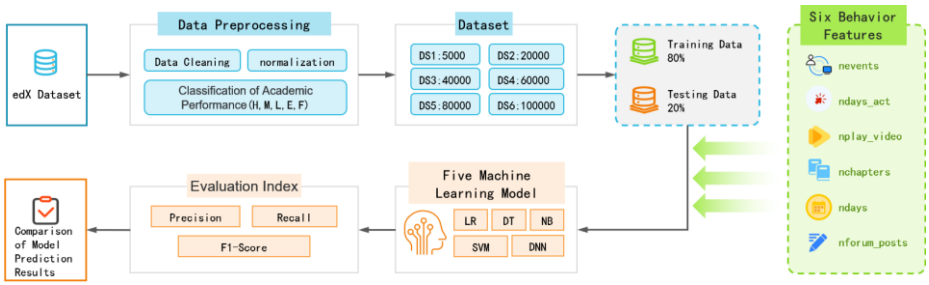


**Fig. 2.** Comparison of the implementation process of machine learning prediction models

In order to further explore the influence of behavioral feature selection on algorithm performance, three methods of SFS, SBS, and multiple regression analysis are used to select the optimal feature combination, and they are combined with five algorithm models to classify the data set and evaluate its performance. The implementation process is shown in Figure 2.

SFS is sequential forward selection, which is a way to select one feature each time from all features to make the evaluation result of feature combination reach the optimal and add feature combination until the evaluation result cannot be optimized.

SBS is sequential backward selection, which takes all features as the initial feature combination and deletes one feature at a time so that the evaluation result of feature combination after deleting this feature can reach the optimal level until the evaluation result cannot be optimized.

In this study, SFS and SBS feature selection methods are combined with Logistic Regression model, and model accuracy is used as the evaluation value of feature combination for feature selection.

Multiple Linear Regression uses all the features of linear fitting and the "step by step" model of SPSS linear analysis to analyze data and get the fitting results of the regression model. The optimal feature combination is selected according to the influence degree of each feature on the linear model. The implemented feature selection framework is shown in Figure 3.

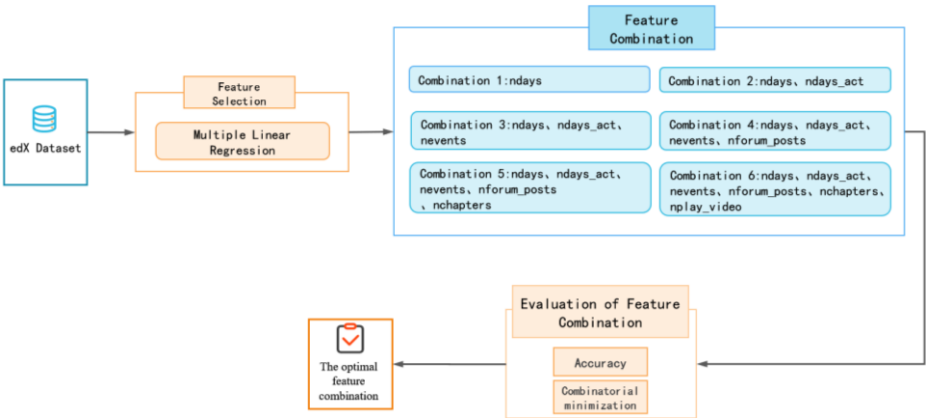


**Fig. 3.** Feature selection frameworks

# 4 Comparison of algorithm results

## 4.1 Baseline model comparison

We use Precision, Recall and F1-Score as evaluation indicators, and uses five machine learning models to predict six different scale datasets. Finally, the prediction results of these models are compared and analyzed.

In the multi-classification problem, the model needs to calculate the evaluation index of each class, and then calculate the overall index of the model according to a certain rule. Assuming that for a certain class, the data belonging to the class is called the positive class, otherwise it is called the negative class, then Precision represents the proportion of data that is correctly predicted in the dataset that is predicted to be the positive class. The formula for Precision is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{9}$$

where TP (True Positive) is the number of positive samples that are predicted as a positive class, and FP (False Positive) is the number of negative samples that are predicted as a positive class. Recall is the proportion of the positive data predicted accurately. The formula for Recall is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{10}$$

where TN (True Negative) represents the number of negative samples that are correctly predicted as negative classes; FN (False Negative) represents the number of positive samples that are incorrectly predicted as negative classes.

F1-Score is an overall assessment of Precision and Recall, its formula is as follows:

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{11}$$

The model evaluation results of the five models are shown in Table 1.

**Table 1.** Model evaluation results of 5 models

| Model Name | Model Evaluation Value | The volume of data | | | | | |
|---|---|---|---|---|---|---|---|
| | | ***8000*** | ***16000*** | ***24000*** | ***32000*** | ***40000*** | ***48000*** |
| LR | Precision | 0.85598 | 0.83831 | 0.84769 | 0.85713 | 0.85806 | 0.86873 |
| | Recall | 0.91250 | 0.90250 | 0.90813 | 0.91344 | 0.91375 | 0.91969 |
| | F1-Score | 0.88334 | 0.86921 | 0.87647 | 0.88370 | 0.88409 | 0.89275 |
| DT | Precision | 0.86988 | 0.87998 | 0.88126 | 0.89077 | 0.87945 | 0.89484 |
| | Recall | 0.87438 | 0.87688 | 0.88313 | 0.89188 | 0.87975 | 0.89219 |
| | F1-Score | 0.87202 | 0.87840 | 0.88217 | 0.89129 | 0.87976 | 0.89348 |
| NB | Precision | 0.89038 | 0.86650 | 0.87901 | 0.88514 | 0.88651 | 0.89529 |
| | Recall | 0.91875 | 0.90156 | 0.90750 | 0.91000 | 0.90525 | 0.90979 |
| | F1-Score | 0.90278 | 0.88242 | 0.89127 | 0.89584 | 0.89418 | 0.90099 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| SVM | Precision | 0.77834 | 0.82017 | 0.77691 | 0.79346 | 0.80983 | 0.84587 |
| | Recall | 0.87750 | 0.86156 | 0.87625 | 0.88406 | 0.88450 | 0.89604 |
| | F1-Score | 0.82495 | 0.83793 | 0.82318 | 0.83002 | 0.83172 | 0.86993 |
| DNN | Precision | 0.98813 | 0.99557 | 0.95894 | 0.99752 | 0.96154 | 0.99393 |
| | Recall | 0.99000 | 0.95145 | 0.99978 | 0.96564 | 0.97968 | 0.98531 |
| | F1-Score | 0.98894 | 0.97301 | 0.97893 | 0.98132 | 0.97053 | 0.98960 |

After visualizing the data in Table 1, the comparison of the prediction performance of different algorithm models can be seen more clearly, as shown in Figure 4.

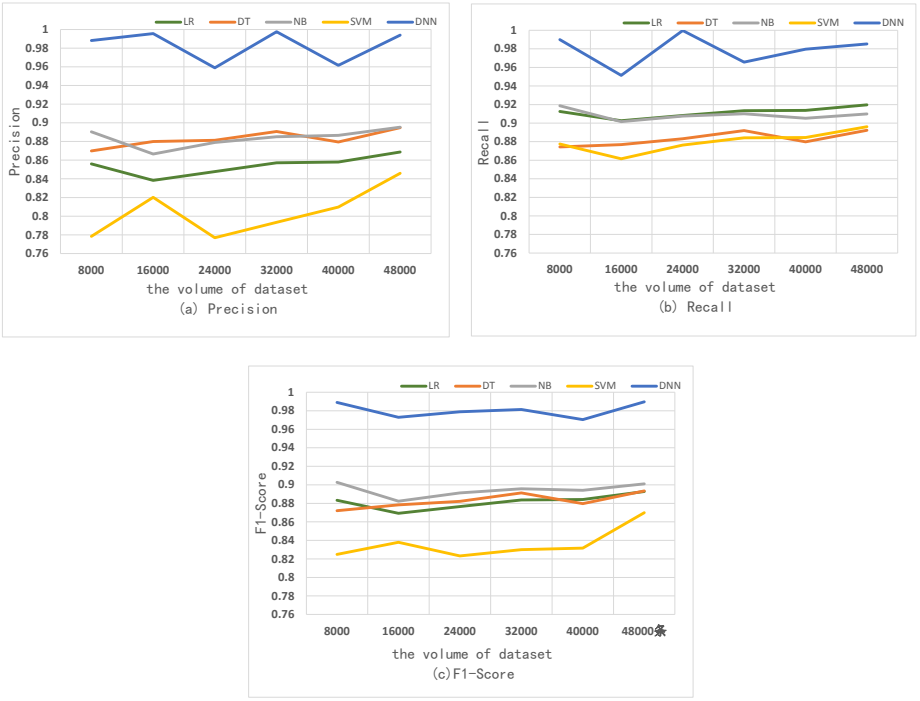


**Fig. 4.** Evaluation results of five models

It can be found from Table 2 and Figure 4 that the evaluation index of the DNN model is higher than the other four models. F1-Score of the DNN model is about 10% higher than the other models, which reflects the strong generalization ability of the DNN model. LR model has relatively good performance when the amount of data is very small. However, as the data size increases, the performance of LR model also shows a small increase. When the data size reaches 48,000, F1- Score reached the highest value (0.89275). In general, as the size of dataset increases, the fitting of LR model to the data set will be better, but the performance improvement is limited. This is because LR model is a linear classifier, and the accuracy of multi-classification is not

high. Some special sample values will greatly interfere with the model, so it is difficult for the model to achieve better performance in large-scale data.

For large-scale datasets, the performance of the DT model is the best, and its F1-Score is 0.89348. As the volume of dataset increases, the performance of the DT model will increase slightly. When dataset size reaches 48,000, its performance will be optimal. DT will produce prediction biases for data sets with inconsistent samples of each class, and at the same time lack of analysis of the correlation between features, so it is difficult to achieve similar effects to the DNN model.

The performance of the NB model changes slightly with the increase of the dataset size. When the data size is 8000, the model performance is the best, and F1-Score is 0.90278. NB model is suitable for multi-classification problems, and its classification efficiency is stable. In the case of specifying the result class, NB model will assume that the feature attributes are independent each other. However, in this experiment, there is a significant correlation between features, so it is difficult for the performance of NB model to exceed 0.9.

The performance of the SVM model is the worst for small-scale data, but as the size of the data set increases, its performance is gradually improving. When the scale of datasets reaches 48,000, F1-Score is 0.86993. SVM has good generalization capabilities. Moreover, it can also solve the classification problem of high-dimensional datasets, but it is essentially a kind of algorithm for binary classification problem. For multi-classification problems, SVM needs to construct multiple binary classification SVMs and combine them, which cannot be directly support multi-classification problems.

With the increase of the data set size, the Precision of the DNN model fluctuates, reaching the lowest value at 24000 data and the highest value at 32000 data. It can be seen that with the increase of the dataset size, there are more similar samples with the previously samples predicted wrongly, and when the samples are large enough, the model can fit them well. Recall of the DNN model shows the opposite trend to Precision. When Precision is relatively high, Recall decreases. Recall reaches the lowest when the dataset volume reaches 16000, and the highest when the dataset volume reaches 24000. The reason for the fluctuation of Recall is that when the dataset volume is 16000, there are more special samples and more difficult samples. In general, compared with the previous four machine learning models, DNN has no major defects in the model, and the requirements for the data set are not as harsh as the previous four models. In general, compared with the previous four machine learning models, DNN model has smaller defects and less requirements for data sets.

The research results show that due to the diversity of the teaching process and the environment, the original dataset of MOOC is complicated, and there are many problems such as uneven distribution and high dispersion of dataset, which leads to poor prediction performance of traditional machine learning methods. Compared with the four shallow machine learning methods of Logistic Regression, Decision Tree, Naive Bayes and SVM, DNN with multiple hidden layers has excellent feature learning capabilities, and the learned features are more relevant to data, which is conducive to prediction. Comprehensively considering the performance and evaluation indicators of the

models, we believe that DNN can achieve the best effect in predicting academic performance of MOOC learners. However, DNN method also has some shortcomings. Because the training process of DNN model is inexplicable, it is not conducive to the analysis of attribution problems in educational practice.

### 4.2 Performance comparison of three behavioral Feature Selection methods

In this study, SFS, SBS and Multiple Linear Regression are used to analyze the six features, from which the optimal feature combination is selected, and the prediction performances of the models under different feature selection methods are analyzed. In the experiment, the optimal feature combinations obtained by three feature selection methods are shown in Table 2.

**Table 2.** Optimal feature combinations of three feature selection methods

| Method Name | Features | | | | | |
|---|---|---|---|---|---|---|
| | ***nevents*** | ***ndays_act*** | ***nplay_video*** | ***nchapters*** | ***ndays*** | ***nforum_posts*** |
| SFS | | √ | | | √ | |
| SBS | | √ | | | √ | |
| Multiple Linear Regression | | √ | | √ | | |

Since SFS and SBS only differ in the direction of selection process, their optimal feature combinations are the same, which are the interaction times of the course and learning days; the optimal feature combination of the Multiple Linear Regression method is the interaction times of the course and the number of chapters studied.

Using the above feature combinations, this research processes 48,000 data and obtains two datasets, which are the dataset composed of SFS/SBS optimal feature combination data and learner performance classification; the dataset composed of Multiple Linear Regression optimal feature combination data and learner performance classification. Five models are used to train and predict the two datasets, and the model evaluation results are shown in Table 3.

**Table 3.** Model evaluation results of the five models under feature selection

| Feature Selection | Algorithm name | Test Set | | |
|---|---|---|---|---|
| | | ***Precision*** | ***Recall*** | ***F1-Score*** |
| SFS/SBS | LR | 0.86113 | 0.91604 | 0.88733 |
| | DT | 0.88312 | 0.88542 | 0.88414 |
| | NB | 0.90853 | 0.90354 | 0.90057 |
| | SVM | 0.86703 | 0.91604 | 0.88998 |
| | DNN | 0.99382 | 0.98595 | 0.98987 |
| Multiple Linear Regression | LR | 0.86499 | 0.91563 | 0.88891 |
| | DT | 0.89924 | 0.91281 | 0.90460 |
| | NB | 0.90744 | 0.90323 | 0.90074 |
| | SVM | 0.86831 | 0.91740 | 0.89122 |
| | DNN | 0.98806 | 0.99704 | 0.99253 |

F1-Score is a comprehensive evaluation of precision and recall. Using data visualization to display F1-Score of each algorithm model, we can more clearly observe the prediction performance comparison of each algorithm before and after feature selection, as shown in Figure 5.

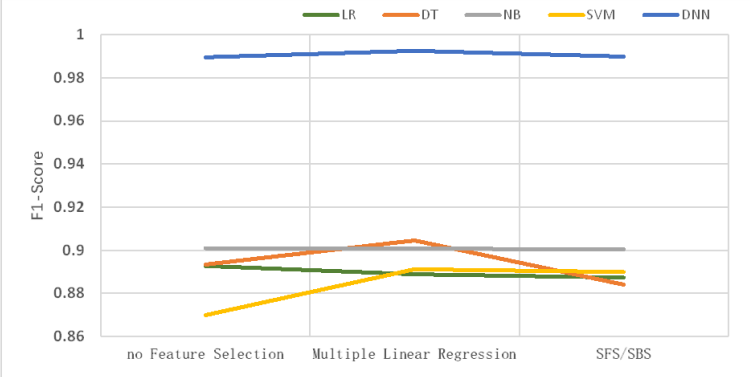


**Fig. 5.** F1-SCORE line chart of five models under different conditions

Combined with Table 3 and Figure 5, it can be found that the model evaluation index of DNN model is about 10% higher than that of the other four models, regardless of whether feature selection method is adopted.

Comparing F1-score of Multiple Linear Regression feature selection and no feature selection, it can be found that the performances of DNN, DT and SVM with Multiple Linear Regression feature selection have been improved, among which SVM has the most performance improvement; for DT and Nb models, after Multiple Linear Regression feature selection, the performances of both models have slightly decreased. The results show that the use of Multiple Linear Regression feature selection for the three models of DNN, DT, and SVM can effectively improve Precision of the models, simplify the models, reduce the running time of the models, eliminate low-correlation or redundant features, and make it easier for researchers to interpret the data; and LR and NB models are not suitable for Multiple Linear Regression feature selection methods, as the number of features increases, the performance of these two models will be better.

Comparing using SFS / SBS feature selection and no feature selection, we find that the performance of SVM model increases, while that of DNN model remains unchanged, and that of LR, DT and NB decreases, of which the degradations of LR and DT models are particularly significant. The results show that Multiple Linear Regression feature selection is better than SFS / SBS combined with Logistic Regression model. Both SFS and SBS belong to greedy algorithm, which selects the features with the best evaluation value. When selecting features, some of them are often ignored due to the dependence between features. So SFS and SBS are not suitable for this study. Finally, this study determined a new performance prediction model, which uses multiple linear regression for feature selection, and then inputs the selected features into the DNN model to predict learners' grades.

To further validate the performance of our model. Among the relevant literature in the field of MOOC prediction, two machine learning-based models and one deep learning-based model were selected to compare the results with those of our model in this study. The three models selected all use the edX dataset. The three MOOC dropout prediction models are Logistic Regression [26], Nonlinear SVM [27], and LSTM [28]. The results of our model compared with the models of related studies are shown in Table 4.

**Table 4.** Comparison of model results among related studies

| Method Name | Precision | Recall | F1-Score |
|---|---|---|---|
| Logistic Regression [26] | 0.78236 | 0.80578 | 0.79390 |
| Nonlinear SVM [27] | 0.76754 | 0.74821 | 0.75775 |
| LSTM [28] | 0.97434 | 0.98589 | 0.98008 |
| Our Model | 0.98806 | 0.99704 | 0.99253 |

Our model outperforms the previous machine learning and deep learning MOOC dropout prediction methods mentioned above in all aspects. The study demonstrates that our model is effective in further improving the accuracy of MOOC learning performance prediction.

The final experimental results show that DNN algorithm combined with feature selection and no feature selection can improve the precision of the model, thereby more effectively predicting the academic performance of MOOC learners.

## 5 Future research and challenges

Machine learning algorithm has been applied to educational scenarios to achieve better prediction results. However, there are some common problems that arise in actual research that bring significant challenges to machine learning algorithms for learning prediction.

MOOC learning sample data usually has the characteristics of imbalance, a large amount of missing sample data, and data discretization. This is because MOOC learners can customize learning according to their own learning progress, learning methods, and learning time, and it includes a large number of registered students who are not really learning. The effectiveness of machine learning algorithms relies on a large number of valid sample data. Significant differences in the observed sample data will affect the generalization ability of the model in the training step and also affect the classification accuracy.

Availability of publicly accessible datasets for MOOCs is also a major challenge for MOOC learning. MOOC platforms tend to be reluctant to release data due to confidentiality and privacy concerns. Public de-identification information is also limited to system-generated events (clickstream data), and most user-provided data is omitted. The lack of user-provided data may hamper successful prediction of the reasons behind dropouts.

Compared with the prediction results, the MOOC learning prediction problem pays more attention to the attribution of the problem. Although the performance of the deep learning algorithm is much better than that of the traditional machine learning algorithm, the inexplicable of deep learning algorithm is the main problem that restricts its application in the practical problem of MOOC prediction. How to reasonably combine the high performance of deep learning algorithm with the statistical method based on feature selection is an important problem to be solved in the future.

The research on MOOC student dropout and performance prediction is still in a continuous developing stage. In order to establish a well-defined and more accurate prediction solution, and to identify, understand and explain high-risk students as soon as possible, there are still a lot of research challenges to be solved in depth.

## 6 Conclusion

This study aims at the high dropout rate in MOOC learning. From the perspective of multi-category performance prediction, we use multiple linear regression to select key behavior data from the edX data set to construct feature combinations, build prediction models with five machine learning algorithms, and analyze the performance of different models. Finally, a new performance prediction model is determined. The experimental results show that our model has the best prediction performance on different feature combinations. The overall prediction accuracy is high (up to more than 97%), and other machine learning algorithms can perform well under the effective feature combinations. In the future research, we will continue to carry out the research on the integrated method of key features extraction and model construction, so as to ensure high accuracy of prediction and simplicity of the model. In general, the research on learning prediction-related issues in large-scale online education is still in its infancy, and the space for research is still very broad. It is necessary to establish clearer and more accurate prediction to identify, understand and explain the reasons for students dropping out. And it could help the curriculum designers and teachers improve course content and conduct effective educational interventions timely.

## 7 Acknowledgment

This material is based upon work supported by Industry-university cooperative education project (202101045002) and the Science and Technology Research Project of Henan Province (“Research on text classification model and application for massive open online courses”).

## 8 References

[1] Wang T (2014). Developing an assessment-centered e-Learning system for improving student learning effectiveness. Comput. Educ, 2014, 73:189-203. https://doi.org/10.1016/j.compedu.2013.12.002

[2] Gutl C, Rizzardini R, Chang V, et al. (2014). Attrition in MOOC: Lessons Learned from Drop-Out Students. In the International Workshop on Learning Technology for Education in Cloud. Springer, Cham, 2014. https://doi.org/10.1007/978-3-319-10671-7_4
[3] Jordan K (2015). Massive open online course completion rates revisited: Assessment, length and attrition. International Review of Research in Open and Distance Learning, 2015, 16(3). https://doi.org/10.19173/irrodl.v16i3.2112
[4] Shingari I, Kumar D, Khetan M (2017). A review of applications of data mining techniques for prediction of students' performance in higher education. JOURNAL OF STATISTICS & MANAGEMENT SYSTEMS, 2017, 20(4): 713-722. https://doi.org/10.1080/09720510.2017.1395191
[5] C. Robinson, M. Yeomans, J. Reich, et al. (2016). Gehlbach.Forecasting student achievement in MOOCs with natural language processing. in Proc. 6 th Int. Conf. Learning Analytics & Knowledge, Edinburgh, UK, 2016, pp. 383–387. https://doi.org/10.1145/2883851.2883932
[6] Dalipi F, Imran A S and Kastrati Z (2018). MOOC dropout prediction using machine learning techniques: Review and research challenges. In 2018 IEEE Global Engineering Education Conference (EDUCON) 1007-1014. https://doi.org/10.1109/EDUCON.2018.8363340
[7] Evangelista, E., A Hybrid Machine Learning Framework for Predicting Students' Performance in Virtual Learning Environment. International journal of emerging technologies in learning, 2021. 16(24): p. 255-272. https://doi.org/10.3991/ijet.v16i24.26151
[8] Tomasevic N, Gvozdenovic N, Vranes S (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & education, 2020, 143(Jan.):103676.1-103676.18. https://doi.org/10.1016/j.compedu.2019.103676
[9] Goel S, Sabitha A S, Choudhury T, et al (2019). Analytical Analysis of Learners' Dropout Rate with Data Mining Techniques: Proceedings of ICETEAS 2018// Emerging Trends in Expert Applications and Security. 2019. https://doi.org/10.1007/978-981-13-2285-3_69
[10] Böttcher A, Thurner V, Häfner T (2020). Applying Data Analysis to Identify Early Indicators for Potential Risk of Dropout in CS Students. In 2020 IEEE Global Engineering Education Conference (EDUCON) 827-836. https://doi.org/10.1109/EDUCON45650.2020.9125378
[11] Zhang G Q (2018). Research on Dropout Rate Prediction of Massive Open Online Courses. Wuhan: Central China Normal University, 2018 (in Chinese).
[12] Vitiello M, Walk S, Helic D, et al. (2018). User Behavioral Patterns and Early Dropouts Detection: Improved Users Profiling through Analysis of Successive Offering of MOOC. Journal of Universal Computer Science, 2018, 24(8):1131-1150.
[13] Alamri A, Alshehri M, Cristea A, et al. (2019). Predicting MOOCs Dropout Using Only Two Easily Obtainable Features from the First Week's Activities. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-22244-4_20
[14] Qiu L, Liu Y (2018). An Integrated Framework With Feature Selection for Dropout Prediction in Massive Open Online Courses. IEEE Access, 2018, PP (99):1-1.
[15] Imran A, Dalipi F, Kastrati Z. (2019). Predicting Student Dropout in a MOOC: An Evaluation of a Deep Neural Network Model. In the 2019 5th International Conference, 2019. https://doi.org/10.1145/3330482.3330514
[16] Qiu L, Liu Y, Hu Q, et al. (2018). Student dropout prediction in massive open online courses by convolutional neural networks. Soft Computing, 2018, 23. https://doi.org/10.1007/s00500-018-3581-3
[17] Tang C Ouyang Y, Rong W, et al. (2018). Time Series Model for Predicting Dropout in Massive Open Online Courses. 2018. https://doi.org/10.1007/978-3-319-93846-2_66

[18] Qi Q, Liu Y, Wu F, et al. (2018). [ACM Press ACM Turing Celebration Conference - China - Shanghai, China (2018.05.18-2018.05.20)] Proceedings of ACM Turing Celebration Conference - China on, - TURC '18 - Temporal models for personalized grade prediction in massive open online courses. Acm International Conference Proceeding, 2018:67-72.
[19] Bognár, L., T. Fauszt and G.Z. Nagy, Analysis of Conditions for Reliable Predictions by Moodle Machine Learning Models. International journal of emerging technologies in learning, 2021. 16(6): p. 106-121. https://doi.org/10.3991/ijet.v16i06.18347
[20] Raschka, S. (2015). Python Machine Learning. Packt Publishing Ltd., Birmingham.
[21] Chen X Y (2020). Research on Machine Learning Algorithm Based on Big Data Background. Computer products and circulation.2020, (3):85 (in Chinese).
[22] Kim H J, Kim J, et al. (2018). Towards perfect text classification with Wikipedia-based semantic Nave Bayes learning. Neurocomputing, 2018, 315(NOV.13):128-134. https://doi.org/10.1016/j.neucom.2018.07.002
[23] Tian Y, Shi Y, Liu X (2012). Recent advances on support vector machines research. Technological & Economic Development of Economy, 2012, 18(1):5-33. https://doi.org/10.3846/20294913.2012.661205
[24] Muniasamy, A. and A. Alasiry, Deep Learning: The Impact on Future eLearning. International journal of emerging technologies in learning, 2020. 15(1): p. 188-199. https://doi.org/10.3991/ijet.v15i01.11435
[25] Jurgen Schmidhuber (2015). Deep learning in neural networks: An overview. NEURAL NETWORKS, 2015, 61:85-117. https://doi.org/10.1016/j.neunet.2014.09.003
[26] He J, Bailey J, Rubinstein B. et al. (2015). Identifying At-Risk Students in Massive Open Online Courses. In the Association for the Advance of Artificial Intelligence, 2015.
[27] Amnueypornsakul B, Bhat S, Chinprutthiwong P. (2014). Predicting Attrition Along the Way: The UIUC Model. In the 2014 Conference on Empirical Methods in Natural Language Processing, 2014. https://doi.org/10.3115/v1/W14-4110
[28] Fei M, Yeung D. (2015). Temporal Models for Predicting Student Dropout in Massive Open Online Courses. In the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, 256-263. https://doi.org/10.1109/ICDMW.2015.174

## 9 Authors

**Huichao Mi** is a lecturer at College of Computer and Information Engineering, Henan University of Economics and Law, China. She received her master's degree from Harbin Engineering University in 2005. Her research focuses on education data mining and analysis of online learning processes.

**Zhanghao Gao** is a postgraduate student at College of Computer and Information Engineering, Henan University of Economics and Law, China. His research interests include big data mining and analysis in education.

**Qiaorong Zhang** is an associate professor at College of Computer and Information Engineering, Henan University of Economics and Law, China. Her research interests include learning analysis technology, education data mining, and AI in education.

**Yafeng Zheng** is an associate professor and M.S. supervisor at Henan University of Economics and Law. Her expertise is learning analysis, data mining and big data visualization in education.

Article submitted 2021-07-19. Resubmitted 2021-12-28. Final acceptance 2022-02-10. Final version published as submitted by the authors.