Scientific
Research
Publishing

Student Performance Prediction via Attention-Based Multi-Layer Long-Short Term Memory

Yanqing Xie

Jinan University, Guangzhou, China
Email: ice13042047710@163.com

Abstract

Online education has attracted a large number of students in recent years, because it breaks through the limitations of time and space and makes high-quality education at your fingertips. The method of predicting student performance is to analyze and predict the student's final performance by collecting demographic data such as the student's gender, age, and highest education level, and clickstream data generated when students interact with VLE in different types of specific courses, which are widely used in online education platforms. This article proposes a model to predict student performance via Attention-based Multi-layer LSTM (AML), which combines student demographic data and clickstream data for comprehensive analysis. We hope that we can obtain a higher prediction accuracy as soon as possible to provide timely intervention. The results show that the proposed model can improve the accuracy of 0.52% - 0.85% and the F1 score of 0.89% - 2.30% on the four-class classification task as well as the accuracy of 0.15% - 0.97% and the F1 score of 0.21% - 2.77% on the binary classification task from week 5 to week 25.

Keywords

Data Science Applications in Education, Distance Education and Online Learning, LSTM

1. Introduction

Online education is a new way of education in the Internet era [1]. Online education platforms, e.g., MOOC, propose massive high-quality learning resources, including lots of classroom videos, exercises and assessments of many

world-renowned schools [2]. In online education platform, students can obtain those courses that suit themselves with low pay even no pay, which provides convenience for students' independent learning [3]. The dependence of online education platforms has injected new vitality into the traditional education industry [4]. Different from the traditional teacher-student face-to-face interaction teaching method, online education method is no longer limited by teaching venues and teachers' conditions. It expands the number of students from finite to infinite and also teachers can ensure the quality of teaching by recording instructional videos and check it in advance [5]. Since online education possesses many of the above advantages, it is improving rapidly and has attracting lots of students [6].

Online education aims to construct an education platform which is open and free for everyone [7]. It hopes to attract more students who can study themselves leading by interest without the limitation of area or time, helping bring learning into everyone's daily life [8].

In 2020, affected by the global epidemic, all schools will change their teaching methods from traditional offline teaching methods to online teaching methods [9]. As a result, a large number of students have poured into online education platforms. This is a test of the online education platform and an opportunity to improve the online education mechanism. How to ensure the quality of each student's learning when a large number of students study a course at the same time is one of the key issues that online education platforms need to consider [10].

We hope to obtain real-time information on the learning status of students so that teachers can intervene in the learning status of students in time and help students better master the content of this course [11]. In order to achieve this goal, we consider establishing a student performance prediction system to evaluate student performance [12].

We collect student data on online education platforms, including student demographic data and student clickstream data, to predict students' final performance [13]. The demographic data of a student includes background information such as each student's age, gender, and highest education level. The student's clickstream data is the interaction log between the student and Virtual Learning Environment (VLE), which is divided into 20 categories, such as web-page click, forum click, quiz attempt and so on [14].

In this article, we propose an Attention-based Multi-layer LSTM (AML) model to analyze the input student demographic data and student clickstream data. We make predictions every 5 weeks and record the accuracy, precision, recall and F1 score of the test set. We hope to make more accurate predictions of students' final performance as soon as possible, so we take the model training and testing every five weeks from week 0 to week 25.

In order to be able to identify students with a tendency to drop out, we divide students' performance into two categories: withdrawn and pass [15]. In order to be able to predict the students' final performance, we divide the students' per-

formance into four categories: withdrawn, fail, pass, and distinction. We separately train and test the two classification methods, and record their model evaluation results. The main contributions of this work are as follow.

- We propose an Attention-based Multi-LSTM model to predict students' final performance. This model utilizes students' demographic data and students' clickstream data, which makes the model can predict on the situation of cold start.
- We did not distinguish between the types of courses when training the model, which made the model perform well in course transfer.

This paper is organized as follows. Section 2 introduces the related work of student performance prediction methods. Section 3 introduces some mathematical notations and formally defines the given problem. Section 4 introduces the model we propose. Section 5 introduces the experiments and results of our work. Section 6 introduces the conclusions of this paper.

2. Related Work

With the development of the online education industry, more and more students have poured into online education platforms [16]. Many educators began to consider how to ensure the quality of online learning for each student when there are a large number of students in a course [17]. Therefore, the concept of student performance prediction system came into being. Most of the input data of the student performance prediction model comes from the back-end data of various online education platforms, which is private.

Many domestic and foreign scholars have been invited to build student performance prediction systems for online education platforms. They can use the private data of online education platforms to build student performance prediction models. References [18] uses the student event stream sequence, such as whether the student submits an assignment at a certain time, whether the student asks a question at a certain time, whether the student completes the exam at a certain time, etc., to build a GritNet model to predict the student's final performance. Modern data mining and machine learning techniques [19] are used for predicting student performance in small student cohorts. References [20] compare the effect of supervised learning algorithms for student performance prediction. References [21] build a decision tree-based algorithm, Logistic Model Trees (LMT) to learn the intrinsic relationship between the identified feature, which are identification of academic and socio-economic features, and students' academic grades. References [22] apply a transfer learning methodology using deep learning and traditional modeling techniques to study high and low representations of unproductive persistence. References [23] extend the deep knowledge tracing model, which is a state-of-the-art sequential model for knowledge tracing, to consider forgetting by incorporating multiple types of information related to forgetting. References [24] propose an attention-based graph convolutional networks model for student's performance prediction. References [25] de-

sign two strategies under Exercise-Enhanced Recurrent Neural Network (EERNN), *i.e.*, EERNNM with Markov property and EERNNA with Attention mechanism for student performance prediction. References [26] propose a method that combines the cluster-based LDA and ANN for student performance prediction and comments evaluation. References [27] establish a model based on discriminative feature selection.

Of course, there are also cases where open datasets are used to predict student performance. For example, the OULA [14] dataset is a common dataset used for student performance prediction research. Many scholars at home and abroad have carried out analysis based on this dataset. Scholars try to use classic machine learning models such as Logistic Regression, Decision Tree [28], linear SVM [29] and other methods to analyze the dynamic impact of demographic characteristics on academic outcomes in the online learning environment. As the effectiveness of deep learning methods is seen by more and more people, some scholars try to use deep learning methods to predict student performance. References [30] use a multi-layer Artificial Neural Network model to predict dropout. LSTM model is proved to be effective in many fields of artificial intelligence as shown in [31] [32] [33] [34] [35]. Some scholars apply multi-layer LSTM models [36] [37] to predict student performance. References [38] investigate ensemble methods, deep learning and regression techniques for predicting student dropout and final result in MOOCs. References [39] propose General Unary Hypotheses Automaton (GUHA) and Markov chain-based analysis to analyze the impact of student activities on the dropout rate.

3. Problem Statement

In this section, we introduce some mathematical notations and formally define the given problem.

Since we need to make a timely assessment of the learning status of each student in the online course, we propose an Attention-based Multi-layer LSTM model for real-time student performance prediction. The mathematical definitions of some concepts involved in the model are as follows.

Suppose that we have m courses, the j_{th} course is denoted as c_j, the set of courses is denoted as $C = \{c_1, c_2, \cdots, c_j, \cdots, c_m\}$. Suppose there be n students enrolled in at least one course, the i_{th} student is denoted as s_i, the set of students is denoted as $S = \{s_1, s_2, \cdots, s_i, \cdots, s_n\}$. For each student s_i, the online education platform will collect his gender, age, highest education level and other background information as his demographic data. There are eight items of background information. The demographic data of student s_i is denoted as the vector d_i. We encode the category data in d_i, and the encoded demographic vector of student s_i is $\bar{d}_i$. Thus, the demographic data set of all students is denoted as $D = \{\bar{d}_1, \bar{d}_2, \cdots, \bar{d}_i, \cdots, \bar{d}_n\}$. Suppose that the course c_j has a total of K weeks, the clickstream data sequence vector of student s_i in the k_{th} week of the course c_j is denoted as q_{ij}^k. Thus, the clickstream dataset of student s_i in

the course c_j is denoted as $Q_{ij} = \left\{ q_{ij}^1, q_{ij}^2, \cdots, q_{ij}^k, \cdots, q_{ij}^K \right\}$. The actual outcome of student s_i in the course c_j is denoted as o_{ij}, which has p possibilities. When we perform a binary prediction, the possible result of o_{ij} is *pass* or *fail*. When we perform a four-class prediction, the possible result of o_{ij} is *distinction*, *pass*, *fail* or *withdrawn*.

According to the definition given above, we build a model $f(\cdot)$ to predict student performance, and the obtained prediction result is denoted as $\bar{o}_{ij}$. The model obtains the best parameters θ through learning, and then substitutes θ into the model to obtain the predicted outcomes. The learning process of the model is shown as Equation (1):

$$T(D, Q, O, f(\cdot)) \to \theta \qquad (1)$$

where $T(\cdot)$ means the learning process of the model, D means the demographic data of students, Q means the clickstream data of students, O means the actual outcomes of students, $f(\cdot)$ means the proposed model, θ means the trained model parameters.

The prediction process of the model is shown as Equation (2).

$$f(D, Q \mid \theta) \to \bar{Q} \qquad (2)$$

where $f(\cdot \mid \theta)$ means the trained model, D means the demographic data of students, Q means the clickstream data of students, θ means the trained model parameters, $\bar{Q}$ means the predicted outcomes of students.

Now, we gave an introduction to all the definitions in the student performance prediction model, and then we will introduce our proposed model $f(\cdot \mid \theta)$.

4. Proposed Model

Our goal is to build a model that can predict the performance of any student in any period of any course. We hope that this model has universal applicability and can be transferred to any course instead of only predicting a single course. We hope that this model can predict the course from before the start of the course, that is, week 0, to the end of the course at any time, not only after the start of the course. Especially in the early and middle of the course, we hope to obtain more accurate forecasts as soon as possible so that the online education platform can issue early warnings in time to urge students to adjust their learning status. We hope that this model can predict the individual performance of any student in the course, not just predict all students in the entire course. In order to achieve the above-mentioned purpose, we propose an Attention-based Multi-layer LSTM (AML) model, whose structure is shown in **Figure 1**, and its specific description is as follows:

In order to obtain a reliable prediction of student results, we consider using student clickstream data, which is inherently a time sequence. Time sequence refers to the input sequence in which the data has a contextual relationship on the time axis, that is, the output state generated at the current time point is not only related to the input data at the current time point, but also related to the
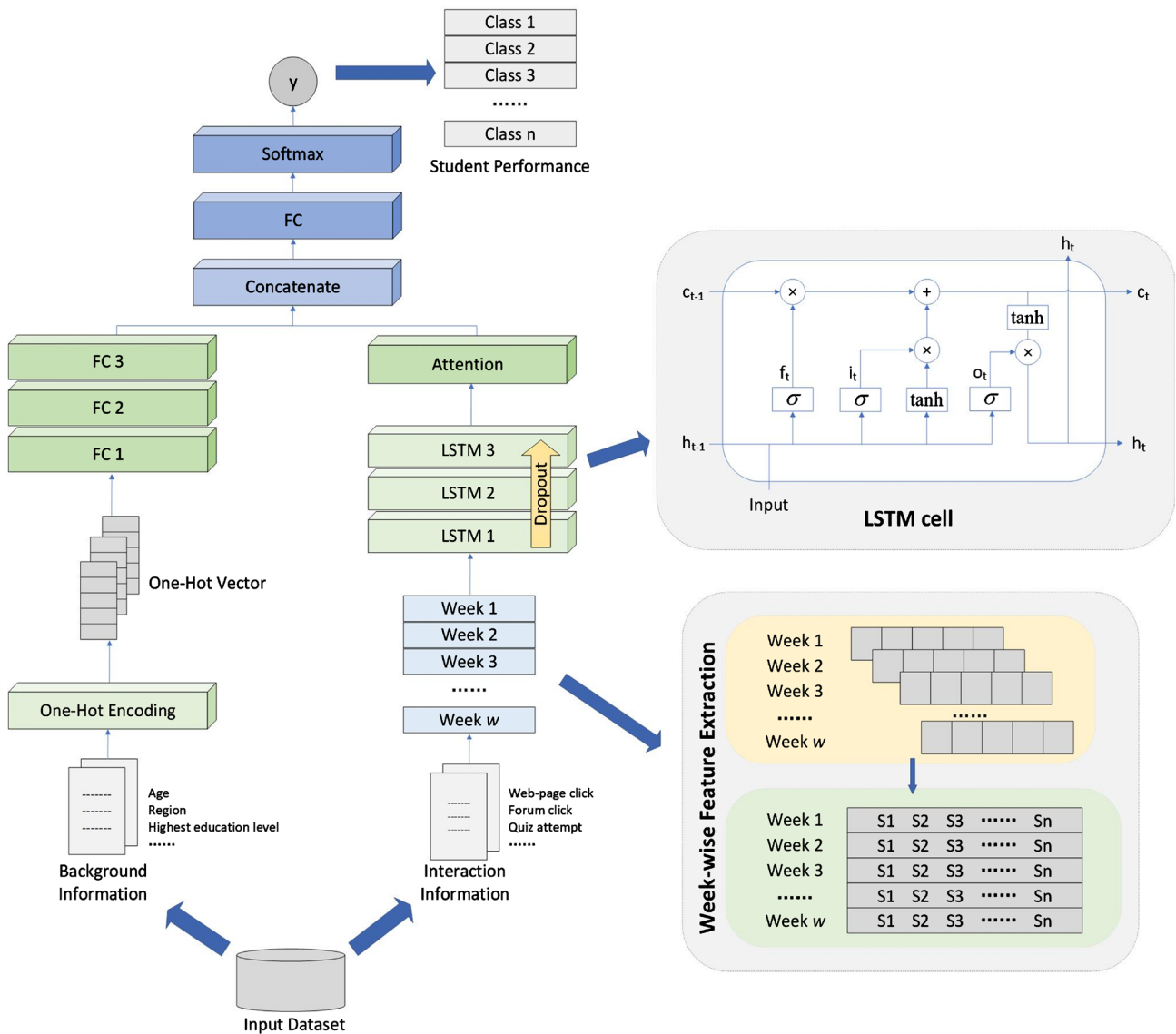
Figure 1. The proposed Attention-based Multi-layer LSTM model (AML).

data input before, and will affect the subsequent time point Output status. Text, voice, etc. are all time sequence data. Student clickstream data is divided into many different categories according to the content of interaction between students and VLE platform. If we simply record the number of interactions between each student and the VLE platform in days or weeks, we ignore the fact that different types of interactions have different effects on student performance. Therefore, we keep the students' clickstream data types and input data into our model on a weekly basis. We utilize the LSTM structure to process the input student clickstream data. LSTM is an effective structure for processing time sequence shown as Equation (3). LSTM selects and memorizes the input information through three gating units, so that the model only remembers the key information, thereby reducing the burden of memory, so it can solve the problem of long-term dependence.

$$I_t = \sigma\left(X_t W_i + H_{t-1} W_i + b_i\right)$$

$$F_t = \sigma\left(X_t W_f + H_{t-1} W_f + b_f\right)$$

$$O_t = \sigma\left(X_t W_o + H_{t-1} W_o + b_o\right) \tag{3}$$

$$\tilde{C}_t = \tanh\left(X_t W_c + H_{t-1} W_c + b_c\right)$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t$$

where I_t, F_t, O_t, H_t and C_t mean input gate vector, forget gate vector, output gate vector, LSTM output unit vector and memory cell vector respectively, W and b mean weight matrix and bias, σ and $\tanh$ mean functions.

If we want to get a better model prediction effect, it is not enough to just use student clickstream data. When the number of weeks of the course is small, the amount of click-stream data of students is small, and the prediction effect of the model is not satisfactory. In particular, when the course is in week 0, that is, when the course has just started, the model cannot receive student clickstream data. Therefore, we introduce the demo-graphic data of students into the model, that is, personal background data of students. Student demographic data is the data collected by the online education platform when students register, which is unique. Student demographic data includes two types: sequence data and category data. We perform one-hot encoding on the category data in the student demographic data, and then concatenate the encoded vector with the sequence data to obtain the processed student demographic data. We input the processed student demographic data into a fully connected layer, then splice the output of the fully connected layer with the output of the LSTM structure, and input the spliced vector into the softmax layer. The softmax layer is a fully connected layer that uses the softmax function to classify. It calculates the probability of each classification so that the class with the largest probability is the predicted classification of the student s_i on the course c_j. The softmax function is shown as Equation (4).

$$S_i = \frac{e^i}{\sum_j e^i} \tag{4}$$

In order to obtain better prediction results, we change the number of fully connected layers and LSTM layers in the model from one layer to multiple layers, and perform multiple tests to obtain the best number of layers. On the basis of the above model, we consider adding an attention mechanism to further improve the prediction performance of the model. The attention mechanism is often used in machine translation tasks in Natural Language Processing. It changes the influence of different content by adding a weight matrix to the input vector, so that the weight of factors that have a greater impact on the student's performance prediction results is increased, and the weight of factors that have less impact on the results is reduced, so as to improve the prediction effect of the model. The attention mechanism is shown as Equation (5).

$$u_{it} = \tanh\left(W_w h_{it} + b_w\right)$$

$$\alpha_{it} = \frac{\exp\left(u_{it}^{\mathrm{T}} u_w\right)}{\sum_t \exp\left(u_{it}^{\mathrm{T}} u_w\right)} \tag{5}$$

$$s_i = \sum_t \alpha_{it} h_{it}$$

where h_{it} means the hidden vector of the student S_i in time t, W_w and b_w mean weight matrix and bias, which are initialed randomly.

As the number of weeks of each course varies, we uniformly take the student data of the first 25 weeks of the course as the input data of the model. We output and record the prediction results every five weeks. Next, we will introduce the experimental process of this article.

5. Experiments

In this section, we conducted some experiments to verify the effect of our proposed model. First, we introduce the dataset used in the paper and our dataset processing scheme. Second, we describe the experimental settings of the proposed model. Finally, we show the experimental comparison results of the proposed model and the baseline model on two classification tasks and a student performance prediction task for the specific course, as well as perform corresponding analysis.

5.1. Dataset

The Open University Learning Analytics (OULA) [14] dataset contains a series of online education related data provided by online education platforms such as student demographic data, student clickstream data, and course data. Student demographic data is background information such as the student's gender, age, and highest education level. It is unique and is the data collected by the online education platform when the student registers. Student clickstream data is the type and frequency information of students interacting with the Virtual Learning Environment (VLE) platform in a course, which includes accessing resources, web-page click, forum click and so on. It reflects the active degree of students participating in the course. The OULA dataset includes 22 courses, 32,593 students, and 10,655,280 data on interactions between students and the VLE platform. Students' output is divided into four categories, including *Distinction* (*D*), *Pass* (*P*), *Fail* (*F*) and *Withdrawn* (*W*). When the student's score is higher than 75 points, his outcome is *D*. When the student's score is higher than 40 but lower than 75, his outcome is *P*. When a student completes the course, but the score is less than 40 points, his outcome is *F*. When the student does not complete the course, his outcome is *W*. We use data in this dataset to train and test our model, and compare the output of the model with the actual results. From this, we can get the accuracy, precision, recall and F1 score of the model in different situations.

When we use the OULA dataset, we divide the dataset differently according to different prediction tasks. When we perform the four-class classification prediction task, we retain the original four-class classification division in the dataset, namely *D*, *P*, *F* and *W*. In the binary classification prediction task of our general experiment, that is, the dropout prediction task, we consider *D* and *P* as *P*, keep *W*, and discard *F*. The students who pass the course are divided into one category, and the students who drop out are divided into another one category. When we perform the binary classification prediction task on a specific course, we consider *D* and *P* as *P*, take *W* and *F* as *F*. Students who will pass the course and those who have not passed the course are divided into two opposite categories.

5.2. Experimental Settings

In this article, we use the AML model to perform two online course student performance prediction tasks, that is, four classification tasks and two classification tasks. We also use the model to test the effect of two classification tasks on the student performance prediction task of specific courses and compare with the results obtained by using the models proposed in other papers to test. According to the above, the final performance of students in the OULA dataset we used is divided into four categories: *D*, *P*, *F*, and *W*. When we perform the four-category prediction task, we divide the prediction results into four categories as described above. When we perform the binary classification task, we consider *D* and *P* in the original four classifications as *P*, keep *W*, and discard *F*. In other words, we classify students who pass the course into one category recorded as *P*, classify those who drop out into another category recorded as *W*, and abandon those students who have completed the course but failed, which is the common dataset classification method for dropout prediction tasks. We use a five-fold cross-check method to train and test the proposed model, which can effectively eliminate the influence of the selection difference between the training set and the test set on the model. The specific steps of the five-fold cross-check are as follows:

- Firstly, we divide OULA dataset into five parts randomly, any two of them have no intersection.
- Secondly, we select one of them without repeating as test set and the others are train sets of the AML model.
- Thirdly, we test our trained model on the test set and obtain its accuracy, precision, recall and F1 score.
- Finally, we average the results of five index evaluations as the final result.

After repeated training and testing, we finally determined the relevant parameters of the proposed model. We set the number of fully connected layers for processing student demographic data and LSTM layers for processing student clickstream data to three layers, the learning rate of the model is 0.001, and the batch size is 100. The proposed model has the best general effect when set as above. We use the model constructed with the above parameters to perform two online course student performance prediction tasks.

The above is a general test of the model. Next, to test the prediction effect of the proposed model on a specific course, we randomly select some courses as the test set, and use the proposed model for training and testing. This article shows the effect of one case as an example. Our prediction tasks for specific courses are still divided into the four categories task and the two categories task. The classification of the prediction results in the four category is the same as that described above. For the two classification tasks, D and P are considered as P, F and W are considered as F, that is, the students who pass the course are classified into one category, as well as the students who fail the course are classified into another category. The model we proposed still uses the same model parameters as the general test on the prediction task in a specific course.

5.3. Results and Discussions

We use the proposed model to train and test the data of the first 25 weeks of each course in the OULA dataset and output the experimental results every five weeks. We use the following models as baseline models and compare their prediction effects with those of the proposed model to prove the effectiveness of our proposed model. In addition to predicting student performance after the start of the course, we also propose and complete the task of predicting student performance before the start of the course, that is, at week 0, which is unique in our paper. We not only test the generality of the proposed model, but also use it to test on a specific course and compare it with the baseline model. Experiments prove that our proposed model is always better than other models.

- Logistic Regression. We train a Logistic Regression (LR) model using scikit-learn package with the maximum number of iterations is 5000.
- ANN. We train a deep Artificial Neural Network (ANN) [30] model with the number of layers is 3.
- LSTM. We train a deep Long-Short Term Memory (LSTM) [36] model with the number of layers is 3.
- DOPP. DOPP [13] model is established by using student demographic data and student clickstream data to predict the performance of students.

5.3.1. Four-Class Classification

According to the experimental settings, we performed a five-fold cross-check on both the proposed model and the baseline model. Since the length of all courses is about 38 weeks, in order to better observe the prediction effect of the model in the early and mid-term of the course, we take the first 25 weeks of the course for training and testing, and output the results of the test set every five weeks, which are recorded as shown in Table 1.

By observing Table 1, we can draw the conclusions as follows:

- As the number of weeks increases, the predictive effect of each model has improved significantly, which is caused by the increase in the amount of input data. The more student clickstream data is input, the more accurately the

Table 1. Four-class classification.

Method	Weeks	Accuracy	Precision	Recall	F1 score
LR	5	44.79	37.99	34.83	29.69
	10	47.88	39.00	37.50	32.74
	15	49.82	40.31	39.75	35.27
	20	54.45	42.80	42.64	38.77
	25	57.21	45.46	44.94	41.74
ANN	5	50.66	40.84	37.20	33.11
	10	55.02	42.90	40.95	36.93
	15	57.81	43.76	43.39	39.45
	20	61.66	46.57	46.06	42.22
	25	63.55	50.49	48.02	43.97
LSTM	5	51.89	38.26	38.20	34.30
	10	56.62	40.44	41.99	38.30
	15	60.47	42.87	45.34	41.82
	20	64.09	45.29	48.19	44.28
	25	66.46	48.81	50.13	46.40
DOPP	5	52.66	47.91	38.95	35.49
	10	57.37	47.59	42.85	39.40
	15	61.15	45.74	45.74	42.51
	20	64.44	49.78	48.69	45.35
	25	66.88	57.43	50.68	47.54
AML	5	53.51	43.29	39.89	37.20
	10	57.79	45.71	43.73	41.70
	15	61.68	49.05	46.49	44.43
	20	65.00	54.98	49.30	46.66
	25	67.40	58.00	51.15	48.43

model can identify students' performance in a specific course, and the more accurate student performance pre-dictions can be made.

- Adding demographic data can help improving the student performance pre-diction effects of the model. The learning status of students is easily affected by the surrounding environment, which brings an inspiration that online education platforms can provide more personalized teaching programs based on the background information of different students. The influence of demographic data on the final prediction results is more obvious when the number of weeks is small, because the amount of student clickstream data entered at this time is low, and the model is more dependent on demographic data when making predictions. Especially when the number of weeks is 0, the prediction of the model is completely dependent on demographic data,

which is also the key to solving the cold start problem in the task of predicting student performance.

- Compared with other baseline models, the AML model has better predictive performance. This is because the AML model adds an attention mechanism to the DOPP model. The attention mechanism allows the model to focus on factors that have a greater impact on the model's student performance prediction effect, thereby improving the model's prediction accuracy, precision, recall and F1 score.

5.3.2. Binary Classification

Consistent with the four-class classification prediction task, we perform a five-fold cross-check for all models under the binary classification prediction task and use data from the first 25 weeks of the course for training and testing. We output and record the results of the test set every five weeks, as shown in the Table 2.

By observing Table 2, we can draw the conclusions as follows:

- The effects of all models on the binary classification prediction task are higher than their effects on the four-class classification prediction task, and the overall effect still shows a trend of increasing with the increase of the number of weeks, indicating that the effect of the student performance prediction task under the binary classification prediction task is still as the amount of student clickstream data increases.

- Since the fifteenth week, the accuracy and F1 score of the LSTM model, the DOPP model and the AML model on the binary classification prediction task don't have a significant gap. We think this is because the prediction effect of the LSTM model has reached a very high level in the fifteenth week, so the improvement of the DOPP model and the AML model based on it is relatively low, but there is still a small improvement. Therefore, we believe that the proposed model is still effective compared to the baseline model.

5.3.3. Evaluation on Week 0

From Table 1 & Table 2, we can see that both the proposed model and baseline models can predict student performance after the start of the course and the proposed model is always effective in prediction tasks compared with other baseline models. However, we are not satisfied with that the model can only predict student performance after the course starts. We hope to predict the student's final performance before the course starts, that is, week 0. We hope to identify students who are at risk of dropping out and failing as quickly as possible before the start of the course, which is not considered in other papers. Therefore, we use the proposed model to predict student performance at week 0 under binary classification task, and the results are shown in Table 3.

From Table 3, we can see that the model we proposed can make predictions for week 0, which is mainly based on the demographic data of students. Compared with other papers, our paper proposes and completes the task of predicting

Table 2. Binary classification.

Method	Weeks	Accuracy	Precision	Recall	F1 score
LR	5	70.10	64.48	68.23	64.30
	10	74.48	84.40	72.25	75.43
	15	78.33	68.18	85.61	76.90
	20	83.63	74.95	90.42	81.39
	25	88.47	81.06	93.69	86.65
ANN	5	76.48	79.64	56.27	65.25
	10	82.61	85.17	69.31	76.04
	15	86.57	85.50	79.50	82.36
	20	91.54	92.54	85.84	89.03
	25	94.61	94.67	91.80	93.20
LSTM	5	77.21	82.90	55.05	65.73
	10	83.24	95.72	70.42	76.69
	15	88.55	92.82	77.28	84.29
	20	93.09	96.94	85.64	90.89
	25	95.58	97.18	91.76	94.37
DOPP	5	77.97	83.63	56.39	67.06
	10	83.94	86.40	71.30	77.62
	15	88.63	92.39	77.83	84.42
	20	93.16	97.02	85.74	91.02
	25	95.64	97.41	91.66	94.44
AML	5	78.94	79.12	62.70	69.83
	10	84.82	88.99	69.86	78.52
	15	88.92	91.55	79.34	84.99
	20	93.31	96.13	86.89	91.26
	25	95.79	97.45	92.01	94.65

Table 3. Evaluation on week 0.

Category	Accuracy	Precision	Recall	F1 Score
Four-class	43.93	33.69	30.39	26.03
Binary	65.62	60.31	35.50	44.33

student performance at week 0, which helps the online education platform to make a preliminary judgment on the students participating in the course before the start of the course, focusing on those students who may drop out or fail and improve the pass rate of the course.

5.3.4. Evaluation on One Case
In the previous article, we complete the generality test of the proposed model.

Next, we will use the DOPP model as the baseline model to test the effect of both the proposed model and the baseline model on the specific course classification task. We display one case to show the effect of both the models. The specific course classification task uses the data of the BBB course opened in the two semesters of 2014B and 2014J as the test set, and the rest of the data as the training set. After such division, it is used as input data for training and testing. Using the experimental process given in [13] as a reference, we conducted experiments when only student clickstream data were used and when student demographic data was added. We record student clickstream data as *cl* and student demographic data as *de*. The results we get are shown in Table 4 & Table 5.

By observing Table 4 & Table 5, we can draw the conclusions as follows:

- In the four-class classification task and binary classification task, as the number of weeks increases, the prediction effects of the baseline model and the proposed model both improve, which is consistent with the results described above.

- In the same situation, the AML model has better predictive performance than the DOPP model, which shows that the AML model still has an advantage in predicting performance in a specific course.

Table 4. Four-class classification evaluation on the case.

Method	Data	Weeks	Accuracy	Precision	Recall	F1 score
DOPP	*cl*	5	56.16	39.81	40.76	38.01
		10	61.13	42.21	44.09	40.91
		15	63.53	44.08	46.27	42.12
		20	66.66	46.67	48.84	45.48
		25	68.01	49.47	50.05	46.59
AML	*cl*	5	56.88	41.89	41.23	38.89
		10	61.51	43.24	44.68	41.98
		15	64.48	42.95	46.24	42.67
		20	66.86	46.62	49.08	45.91
		25	68.71	48.66	51.28	48.41
DOPP	*cl+ de*	5	57.08	40.36	41.51	39.10
		10	61.79	43.29	44.75	42.04
		15	64.74	45.81	47.08	44.90
		20	67.02	62.43	50.72	46.97
		25	69.12	49.57	52.02	49.91
AML	*cl+ de*	5	57.75	40.72	41.98	39.95
		10	61.90	43.97	45.67	43.49
		15	64.84	45.60	48.04	45.72
		20	67.48	59.09	51.53	50.78
		25	69.55	62.95	52.84	51.50

Table 5. Binary classification evaluation on the case.

Method	Data	Weeks	Accuracy	Precision	Recall	F1 score
DOPP	cl	5	71.45	77.46	63.43	69.74
		10	77.41	82.99	71.03	76.54
		15	81.82	86.84	76.55	81.37
		20	85.12	91.03	79.12	84.66
		25	88.22	90.49	86.38	88.38
AML	cl	5	71.88	77.65	64.31	70.36
		10	78.00	84.59	70.43	76.87
		15	81.92	85.99	77.84	81.71
		20	85.33	90.98	79.62	84.92
		25	88.48	91.17	86.13	88.58
DOPP	cl+ de	5	73.14	79.20	65.40	71.64
		10	78.26	84.88	70.68	77.13
		15	82.33	87.40	77.05	81.90
		20	85.56	89.81	81.39	85.40
		25	88.91	93.93	84.06	88.72
AML	cl+ de	5	73.47	76.50	70.53	73.39
		10	78.51	85.60	70.43	77.28
		15	82.74	88.50	76.70	82.18
		20	85.66	90.27	81.10	85.44
		25	89.12	92.07	86.48	89.18

6. Conclusions

Different from the traditional face-to-face teaching method, the online education method relies on the powerful Internet technology to get rid of the time and place constraints of students in the learning process, and truly bring high-quality education to everyone. Online education has attracted a large number of students, and the number of students in each course far exceeds the number of students in traditional classrooms. Due to this situation, we need to propose a method, which is to build a student performance prediction system, to ensure the quality of online education for students. The online education platform collects student demographic data and student clickstream data to use student performance prediction models for tracking and analyzing student learning status in real time. Once the student's final performance prediction is found to be a failure or withdrawal, we can intervene in time to help students adjust their learning status and better master this course.

This article uses the Open University Learning Analytics (OULA) dataset for analysis and proposes an Attention-based Multi-layer LSTM (AML) model. We use student demographic data and student clickstream data to predict student

performance at the end of the period. The results show that the proposed model is always better than other models. In other words, the AML model can predict the student's final performance earlier and more accurately than other models. The reasons for the results are as follows. First, the AML model combines students' background information and interaction information with the online learning platform. Second, it adds an attention layer into multi-layer LSTM model, which helps the model pay more attention to those data that impact the prediction effect more deeply. Therefore, it can be used to intervene in the student's learning state earlier to reduce the dropout rate and failure rate of the course.

In the future, we will consider adding unused data in the OULA dataset to the model, such as course information, students' pre-course learning conditions, and the time when students submit classroom tests. We try to further improve the model's accuracy, precision, recall and F1 score when facing different students and different courses, especially in the initial stage of the course.

Acknowledgements

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Volery, T. and Lord, D. (2000) Critical Success Factors in Online Education. *International Journal of Educational Management*, **14**, 216-223.
https://doi.org/10.1108/09513540010344731

[2] Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D. and Emanuel, E. (2013) The MOOC Phenomenon: Who Takes Massive Open Online Courses and Why? *SSRN*, Article ID: 2350964.

[3] Bettinger, E. and Loeb, S. (2017) Promises and Pitfalls of Online Education. *Evidence Speaks Reports*, **2**, 1-4.

[4] Mackness, J., Mak, S. and Williams, R. (2010) The Ideals and Reality of Participating in a MOOC. *Proceedings of the 7th International Conference on Networked Learning*, Aalborg, 3-4 May 2010, 266-275.

[5] Salal, Y., Abdullaev, S. and Kumar, M. (2019) Educational Data Mining: Student Performance Prediction in Academic. *International Journal of Engineering and Advanced Technology*, **8**, 54-59.

[6] Arasaratnam-Smith, L.A. and Northcote, M. (2017) Community in Online Higher Education: Challenges and Opportunities. *Electronic Journal of e-Learning*, **15**, 188-198.

[7] Larreamendy-Joerns, J. and Leinhardt, G. (2006) Going the Distance with Online Education. *Review of Educational Research*, **76**, 567-605.
https://doi.org/10.3102%2F00346543076004567

[8]   Gargano, T. and Throop, J. (2017) Logging on: Using online Learning to Support the Academic Nomad. *Journal of International Students*, **7**, 918-924. https://doi.org/10.32674/jis.v7i3.308

[9]   Bao, W. (2020) Covid-19 and Online Teaching in Higher Education: A Case Study of Peking University. *Human Behavior and Emerging Technologies*, **2**, 113-115. https://doi.org/10.1002/hbe2.191

[10]   Korkmaz, G. and Toraman, C. (2020) Are We Ready for the Post-Covid-19 Educational Practice? An Investigation into What Educators Think as to Online Learning. *International Journal of Technology in Education and Science* (*IJTES*), **4**, 293-309. https://doi.org/10.46328/ijtes.v4i4.110

[11]   Liu, D., Zhang, Y., Zhang, J., Li, Q., Zhang, C. and Yin, Y. (2020) Multiple Features Fusion Attention Mechanism Enhanced Deep Knowledge Tracing for Student Performance Prediction. *IEEE Access*, **8**, 194894-1944903. https://doi.org/10.1109/ACCESS.2020.3033200

[12]   Pandey, M. and Taruna, S. (2016) Towards the Integration of Multiple Classifier Pertaining to the Student's Performance Prediction. *Perspectives in Science*, **8**, 364-366. https://doi.org/10.1016/j.pisc.2016.04.076

[13]   Karimi, H., Huang, J. and Derr, T. (2020) A Deep Model for Predicting Online Course Performance. *34th AAAI Conference on Artificial Intelligence*, New York, 7-12 January 2020.

[14]   Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017) Open University Learning Analytics Dataset. *Scientific Data*, **4**, Article No. 170171. https://doi.org/10.1038/sdata.2017.171

[15]   Halawa, S., Greene, D. and Mitchell, J. (2014) Dropout Prediction in Moocs Using Learner Activity Features. *Proceedings of the Second European MOOC Stake-Holder Summit*, Lausanne, 10-12 February 2014, 58-65.

[16]   Sun, A. and Chen, X. (2016) Online Education and Its Effective Practice: A Research Review. *Journal of Information Technology Education*: *Research*, **15**, 157-190.

[17]   Kebritchi, M., Lipschuetz, A. and Santiague, L. (2017) Issues and Challenges for Teaching Successful Online Courses in Higher Education: A Literature Review. *Journal of Educational Technology Systems*, **46**, 4-29. https://doi.org/10.1177%2F0047239516661713

[18]   Kim, B.-H., Vizitei, E. and Ganapathi, V. (2018) GritNet: Student Performance Prediction with Deep Learning. arXiv:1804.07405. https://doi.org/10.1111/bjet.12836

[19]   Wakelam, E., Jefferies, A., Davey, N. and Sun, Y. (2020) The Potential for Student Performance Prediction in Small Cohorts with Minimal Available Attributes. *British Journal of Educational Technology*, **51**, 347-370. https://doi.org/10.1111/bjet.12836

[20]   Mohammadi, M., Dawodi, M., Tomohisa, W. and Ahmadi, N. (2019) Comparative Study of Supervised Learning Algorithms for Student Performance Prediction. 2019 *International Conference on Artificial Intelligence in Information and Communication* (*ICAIIC*), Okinawa, 11-13 February 2019, 124-127. https://doi.org/10.1109/ICAIIC.2019.8669085

[21]   Aman, F., Rauf, A., Ali, R., Iqbal, F. and Khattak, A.M. (2019) A Predictive Model for Predicting Students Academic Performance. 2019 10*th International Conference on Information, Intelligence, Systems and Applications* (*IISA*), Patras, 15-17 July 2019, 1-4. https://doi.org/10.1109/IISA.2019.8900760

[22]   Botelho, A.F., Varatharaj, A., Patikorn, T., Doherty, D., Adjei, S.A. and Beck, J.E.

(2019) Developing Early Detectors of Student Attrition and Wheel Spinning Using Deep Learning. *IEEE Transactions on Learning Technologies*, **12**, 158-170. https://doi.org/10.1109/TLT.2019.2912162

[23] Nagatani, K., Zhang, Q., Sato, M., Chen, Y.-Y., Chen, F. and Ohkuma, T. (2019) Augmenting Knowledge Tracing by Considering Forgetting Behavior. *The World Wide Web Conference*, San Francisco, 13-17 May 2019, 3101-3107. https://doi.org/10.1145/3308558.3313565

[24] Hu, Q. and Rangwala, H. (2019) Reliable Deep Grade Prediction with Uncertainty Estimation. *9th International Conference on Learning Analytics & Knowledge*, 4-8 March 2019, 76-85. https://doi.org/10.1145/3303772.3303802

[25] Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., *et al.* (2018) Exercise-Enhanced Sequential Modeling for Student Performance Prediction. *32nd AAAI Conference on Artificial Intelligence*, New Orleans, 2-7 February 2018, 2435-2443.

[26] Sood, S. and Saini, M. (2021) Hybridization of Cluster-Based LDA and ANN for Student Performance Prediction and Comments Evaluation. *Education and Information Technologies*, 26, 2863-2878. https://doi.org/10.1007/s10639-020-10381-3

[27] Lu, H. and Yuan, J. (2018) Student Performance Prediction Model Based on Discriminative Feature Selection. *International Journal of Emerging Technologies in Learning*, **13**, 55-68. https://doi.org/10.3991/ijet.v13i10.9451

[28] Rizvi, S., Rienties, B. and Khoja, S.A. (2019) The Role of Demographics in Online Learning; A Decision Tree Based Approach. *Computers & Education*, **137**, 32-47. https://doi.org/10.1016/j.compedu.2019.04.001

[29] Kloft, M., Stiehler, F., Zheng, Z. and Pinkwart, N. (2014) Predicting MOOC Dropout over Weeks Using Machine Learning Methods. *Proceedings of the EMNLP* 2014 *Workshop on Analysis of Large Scale Social Interaction in MOOCs*, Doha, October 2014, 60-65. http://dx.doi.org/10.3115/v1/W14-4111

[30] Waheed, H., Hassan, S.-U., Aljohani, N.R., Hardman, J., Alelyani, S. and Nawaz, R. (2020) Predicting Academic Performance of Students from VLE Big Data Using Deep Learning Models. *Computers in Human Behavior*, **104**, Article ID: 106189. https://doi.org/10.1016/j.chb.2019.106189

[31] Huang, F., Zhang, X. and Li, Z. (2018) Learning Joint Multimodal Representation with Adversarial Attention Networks. *Proceedings of the* 26*th ACM International Conference on Multimedia*, New York, 22-26 October 2018, 1874-1882. https://doi.org/10.1145/3240508.3240614

[32] Huang, F., Xu, J. and Weng, J. (2020) Multi-Task Travel Route Planning with a Flexible Deep Learning Framework. *IEEE Transactions on Intelligent Transportation Systems*, **22**, 3907-3918. https://doi.org/10.1109/TITS.2020.2987645

[33] Huang, F., Wei, K., Weng, J. and Li, Z. (2020) Attention-Based Modality-Gated Networks for Image-Text Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications* (*TOMM*), **16**, Article No. 79. https://doi.org/10.1145/3388861

[34] Huang, F., Jolfaei, A. and Bashir, A.K. (2021) Robust Multimodal Representation Learning with Evolutionary Adversarial Attention Networks. *IEEE Transactions on Evolutionary Computation*, 17 March 2021, p. 1. https://doi.org/10.1109/TEVC.2021.3066285

[35] Huang, F., Li, C., Gao, B., Liu, Y., Alotaibi, S. and Chen, H. (2021) Deep Attentive Multimodal Network Representation Learning for Social Media Images. *ACM Transactions on Internet Technology* (*TOIT*), **21**, Article No. 69. https://doi.org/10.1145/3417295

[36] Aljohani, N.R., Fayoumi, A. and Hassan, S.-U. (2019) Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment. *Sustainability*, **11**, Article No. 7238. https://doi.org/10.3390/su11247238

[37] Hassan, S.-U., Waheed, H., Aljohani, N.R., Ali, M., Ventura, S. and Herrera, F. (2019) Virtual Learning Environment to Predict Withdrawal by Leveraging Deep Learning. *International Journal of Intelligent Systems*, **34**, 1935-1952. https://doi.org/10.1002/int.22129

[38] Jha, N.I., Ghergulescu, I. and Moldovan, A.-N. (2019) OULAD MOOC Dropout and Result Prediction Using Ensemble, Deep Learning and Regression Techniques. *Proceedings of the* 11*th International Conference on Computer Supported Education*: *CSEDU*, Vol. 2, Heraklion, 2-4 May 2017, 154-164. https://doi.org/10.5220/0007767901540164

[39] Hlosta, M., Herrmannova, D., Vachova, L., Kuzilek, J., Zdrahal, Z. and Wolff, A. (2018) Modelling Student Online Behaviour in a Virtual Learning Environment. arXiv: 1811.06369.