Pedagogical Intervention Practices: Improving Learning Engagement Based on Early Prediction

Han Wan, *Member, IEEE*, Kangxu Liu, Qiaoye Yu, and Xiaopeng Gao, *Member, IEEE*

***Abstract*—Most educational institutions adopted the hybrid teaching mode through learning management systems. The logging data/clickstream could describe learners' online behavior. Many researchers have used them to predict students' performance, which has led to a diverse set of findings, but how to use insights from captured data to enhance learning engagement is an open question. Furthermore, identifying students at risk of failure is only the first step in truly addressing this issue. It is important to create actionable predictive model in the real-world contexts to design interventions. In this paper, we first extracted features from students' learning activities and study habits to predict students' performance in the Kung Fu style competency education. Then, we proposed a TrAdaBoost-based transfer learning model, which was pretrained using the data of the former course iteration and applied to the current course iteration. Our results showed that the generalization ability of the prediction model across the teaching iterations is high, and the model can achieve relatively high precision even when the new data are not sufficient to train a model alone. This work helped in timely intervention toward the at-risk students. In addition, two intervention experiments with split-test were conducted separately in Fall 2017 and Summer 2018. The statistical tests showed that both behavior-based reminding intervention and error-related recommending intervention that based on early prediction played a positive role in improving the blended learning engagement.**

***Index Terms*—Data mining, learning performance prediction, pedagogical intervention, small private online course (SPOC), transfer learning.**

I. Introduction

MASSIVE open online courses (MOOCs) covered several aspects including online technologies, open educational resources, and abundant educational data that giving clues about how people learn. It became the accelerant that researchers paid more attention to the instructional strategies and assessment.

On the other hand, there were more on-campus courses that adopted the hybrid teaching mode. These small private online courses (SPOCs) combined online resources and technology with engagement between faculty and students based on online platforms (such as Open edX). It is important to observe the learning activities both inside and outside classrooms for the teaching quality monitoring. When developing capacities for quantitative educational researches, following problem need to be solved: What learning-related data should be collected? How could these educational data be gathered? How to extract meaningful information from them? What pedagogical intervention could be implemented? And finally, how does the intervention improve learning and teaching activities?

There is strong interest in identifying the at-risk students to reduce the attrition of the students in MOOCs. The timely early prediction can help instructors provide proper supports toward at-risk students in SPOC. For successful timely interventions in a SPOC, predictive models must be transferable—that is, they must perform well in the new course iteration.

Most predictive analytics on MOOCs have focused on training and evaluating models on the same course offering. Some models were trained on data retrospectively collected from completed courses. This made it difficult to conduct real-time prediction in the ongoing courses that are different from the previous course.

Therefore, transfer statistical knowledge between courses is of crucial importance if one wants to do real-time prediction. Unfortunately, it is usually difficult to apply the models that were built on the past course to a new one. Boyer [1] showed that models built on the previous offering did not always yield good predictive performance when applied to new offerings of the same MOOC.

Furthermore, the data-driven approach to address the personalization in an ongoing course could be structured in several steps as follows.

1) Build a transferable predictive model for possible behavioral outcomes.
2) Design intervention strategies that might deliver positive outcomes.
3) Execute the intervention based on the timely prediction.
4) Evaluate whether the intervention improved outcomes.

In this paper, we make use of the advantages of open edX, acquiring the learning interaction data of the on-campus students in our SPOC. Based on the machine learning methodology, we could find out at-risk students in real time. A transfer learning model was applied in the Summer 2018 semester to

Manuscript received September 15, 2018; revised April 5, 2019; accepted April 9, 2019. Date of publication April 15, 2019; date of current version June 17, 2019. This work was supported in part by the Computer Education Research Association of Chinese Universities (CERACU2019R12) and in part by the Research and Practice Program of New Engineering Disciplines (Beihang–Tencent Cooperation). (*Corresponding author: Xiaopeng Gao.*)

The authors are with the School of Computer Science and Engineering, Beihang University, Beijing 100083, China. (e-mail: wanhan@buaa.edu.cn; liukangxu@buaa.edu.cn; yuqiaoy@buaa.edu.cn; gxp@buaa.edu.cn).

Digital Object Identifier 10.1109/TLT.2019.2911284

obtain the benefit from the Fall 2017 prediction model. Then we posited and designed interventions that were likely to deliver positive outcomes. After using the model to produce predictions in real time, we executed pedagogical intervention toward at-risk students. We evaluated the differences in learning behavior under reminding intervention in Fall 2017 (study 1). The impacts of the error-related recommendation on learning engagement were elaborated in study 2 (intervention in Summer 2018). In this paper, we focus on the learning engagement that could be measured through the online interactions in the following dimensions: duration of online study, completion of assignments, and participation in the discussion forum.

The remainder of this paper starts with a review of the recent studies in the field of performance prediction and pedagogical intervention in Section II. Then, Section III lays the contexts including the structure, schedule, and the grading mechanism of our course. Section IV elaborates the learners' performance prediction in Fall 2017 semester. Following this, Section V presents the transfer learning model built in Summer 2018. The split-tests conducting the pedagogical interventions are discussed in Section VI. Section VII analyzes the result of the transfer learning model and the interventions. Section VIII draws the conclusions and future work.

II. Background

The online learning management system (LMS) could provide instructors or researchers data about students' learning and interaction behavior. Initial research aimed toward MOOC personalization mostly focuses on using past observations to build a predictive model [2]. Moreno-Marcos *et al.* [3] describe most common characteristics of the MOOCs that have been used for prediction, analyzed the outcomes have been predicted in contributions, discussed prediction features, elaborated techniques/models used for prediction in MOOCs, and highlighted the future research directions.

Many researchers analyzed students' learning behavior based on the data and tried to predict students' performance [4], [5] using data mining and machine learning method. These studies focused on training and evaluating models on the same course offering.

He *et al.* [6] described the prediction of at-risk students in MOOC using transfer learning models. They implemented and evaluated two logistic regression (LR) algorithms on different offerings of a MOOC. The results indicated that using data from the previous offering combined with early data of the current offering could accurately identify at-risk students in ongoing courses. However, they focused on the implementation of the predictive models which was only the first step to provide effective and appropriate prevention strategies.

Harvard research group [7] used multinomial logistic regression to identify students that at risk of dropping out in the MOOC and send them an email to ask about their lack of engagement. Their reports showed that survey motivated some students to re-engage in the class and increased the comeback rate in certain cases.

On the one hand, other researchers aimed at enhancing the pedagogical effectiveness of MOOC or SPOC. Chudzicki [8] focus on the impact of problem formats and pretest feedback on learners. The entire learner population was partitioned into two or more groups and each group was given separate course material in a single split-test. They found that students who involved in drag-and-drop activities performed better than the learners that involved in the multiple-choice counterparts. And their experiments also showed little evidence for enhancement of posttest scores due to students seeing the same items on the pretest, even though the pretest gave feedback.

Furthermore, there are also some research works about social media in education, which refers to the practice of using social media platforms as a way of enhancing the education of students.

Bicen [9] found that the participants of MOOC courses tend to obtain information from the social media instead of from the pages related to MOOCs when they encountered problems. Tampere University of Technology [10] provided social networking tools to support collaborative study. They described how interventions could motivate students to use a social networking and students could take advantage from such an environment with social network.

Compared with the forum in MOOCs, external social tools also promoted the discussion and sharing resources related to the MOOC. Typical built-in social tools of MOOCs include wiki, discussion forum, and microblogging. Among them, Facebook and Twitter are the most popular third-party social media. Ternauciuc and Mihaescu [11] introduced the built-in social media tools in MOOCs and compared the difference between Moodle LMS and MOOC. They proposed that Moodle LMS with integrated social tools could work as an efficient MOOC platform. Purser *et al.* [12] analyzed the learners' activities on the MOOC related Facebook group, and found that social media worked well as a catalyst for learner agency. For online courses, collaborative learning is an effective replacement of instructors, and how to engage and motivate the students might influence their performance. Chen *et al.* [13] retrieved the social network of students on the discussion board in SPOC, and then grouped the students based on the academic results. Furthermore, they developed the group division modules based on the online discussion of the edX platform, and found that this method did a better job in learning performance despite the worse satisfactions. Konstantinou and Epps [14] integrated a third-party social media application in UNSW's LMS, which could create the online community to facilitate the interactions. The students' feedback showed their work increased engagement and assisting with learning.

These studies had shown that social tools were the basis for supporting the connections among MOOC participants, and they were sometimes helped increase engagement and learning.

III. Contexts

Computer structure is a second-year course in the School of Computer Science and Engineering, Beihang University. This course is designed to help students to comprehend MIPS architecture and assembly through a series of laboratory-based projects [15]. In Fall 2017, three tutorials including lecture videos,

e-texts, quizzes, and worked examples were created for on-campus students to grasp the fundamental knowledge. Furthermore, a full-featured autograding submission system was integrated with the Open edX platform. This testing system could judge learners' submission automatically and then return feedback with the grade to learners.

A. Structure of the Course

The structure of Open edX courseware looks like the hierarchical structure of a traditional textbook [16]:

1) At the top level is the chapter, and each chapter represents for a tutorial or a project.
2) Each chapter contains several sequentials, and each sequential represents a section of the tutorial or the project, which could contain several verticals.
3) Each vertical could contain an unlimited number of XBlocks [17], such as HTML block (for e-text), video block (for lecture video), problem block (for various kinds of quizzes and project work (PW) submission), and discussion block (for accessing the posts with the same tag).

When the content is released, it becomes visible to the learners. Usually, we release all sequentials within a chapter simultaneously, and we also could set release and due dates sequential by sequential.

B. Course Schedule

The Fall 2017 iteration contains six weeks of tutorials, seven weeks of graded material, and five weeks of optional content. At the beginning of the semester, the materials for tutorials were released simultaneously, and students were prompted to learn those tutorials at their own pace. At the end of week 6, an in-class test containing three problems chosen from tutorials was conducted to examine the students' learning outcomes. After that, learners were required to challenge nine projects (from Project 0 to Project 8).

From week 7, the material for each week corresponds to one chapter within the online course—one sequential for instruction material, one sequential for that week's PW at home, and one sequential for a weekly in-class project test (PT). The material for a given week was released one week before it was due, with the automatic testing of the homework for a given week due simultaneously.

The in-class PT usually was conducted at the end of each week. Students need to implement several homework-like works. Or they need to expand their homework to meet the requirements. During the test, we used the visibility control to hide the PW and discussion from the learners. Content groups and cohorts were also used to make the PT only available to the learners in the corresponding progress. After passing the autograding submission test, students still need to answer several questions related to the project with the instructor or teaching assistant (TA). A grade (from "F" to "A+") that reflects the mastery of project content would be given to each student.

Like Kung Fu belt test, the learners who passed the in-class PT could pace on to the next project [15]. Students those who failed in the in-class session could get the face-to-face help from TA or teachers and should rechallenge the current project in the next week.

The summer iteration is prepared for the learners who failed in the previous iterations, which means that all students enrolled in the summer semester are retakers. In the Summer 2018 course iteration, there were no self-paced tutorials and only 6 weeks long which contain 12 in-class tests (PTs). Learners could pace on their study from the project that they failed in the previous course iterations and they could proceed to redo the in-class test.

C. Course Grading

Our previous study [15] showed that the Kung Fu style competency education prompts students to own their learning as the pace and/or the path of learning. In this case, we mainly graded students according to the quantity of the in-class projects that they completed and the quality of each completed project.

Furthermore, the completion degree of the tutorials and the participation of the discussion forum were also taken into consideration in the Fall 2017 semester. First, a grade level would be attached to each student based on the project that he/she eventually passed. For instance, if a learner passed the in-class PT of the Project 5, he/she would get a score between 60 and 69. Second, the baseline score depended on learners' performance of the in-class session. Finally, the performance of the tutorials and the discussion forum were added as an extra bonus.

IV. Behavior-Based Performance Prediction

We had built a model to predict students' final performance based on their learning behaviors [18], which achieved an area under the receiver operating characteristic (AUROC) curve in the range 0.927–0.984. However, to conduct timely pedagogical interventions, it is crucial to predict whether the learner could pass the in-class test each week.

A. Feature Engineering

When students interact with the online LMS, their learning behavior will be recorded via the tracking log. The sources of those log events can be grouped into the following types:

1) course resources interaction: "page_closed," "load_video," "play_video," "pause_video," and so on;
2) problem interaction: "problem_get," "problem_save," "problem_check," "problem_graded," and so on;
3) discussion forum interaction: "thread.created," "thread.opened," "response.created," "searched," and so on;
4) survey and other modules.

We had conducted 26 features which were either indicators of students' learning behavior or indicators of students' learning habits.

The features listed in Table I focus on students' online learning behavior, for instance:

1) features x2 and x14 measure the online learning time of the learner;
2) features x3, x4, and x16 show the strength of the learner's engagement in the forum.

Furthermore, students' learning habit related features were calculated as given in Table II.

TABLE I
LEARNING BEHAVIOR RELATED FEATURES

	Name	Definition
x2	total_duration	Total time spent on all resources
x3	number_forum_posts	Number of forum posts posted
x4	number_forum_browse	Number of forum posts browsed
x5	number_distinct_problems_submitted	Number of distinct problems attempted
x6	number_submissions	Number of submissions
x7	number_distinct_problems_submitted_correct	Number of distinct problems attempted and corrected
x8	average_number_submissions	Average number of submissions per problem (x6/x5)
x9	oberserved_event_duration_per_correct_problem	Total time spent / number of distinct correct problems (x2/x7)
x10	submissions_per_correct_problem	Number of submissions / number of correct problems (x5/x7)
x11	average_time_to_solve_problem	Average time from the first submission to the last submission for each problem
x12	observed_event_variance	Variance of a student's observed event timestamps
x13	max_observed_event_duration	Maximum duration of observed events
x14	total_video_duration	Total time spent on video resources
x15	total_courseware_access	Number of courseware access
x16	number_forum_responses	Number of forum responses
x17	average_number_of_submissions_percentile	A student's submissions count / the average of all the students' submissions count
x18	average_number_of_submissions_percent	A student's submissions count / the maximum of all the students' submissions count
x19	number_submissions_final_correct_problems	Number of submissions for final correct problems
x20	correct_submissions_percent	Percentage of the total submissions that were correct (x8/x7)

TABLE II
LEARNING HABITS RELATED FEATURES

	Name	Definition
x201	average_start_submission_time	Average first submitting time after the problem released
x202	total_courseware_access_after_incorrect_submission	Number of courseware access after the problem submitted incorrect
x203	total_video_time_after_incorrect_submission	Video watching time after the problem submit incorrect
x204	average_time_between_problem_submission	Average duration between problem submissions
x205	time_first_visit	Minimum (first problem_get_time, first e-text_access_time) – project_release_time
x206	average_time_till_first_check	Average time between problem_first_check and problem_first_get of all problems
x207	discussion_duration_after_incorrect_submission	Total discussion duration after incorrect submission

1) Feature x201 shows that students tend to finish their assignments as soon as possible or delay to the deadline.
2) Features x202 and x203 indicate whether the learner review resources if submitted a wrong answer.
3) Feature x204 is the average submitting interval. It covers the possible situation that the student just resubmits another answer quickly when the former submission is incorrect.
4) Features x205 and x206 show students who attempt to do the homework soon after the project has been released or not.
5) Feature x207 indicates how much time the learner spent in the discussion forum if the submission is incorrect.

As shown in Fig. 1, we assembled the features of one week to a vector. In order to consider the effect of students' history learning behavior, each feature includes two parts: history feature and current feature. The way of calculating history feature is similar to exponential moving average

$$\mathrm{h}_k[1] = 0$$

$$\mathrm{h}_k[i+1] = \mathrm{x}_k[i]{*}\beta + \mathrm{h}_k[i]{*}(1-\beta)$$

where k is the feature index, from x1 to x207. $\mathrm{x}_k[i]$ is one feature of week i ($i = 1, 2, \ldots, 12$) and $\mathrm{h}_k[i]$ is its relevant history feature. β is a parameter and equals to 0.5 in our model.

B. Model Selection

When building the prediction model in Fall 2017 semester, we compared several classification algorithms, including LR, support vector machine (SVM), and gradient boosting decision tree (GBDT). GBDT classifier had been finally chosen to build the predictive model for the following reasons.

1) Compared to Regression/Bayes/SVM models, GBDT allows the combination of different features to have different discriminant.
2) GBDT has inherited the characteristics of the boosting algorithm with low degree of overfitting.

C. Model Evaluation

Here we use the AUROC score of prediction model in each week as the evaluation metric. For each week in the course, we executed the following steps to perform the GBDT analysis.

1) Assemble the features as described in Fig. 1.
2) Divide the data into ten folds, nine for training and the rest for testing.
3) Train a GBDT model, and evaluate the model using AUROC score.
4) Evaluate the model using the mean of AUROC score.
5) Calculate the mean of AUROC scores from each week.

Finally, the average AUROC score of the GBDT prediction model has reached 0.86.

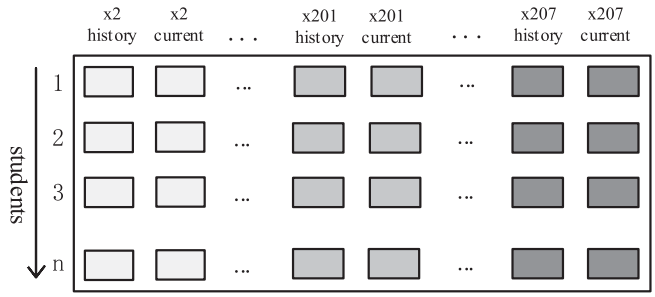

Fig. 1. Feature flattening process.

TABLE III
REMOVED FEATURES

Name	Definition
number_forum_posts	Number of forum posts posted
number_forum_browse	Number of forum posts browsed
number_forum_responses	Number of forum responses
average_start_submission_time	Average first submitting time after the problem released
time_first_visit	Minimum (first problem_get_time, first e-text_access_time) – project_release_time
discussion_duration_after_incorrect_submission	Total discussion duration after incorrect submission

V. TRANSFER LEARNING PREDICTION MODEL

In Summer 2018 semester, the course schedule changed as mentioned in Section III-B. There were two in-class checking sessions each week (Monday and Wednesday). When predicting the students' performance at each in-class test, the prediction performance is poor when trained the classifier built in Section IV using the data of Fall 2017 semester directly. Therefore, we rebuilt the prediction model based on the transfer learning algorithm in order to improve the portability of the prediction models across teaching iterations.

A. Feature Screening

There were some learning behavior differences in course between Fall 2017 and Summer 2018. Since the summer course is open to students who retake this course, all the course content was released at the beginning of the course. Besides that, students almost never participated in the course forum. As a result, those features (given in Table III) that related to forum and releasing time would be absent in the prediction model.

B. Transfer Learning Procedure

The training dataset (Fall 2017) and predicting dataset (Summer 2018) have different distributions; furthermore, the dataset in Summer 2018 is too small to train a prediction model. Therefore, we considered using transfer learning to solve this problem.

TrAdaboost [19] is a transfer learning algorithm which can deal the problem of different data distributions. The key points of TrAdaboost are giving each instance a weight and training the classifier several rounds. It also needs a base classifier. In each training round, train the base classifier using dataset with weights and calculate the error rate, then update the weights of instances for next training round. The training dataset of TrAdaboost contains two parts: source data which are diff-distribution training instances, and target data which are the same-distribution ones.

We extracted students' learning behavior data of Fall 2017 semester as the source data, and the data of Summer 2018 semester was used as the target data. When an instance is classified wrongly in one training round, we conclude the following.

1) If this instance is from the source data, we thought this instance is dissimilar to the target data and it should hold a lower weight. Thus, in the next round, the misclassified diff-distribution training instances will affect the learning process less than the current round.
2) If this sample is from the target data, we thought this instance is difficult to classify and it should hold a higher training weight. The instances with higher training weights will intent to help the learning algorithm to train better classifiers.

We built the model with an open-source machine learning library scikit-learn in Python 3.5.

VI. PEDAGOGICAL INTERVENTIONS

In the Fall 2017 and Summer 2018 semesters, two teaching interventions had been conducted through the platform built with Enterprise WeChat. The first one focused on using performance predictions to provide reminder interventions for at-risk learners. The second one attempted to deliver retakers who were at-risk with error-related recommending intervention. In both experiments, randomized controlled trials were conducted to minimize selection bias. The goal of both interventions is to improve learners' engagement.

A. Intervention Platform

Enterprise WeChat is an instant messaging (IM) application developed by Tencent. It aims to help colleagues to communicate and allow companies to develop modules that meet their own requirements. In terms of supporting pedagogical interventions, we used Enterprise WeChat to develop the intervention platform. The reasons for using Enterprise WeChat instead of traditional email are as follows.

1) Timely: As an IM tool, Enterprise WeChat can deliver messages to users in real time.
2) Easy to Track: The highly customizable API provided by Enterprise WeChat makes it easy to track user interactions with the intervention message accurately based on its authentication mechanism. Meanwhile, tracking emails are much more complicated, and some email service providers may block tracking links.
3) Free of Charge: Using Enterprise WeChat as an intervention platform is completely free.

B. Controlled Trial Design

Typically, those who participated in the trial were randomly allocated to either the group with additional intervention or to a group only receiving conventional teaching as the control. Considering that some individual learner characteristics (such as

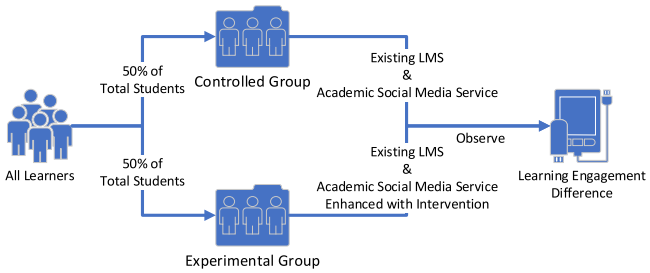

Fig. 2. Schematic diagram of the randomized controlled trials.

grade point average (GPA) of the previous academic year, the grade of the prerequisite courses, etc.) could be obtained at the beginning of course, we grouped the learners using blocked randomization sampling method. It ensures that the initial ability distribution of each group is approximately equal.

As shown in Fig. 2, throughout the teaching iteration, the experimental groups (group B in Fall 2017 and group D in Summer 2018) are injected with additional intervention. Meanwhile, the controlled groups (group A in Fall 2017 and group C in Summer 2018) only participate in basic teaching procedures without intervention. Furthermore, the controlled trial is stable, that is, the same partitioning of learners was used throughout the course. Comparison groups allow the teachers to determine the effects of the intervention when compared with the on intervention (experimental) group, while other variables are identical.

C. Participants and Partition

Considering our large learner population and the wide distribution of students' initial ability, the best way to conduct the partition is using stratified sampling. In order to make the experiment more reliable and reduce the impact of individual differences, learners have been divided into groups based on their learning ability. Freshman year GPA had been used in the 2017 Fall semester trial, while the performance of last iteration had been used in the 2018 Summer semester trial.

As shown in Fig. 3, 438 sophomores who enrolled in the 2017 Fall semester were partitioned into several groups according to their GPA. Each group covered a grade segment of five points. Starting with the group of the highest average score, if the size of the group is not even, the member with the lowest score in the group would be moved into the neighboring group. After that, group A and group B picked up the student randomly from each grade segment in turn. Eventually, two groups both had 219 members.

There were 124 learners who failed in previous semesters enrolled in the 2018 Summer teaching iteration. Their historical performance of the course is more suitable to reflect the students' learning abilities. Therefore, according to historical performance, they were divided into two groups using the same procedure used in Fall 2017. Eventually, two groups both had 62 retakers.

D. Behavior-Based Intervention in Fall 2017

The experiment was conducted among 438 sophomores who majored in computer science. To measure the effects of

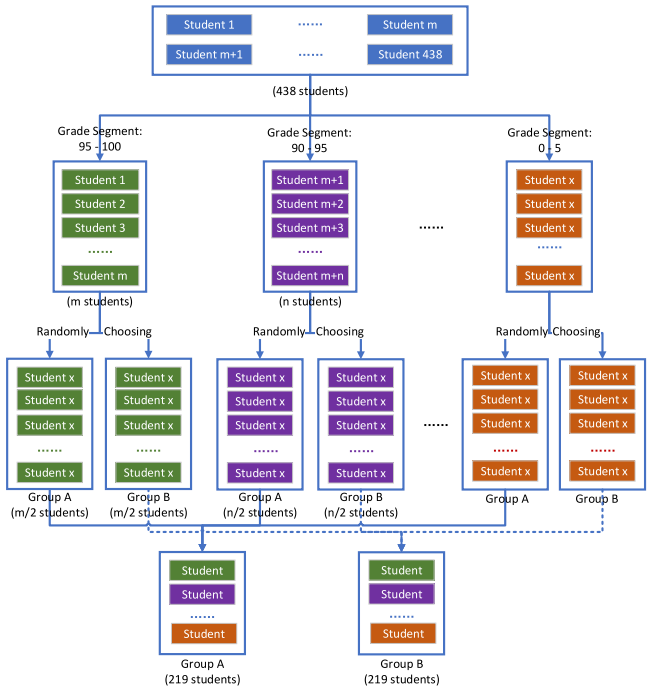

Fig. 3. Partitioning process of the participants in the Fall 2017 semester.

the intervention, we conducted an intervention experiment with split-test (also known as A/B testing).

1) Feature Importance Analytics: Based on the predictive model built in Section IV, we calculated the importance of features using data from week 1 to week 6 of the last Fall semester course iteration [18]. As shown in Table IV, there are some features that show higher importance in the prediction.

1) *observed_event_variance* shows the stability of the time distribution of learner's online activities during the week.
2) *total_duration* measures the time a learner spent on all resources.
3) *total_lecture_duration* indicates the time a learner spent on watching lecture videos.
4) Some features are related to other features, for example, *observed_event_duration_per_correct_problem* is associated with *max_observed_event_duration* and *number_distinct_problems_submitted_correct*.
5) Some features are relevant to student's initial knowledge level, like the *correct_submissions_percent* (reflects the correct rate of submissions).

From our early study, we found that learners' activities in the tutorial had affected their final performance directly. We also learned the lesson that we need to do intervention as early as possible during the self-paced tutorial learning stage (from week 1 to week 6) instead of just warning students using the in-class test at the end of week 6. Furthermore, most failed student did not contribute in the online forum, and their number of forum posts read also far below the average. In this case, we focused on analyzing at-risk learners' behavior and how to stimulate them in learning with intervention.

2) Warning Inactive Students: By analyzing the feature importance mentioned above, we selected several features that

TABLE IV
FEATURE IMPORTANCE OF THE PREDICTION MODEL USING DATA FROM WEEK 1 TO WEEK 6

Name	Importance
observed_event_variance	0.088165
total_duration	0.078527
max_observed_event_duration	0.061522
total_lecture_duration	0.061385
total_book_duration	0.055852
observed_event_duration_per_correct_problem	0.053234
pset_grade_overtime	0.045696
time_on_problem_molecular	0.045341
time_on_problem_atomic	0.044301
average_time_till_first_check	0.042233
total_video_duration_before_submit	0.041671
time_first_visit	0.037739
average_predeadline_submission_time	0.037329
problem_finish_num_pre_deadline48h	0.036012
correct_submissions_percent	0.035489
total_video_duration_after_incorrect_submission	0.034355
number_distinct_problem_submit	0.027037
number_submissions_correct	0.026670
average_time_to_solve_problem	0.020995
total_ebook_duration_before_submit	0.014928
total_ebook_duration_after_incorrect_submission	0.01342
number_distinct_problems_submitted_correct	0.012892
submissions_per_correct_problem	0.011881
average_time_betwen_problem_submission	0.011534
total_discussion_duration_after_incorrect_submission	0.010617
pset_grade	0.009149
average_number_of_submissions_percentile	0.008629
number_submissions	0.007725
number_forum_posts	0.007517
average_number_of_submissions_percent	0.006542
problem_finish_num_pre_start96h	0.004899
average_number_submissions	0.004503
number_forum_responses	0.002209

are interpretable and observable with higher importance, including the total time spent on online learning, the time spent on lecture videos, and the number of distinct problems attempted. Besides, we also added a feature that measures the number of reading times in the discussion forum based on practical experience.

We conducted performance prediction multiple times during the tutorials to identify the at-risk students. After observing their behavior distribution in each dimension, if 90% of the at-risk learners' activity metric were below a certain value, then it would be set as the threshold of current dimension. Form the inactive learner set on each dimension using corresponding threshold, and the intersection of above five dimensions composed the warning list. The students whose engagement was below the thresholds were called latecomers. Therefore, for the experimental group (group B), if a student was in the warning list, we would push a message which contained his/her current study status, weak points, and possible engagement improving suggestions to his/her Enterprise WeChat client. For the controlled group (group A), nothing would be done. We also kept on monitoring their learning engagement in the rest of the course.

E. Error-Related Intervention in Summer 2018

The experiment was conducted among 124 learners who failed in previous semesters and enrolled in the 2018 Summer teaching iteration. The randomized controlled trial was used to enhance the creditability of the result. The partition procedure is described in Section VI-C.

All problems in PW and PT could be autoevaluated by the grading system that was developed based on the Open edX's external grader service mechanism. The autograding submission system reduced the evaluation burden of instructors, and it also improved students' learning experience since they could get timely feedback and resubmit improved assignments until the deadline. On the other side, the autograding system enabled staff to provide customized recommendations that related to learners' errors.

The most recent extension in the autograding system is to support setting up multiple test cases in a single problem. The system will evaluate each submission using all the test cases that belong to the problem and provide the learner with the evaluated status of each case. For instructors and TAs, the system could generate reports including pass rates and detailed error messages for each testcase, which can be used to analyze common mistakes.

In Summer 2018, we used the transfer learning model proposed in Section V to identify at-risk learners of each in-class test. At the same time, based on the automatic evaluation system enhanced with refined test cases, our TAs analyzed and organized the common mistakes of each in-class test. In order to improve the accuracy of the intervention, common mistakes were analyzed at the level of testcases, and the intervention-related messages were customized according to the individual's answer status. Therefore, for the learners in the experimental group (group D), if he/she was predicted to be at risk, then we would provide him/her with a common mistake solution written by the TA team, which contained his/her failed test cases. For the students in the controlled group (group C), nothing would be done.

Those error-related interventions were pushed to the experimental group on Day 5, Day 12, and Day 19, respectively, in Summer 2018.

VII. RESULTS AND DISCUSSION

A. Evaluation of the Transfer Learning Model

The AUROC score is used for model evaluation. A GBDT-based prediction model (described in Section IV) trained with the data of Fall 2017 semester was used as the baseline model. We executed the following steps 36 h before the in-class checking session to perform predicting.

1) First, assemble the features of different weeks into a composite vector according to the process described in Section IV-A.
2) Second, use the data of the Fall 2017 semester as the source data, and the data of the first two weeks in Summer 2018 semester as the target data to train a TrAdaboost classifier. Here, we chose Naïve Bayes classifier as the base classifier, and the number of training rounds is 10.
3) Then, predict students' performance 36 h before each in-class checking session using both the baseline model and the transfer learning model.

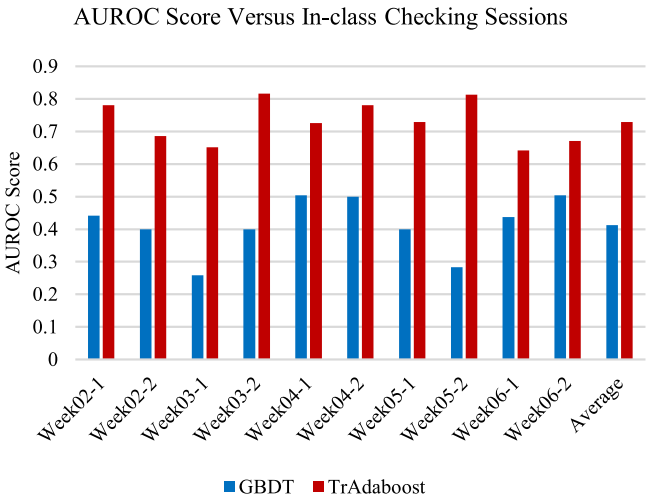

Fig. 4. AUROC score results for predicting student' performance.

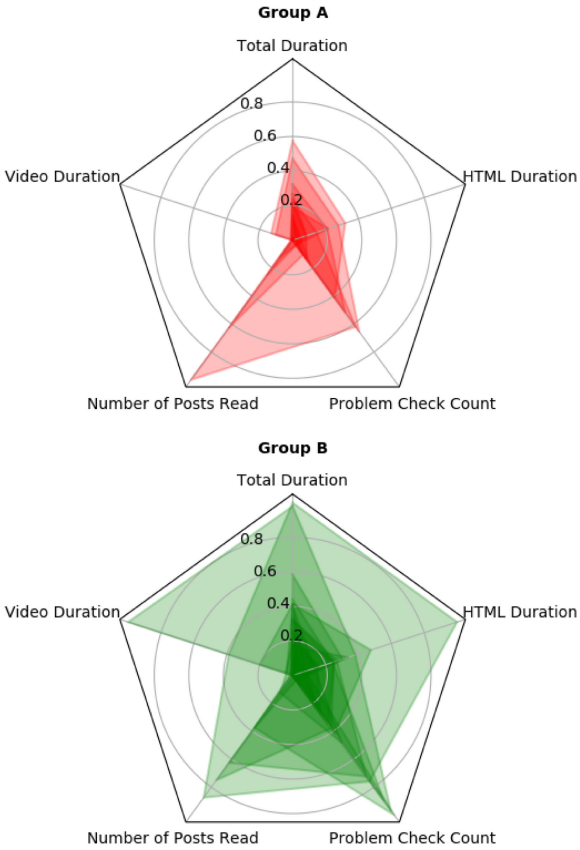

Fig. 5. Learning engagement distribution of the latecomers in each group from the first reminder to the end of the tutorial learning phase in the following aspects: the total time spent online, the time spent on lecture videos, the time spent on reading e-texts, the number of attempts to solve problems, and the number of the forum posts read.

4) At last, evaluate the performance of the baseline model and the transfer learning model using the AUROC score for each in-class checking session.

As shown in Fig. 4, the *x*-axis indicates the in-class checking session, and the *y*-axis shows the AUROC score of the corresponding prediction.

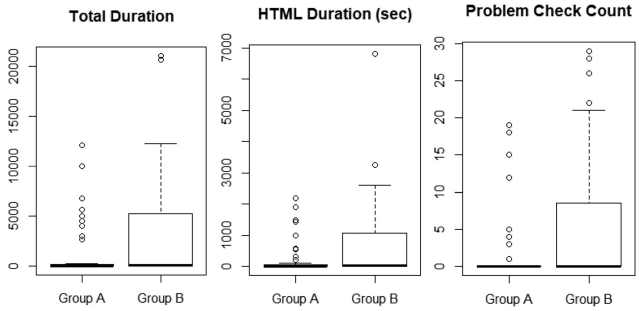

Fig. 6. Boxplots of online interaction behavior statistics of the latecomers from the first reminding push to the end of the tutorial.

We predicted whether the learner failed in the in-class checking session from week 1 to week 6 of the 2018 Summer semester. We noticed that the AUROC score of GBDT (the model built in Section IV) is less than 0.5, and that means the performance of the baseline model is worse than random classifier when predicting student performance in a new environment (course).

The AUROC score of the transfer learning model is higher than 0.7 in most time, and it has better performance than the baseline model. The average AUROC score has reached 0.73.

B. Impact of the Behavior-Based Intervention

In order to investigate whether the behavior-based intervention had a positive impact on learners' engagement, the online learning behavior of students before and after the intervention is measured. As the goal of our study is to improve the learning engagement of latecomers, we focused on observing features that are directly related to learning engagement, including the total time spent online, the time spent on lecture videos, the time spent on reading e-texts, the number of problem attempts, and the number of the forum posts read.

During the Fall 2017 tutorials learning phase, there were 42 learners in group A and 41 learners in group B were marked as latecomers using the thresholds concluded from the prediction result. The engagement of those students will be analyzed in the following paragraphs.

Fig. 5 shows the radar charts of the total amount of individual behavior for each latecomer in the two groups from the first reminding intervention to the end of the self-paced tutorials learning stage. Each layer in the radar chart indicates the engagement of a student in five dimensions. The closer the layer is to the edge, the more active the student is. Therefore, the figure indicates that most latecomers in the experimental group (group B) showed a higher engagement than the controlled group (group A) in the following dimensions: *Total Duration*, *HTML Duration,* and *Problem Check Count*. The boxplots in Fig. 6 show the difference in the latecomers' learning engagement between the experimental group and the controlled group after the first intervention, which is consistent with the trend shown in the radar charts.

Fig. 7 gives a line chart with the trends of learning engagement of latecomers in the two groups during the tutorial learning phase. It can be observed that after the interventions, the overall online activities of Group B were higher than Group A in all five aspects. To investigate whether there are statistically

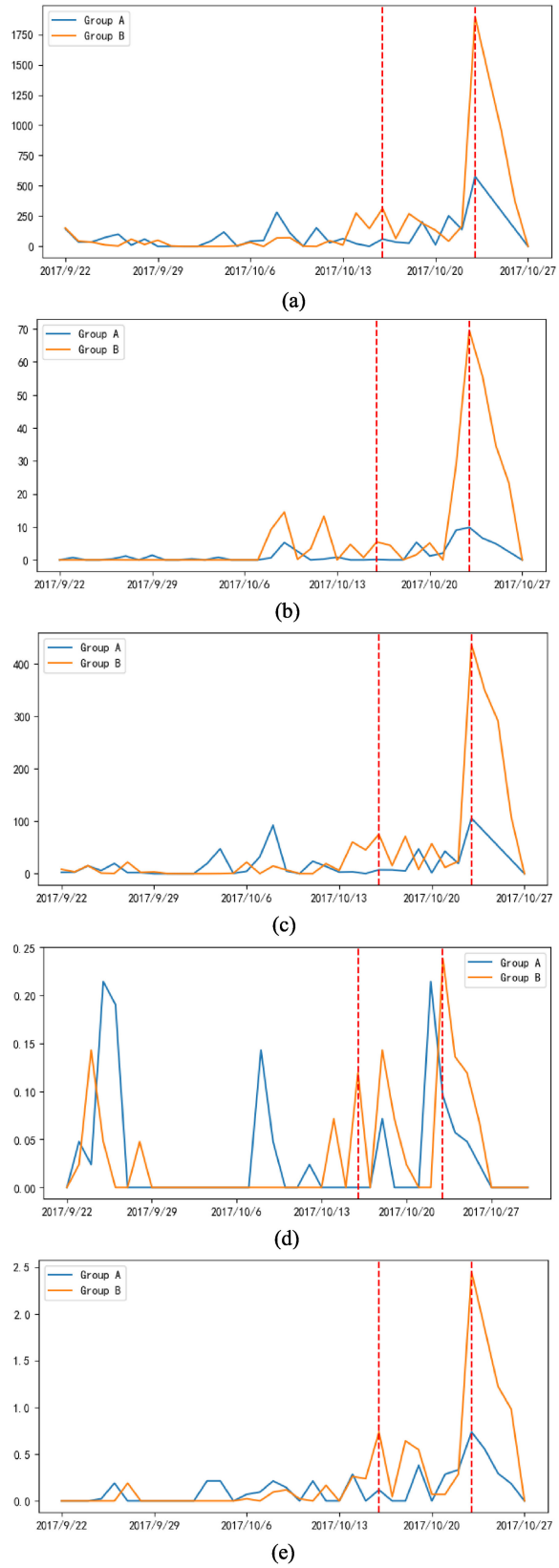

Fig. 7. Engagement statistics (per capita) of learners in group A and group B who were under the threshold during the tutorials in Fall 2017. The red vertical lines indicate the days when the interventions were given. Learners with zero activity are included in the statistics. (a) Total duration (s). (b) Video duration (s). (c) E-text duration (s). (d) Discussion posts read count. (e) Problem check count.

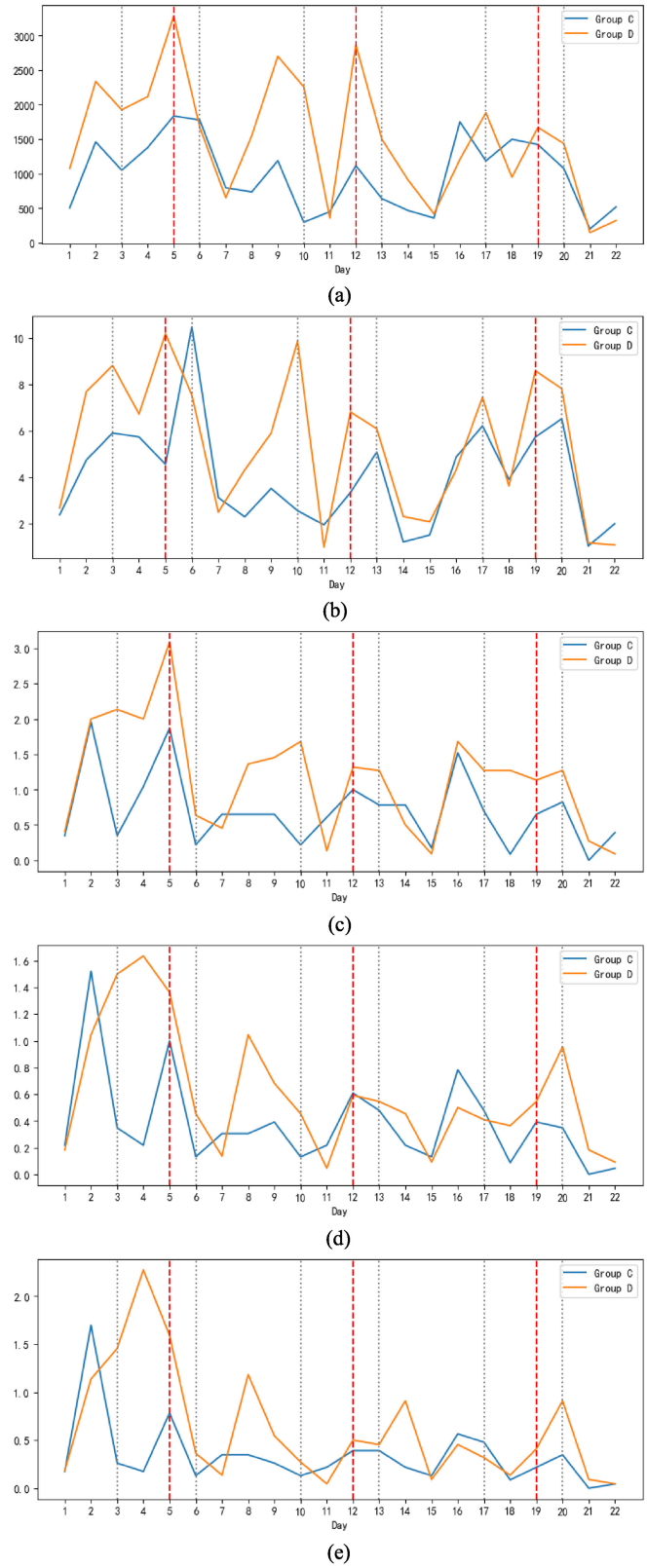

Fig. 8. Engagement statistics (per capita) of the retakers who were at-risk in group C and group D in Summer 2018. (a), (b) are the features reflecting learning behavior, (c)–(e) are the features that directly related to problem solving. The gray vertical lines indicate the in-class PT days. The red vertical lines indicate the days when the interventions were given. Learners with zero activity are included in the statistics. (a) Total duration (s). (b) E-text visiting count. (c) Total submissions. (d) Correct submissions. (e) Problem attempts.

TABLE V
DESCRIPTIVE STATISTICS ON LATECOMERS' ENGAGEMENT BEFORE/AFTER THE FIRST INTERVENTION DURING THE TUTORIALS

Name		Group A		Group B		W	p-value
		Mean	SD	Mean	SD		
Total Duration	Before Intervention	56.110	95.979	43.305	83.369	803	0.875
	After Intervention	166.64	358.14	403.441	665.319	640	0.066
Video Duration	Before Intervention	0.483	1.736	1.467	4.27	855	0.911
	After Intervention	2.741	9.27	10.984	36.867	776	0.238
E-text Duration	Before Intervention	11.985	26.857	9.688	21.456	813	0.948
	After Intervention	30.141	68.504	91.601	165.278	613	0.034*
Discussion Posts Read Count	Before Intervention	0.029	0.091	0.014	0.064	941	0.157
	After Intervention	0.048	0.183	0.076	0.211	817	0.484
Problem Check Count	Before Intervention	0.068	0.173	0.047	0.119	842	0.764
	After Intervention	0.238	0.612	0.638	1.087	643	0.044*

*: Significant at the 0.05 level (alternative hypothesis: true location shift is not equal to 0).

significant differences in the distribution of learning engagement, several Shapiro–Wilk tests were performed on the latecomers' learning behavior statistics before and after the first intervention in group A and group B. The results indicate that the data were not from normally distributed populations ($p < 0.05$). Therefore, two-sample Wilcoxon rank-sum (Mann–Whitney U) nonparametric tests were conducted to determine that whether the distribution of two groups was significant different ($p < 0.05$). Table V shows the means and standard deviations of the latecomers' engagement in five dimensions of both groups, along with the test statistic W and the p-value.

First, for the *Total Duration* dimension, as listed in Table V, the nonparametric test results showed that there was no significant difference in behavior between the controlled group and the experimental group before and after the intervention ($W = 803$, $p = 0.875$; $W = 640$, $p = 0.066$, respectively). However, the increment of mean in group B was more than three times that of group A, and the p-value was close to 0.05, which indicates that the interventions may have an impact on the time invested in learning, although not significant at the 0.05 level.

Next, for the *Video Duration* and *Discussion Posts Read Count* dimensions, although the increments of the mean of group B were higher than group A, the test results suggested that the effect of interventions on learning engagement was not significant in these dimensions.

Then, for the *E-text Duration* dimension, the Wilcoxon rank-sum test results indicated a statistically significant difference between group A and group B after the first intervention ($W = 613$, $p = 0.034$) while the difference before the intervention was not significant ($W = 813$, $p = 0.948$). Thus, learners showed more learning engagement in reading course materials after the interventions.

Finally, the statistics also indicated that the interventions significantly improve the students' attempts in solving problems ($W = 643$, $p = 0.044$).

C. Impact of the Error-Related Intervention

Since the 124 learners in the 2018 summer semester were all retakers, there were no tutorial learning weeks. Furthermore, we found that nearly no student watched the lecture videos after analyzing log data. Therefore, we selected the following five dimensions to measure students' learning engagement in this experiment: *Total Duration*, *E-text Visiting Count*, *Total Submissions*, *Correct Submissions*, and *Problem Attempts*. *Total Duration* and *E-text Visiting Count* reflect the learners' efforts to engage in learning. *Total Submissions*, *Correct Submissions,* and *Problem Attempts* are directly related to problem-solving attributes.

During the 2018 summer semester, there were 22 retakers in group C and 23 retakers in group D were identified at-risk by the transfer learning model. The engagement of those students will be analyzed in the following paragraphs.

Fig. 8 gives several trends of the experimental group (group D) and the controlled group (group C) observed during the course iteration. The figure shows that on most days, the average engagement of the experimental group was higher than those of the controlled group. Fig. 8(d) and (e) demonstrates the influence of the interventions on problem-solving efficiency—learners in group D made more correct submissions and attempted more problems than those in group C. Meanwhile, while improving the efficiency of problem solving, the interventions also promoted students to review e-texts and spend more time on learning, as shown in Fig. 8(a) and (b).

To investigate whether the intervention produced significant differences in learning engagement, a set of Shapiro–Wilk tests were first performed to evaluate the normality of the samples. Applying this test to the data of *Total Duration* and *E-text Visiting Count* dimensions, the test results indicate that the samples may come from normally distributed populations ($p > 0.05$). When applying this test to the remaining three dimensions, the results show that the data may not follow normal distribution. Therefore, the parametric two-sample Welch t-test and the nonparametric two-sample Wilcoxon rank-sum (Mann–Whitney U) test were performed according to the normality of the samples, respectively.

As shown in Table VI, there was significant difference between the experimental group ($M = 987.824$, $SD = 514.963$) and the controlled group ($M = 1511.382$, $SD = 867.927$) in the *Total Duration* dimension ($t = -2.433$, $p < 0.05$). Furthermore, the nonparametric test results also indicate that the error-related interventions might have an impact on engagement in the *Total Submissions*, *Correct Submissions,* and *Problem Attempts* dimensions, although the differences are not significant at 0.05 level ($p = 0.056$; $p = 0.065$; $p = 0.069$, respectively). Yet, the differences in *E-text Visiting Count* are nonsignificant between the experimental group and the controlled group ($p = 0.095$).

TABLE VI
DESCRIPTIVE STATISTICS ON AT-RISK LEARNERS' ENGAGEMENT AFTER THE FIRST INTERVENTION DURING THE SUMMER 2018 SEMESTER

Name	Group C		Group D		Statistical	Statistical	
	Mean	SD	Mean	SD	Test Method	Test Result	
Total Duration	987.824	514.963	1511.382	867.927	Parametric t-test	t = -2.433	p = 0.020*
E-text Visiting Count	4.036	2.241	5.393	2.980	Parametric t-test	t = -1.707	p = 0.095
Total Submissions	0.704	0.527	1.161	0.779	Non-parametric Wilcoxon	W = 160	p = 0.056
Correct Submissions	0.379	0.350	0.603	0.466	Non-parametric Wilcoxon	W = 163	p = 0.065
Problem Attempts	0.391	0.380	0.680	0.610	Non-parametric Wilcoxon	W = 164	p = 0.069

*: Significant at the 0.05 level (alternative hypothesis: true difference in means is not equal to 0).

VIII. CONCLUSION AND FUTURE WORK

In this paper, we described the background about an SPOC delivered via the Open edX platform, including course structure, schedule, grading, and our pervious study of predicting the learners' performance based on their study behavior features. Our transfer learning model utilized a small amount of newly labeled data (Summer 2018 iteration) to leverage the old data (Fall 2017 iteration) to construct a high-quality classification model for the new iteration course. Results showed that the generalization ability of the prediction model across the teaching iterations is high, and the model can achieve relatively high precision even when the new data are not sufficient to train a model alone.

Second, we designed and conducted two pedagogical interventions with controlled trial in different course iterations. The first one focused on using performance predictions to provide reminder interventions for learners in the warning list. The second one attempted to deliver learners who were at-risk with error-related recommending interventions.

Furthermore, we compared the transition of latecomers in learning behavior between the experimental group and the controlled group. It turned out that reminding inactive learners did make a difference in their learning engagement. Besides, the results showed that the error-related interventions played a positive role in improving learning engagement.

However, our work is only the first step to improve learning engagement through Enterprise WeChat. The findings of this paper also serve to open new lines of use social tools in on-campus courses. One of these research lines refers to encouraging everyone, including mentors, to further contribute in the discussion forum. Early studies showed that participants tend to obtain information from the social media instead of searching the encountered problem in the discussion forum. Searching and waiting the reply from other learners or mentors are time consuming. It would be valuable if it is possible to remind others about new-occurring question and reply through social tools. Another line of research of concerns how to find out at-risk learner as early as possible with higher accuracy. Further, there is a need for mechanisms that help developing personalized recommending systems assist in the education intervention.

As an initial attempt to conduct pedagogical intervention, we propose many possible directions for the future work. In the next course iteration, we will extend the functions of the LMS and autograding system to collect more detailed data which will be used in improving the transfer learning model. Furthermore, we will improve the intervention platform so that learners could customize the reminding messages to the concerned discussion forum activities.

ACKNOWLEDGMENT

The authors would like to thank Principles of Computer Organization teaching staff members: L. Xudong, X. Limin, L. Zhongzhi, N. Jianwei, Z. Liang, F. Cuijiao, L. Huiyong, and Y. Jianlei. They would also like to thank the aid of the TA group that helped the course improvements, including Z. Mingyuan, L. Ziyuan, G. Zhiyuan, W. Chenyu, Y. Yang, C. Yue, Z. Wenbin, C. Xingzhou, L. Jiaxin, and D. Jun, for their attempts in building performance prediction model in Fall 2016.

REFERENCES

[1] S. Boyer and K. Veeramachaneni, "Transfer learning for predictive models in massive open online courses," in *Proc. Artif. Intell. Edu.*, 2015, pp. 54–63.
[2] J. Gardner and C. Brooks, "Student success prediction in MOOCs," *User Model. User-Adapted Interaction*, vol. 28, no. 2, pp. 127–203, Jun. 2018.
[3] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, "Prediction in MOOCs: A review and future research directions," *IEEE Trans. Learn. Technol.*, to be published, doi: 10.1109/TLT.2018.2856808
[4] C. Taylor, "Stopout prediction in massive open online courses," Ph.D. dissertation, Massachusetts Inst. Technol., Department of Electrical Engineering and Computer Science, Cambridge, MA, USA, 2014.
[5] Z. Ren, H. Rangwala, and A. Johri, "Predicting performance on MOOC assessments using multi-regression models," in *Proc. Int. Conf. Edu. Data Mining*, 2016, pp. 484–489.
[6] J. He, J. Bailey, B. I. Rubinstein, and R. Zhang, "Identifying at-risk students in massive open online courses," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1749–1755.
[7] J. Whitehill, J. Williams, G. Lopez, C. Coleman, and J. Reich, "Beyond prediction: First steps toward automatic intervention in MOOC student stopout," in *Proc. Int. Conf. Edu. Data Mining*, 2015, pp. 171–178.
[8] C. A. Chudzicki, "Learning experiments in a MOOC (massive open online course)," M.S. thesis, Massachusetts Inst. Technol., Department of Physics, Cambridge, MA, USA, 2015.
[9] H. Bicen, "Determining the effect of using social media as a MOOC tool," *Procedia Comput. Sci.*, vol. 120, pp. 172–176, Jan. 2017.
[10] K. Silius, T. Miilumäki, J. Huhtamäki, T. Tebest, J. Meriläinen, and S. Pohjolainen, "Social media enhanced studying and learning in higher education," in *Proc. IEEE EDUCON Conf.*, 2010, pp. 137–143.
[11] A. Ternauciuc and V. Mihaescu, "Use of social media in MOOC—Integration with the Moodle LCMS," *eLearning Softw. Edu.*, vol. 1, no. 1, pp. 298–303, 2014.
[12] E. Purser, A. Towndrow, and A. Aranguiz, "Realising the potential of peer-to-peer learning: taming a MOOC with social media," *e-Learning Papers*, vol. 33, pp. 1–5, May 2013.
[13] Y. Chen, Y. Lin, L. Chu, Y. Chiou, and T. K. Shih, "Team formation for collaborative learning with social network consideration based on edX's online discussion board," in *Proc. 8th Int. Conf. Ubi-Media Comput.*, 2015, pp. 146–151.
[14] G. Konstantinou and J. Epps, "Facilitating online casual interactions and creating a community of learning in a first-year electrical engineering course," in *Proc. IEEE 6th Int. Conf. Teaching, Assessment, Learn. Eng.*, 2017, pp. 128–133.

[15] H. Wan, X. Gao, and Q. Liu, "Hybrid teaching mode for laboratory-based remote education of computer structure course," in *Proc. IEEE Front. Edu. Conf.*, 2016, pp. 1–8.
[16] "Understanding course building blocks," Jan. 28, 2018. [Online]. Available: http://edx.readthedocs.io/projects/open-edx-building-and-running-a-course/en/latest/developing_course/workflow.html#id7
[17] "Working with HTML components," Jan. 28, 2018. [Online]. Available:: http://edx.readthedocs.io/projects/open-edx-building-and-running-a-course/en/latest/course_components/create_html_component.html#working-with-html-components
[18] H. Wan, D. Jun, X. Gao, and K. Liu, "Supporting quality teaching using educational data mining based on OpenEdX platform," in *Proc. IEEE Front. Edu. Conf.*, 2017, pp. 1–7.
[19] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. 24th Int. Conf. Mach. Learn.*, New York, NY, USA, 2007, pp. 193–200.

Han Wan (M'2013) received the Ph.D. degree in computer architecture from Beihang University, Beijing, China, in 2011.

She has been a Lecturer with Beihang University since 2011, and was a Visiting Scholar with the Education Research Group, Massachusetts Institute of Technology (MIT) from 2015 to 2016. Her work at MIT was on the topic of drop-out prediction in MOOCs using learners' study habits features. She then returned to Beihang University where she worked as a Lecturer with the School of Computer Science and Engineering. Her research interests lie at educational intervention and adaptation mechanisms to support individualized learning in blended teaching mode. She is a member of the IEEE Education Society.

Kangxu Liu received the B.E. degree in computer science and technology in 2018 from Beihang University, Beijing, China, where he is currently working toward the Master's degree with the School of Computer Science and Engineering.

His research interests including autograding system, plagiarism detection, educational data mining, and recommendation system.

Qiaoye Yu is currently working toward the Postgraduate degree with the School of Computer Science and Engineering, Beihang University, Beijing, China. He is majoring in data mining and machine learning on educational area.

His research interests include learning performance prediction in MOOCs, recommendation system of online course forum, and deep neural network for knowledge tracing.

Xiaopeng Gao (M'19) received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2002.

He is currently the Associate Dean with the School of Computer Science and Engineering, Beihang University, Beijing. His educational research is supported by the Beijing Municipal Education Commission and Tencent, Ltd. He has authored/coauthored more than 60 papers in national and international conference proceedings and journals. His current research interest is supporting virtual lab online and accurate educational assessment.

Dr. Gao received the Second Prize of National Teaching Achievement (the highest prize in teaching and education granted by Chinese government) in 2014 and the top Prize of Beijing's Teaching Achievement in 2017. He is a member of the ACM.