# Designing a Transferable Predictive Model for Online Learning Using a Bayesian Updating Approach

Wanli Xing, Dongping Du, *Member, IEEE*, Ali Bakhshi, Kuo-Chun Chiu, and Hanxiang Du

***Abstract*—Predictive modeling in online education is a popular topic in learning analytics research and practice. This study proposes a novel predictive modeling method to improve model transferability over time within the same course and across different courses. The research gaps addressed are limited evidence showing whether a predictive model built on historical data retrospectively can be directly applied to a future offering of the same course or to another different course; lacking interpretable data mining models to improve model transferability over time and across courses. Three datasets from two distinct online courses with one course offered two times over two years were applied using direct transferring of the predictive model and the proposed Bayesian updating technique for model transfer. The results showed that the direct transferring of predictive model to the subsequent offering of the course and to a totally different course did not work effectively. By contrast, the proposed Bayesian updating provided a robust and interpretable approach with improved model transferability results for both situations. This Bayesian updating model can be continuously updated with new collected data rather than building prediction model from scratch every time, which can serve as a new methodological framework to carry experience and knowledge from past and other courses forward to new courses.**

***Index Terms*—Learning analytics (LA), machine learning, model transferability, online learning, performance prediction.**

## I. Introduction

A RECENT report of higher education showed that distance student enrollments have increased for the past 16 years straight [1]. A third of all students in postsecondary education now take at least one online course [2]. Furthermore, 65% of higher education institutions have indicated that online learning is essential to their long-term strategy [3]. Similar trends are also observed in K–12 education. According to the 2016 NMC Horizon Report, more than 2.7 million American K–12 students chose to use online learning [4]. With the rapid growth of online education, a critical concern is that teachers find it difficult to monitor students' progress as compared to face-to-face classrooms. A lack of timely teacher support and intervention is one of the major reasons for low retention rate and low student achievement in online learning [5].

As learning management systems can track every click action when students study online, LA researchers and practitioners have utilized such behavioral data traces along with other static factors (e.g., education background, socioeconomic status) to build predictive models of student performance. This predictive modeling enables teachers to monitor student progress and to identify at-risk students in advance in order to support early intervention [6], [7]. The objective of performance prediction is to estimate an unknown value, often the final performance of the student. To meet this goal, various data are transformed to features and then input into machine learning algorithms to predict student performance. According to a review of LA research, the prediction of student performance is one of the most popular topics in the context of online education [8], [9].

Most predictive analytics research for online learning has centered on training and evaluating models on one course retrospectively. That is, they build the predictive model of student performance using data from a completed course by following a standard data mining procedure: splitting the historical data into a training set and a testing set, applying machine learning to the model from the training data with cross validation, and further testing the model's accuracy on a test dataset as a proxy for how the model may perform on unseen data [10]. This retrospective modeling approach makes it difficult to examine whether this model is effective in subsequent offerings of the same course or in a different course because courses evolve over time, the learner cohort also differs, and the differences among courses can be tremendous [11]. From a practice perspective, the quality and quantity of data used to calibrate predictive models is essential to their success or failure [12]. It is important to rely on a sufficiently large dataset to generate adequate model accuracy, for example, building a prediction model using multiple years' worth of course data. Models developed with limited data will fundamentally undermine their performance. Therefore, it is vital to carry the intelligence and knowledge learned from historical data forward. Moreover, matching data from different sources and combining them is also a substantial effort [13].

Manuscript received June 11, 2020; revised April 28, 2021 and August 17, 2021; accepted August 21, 2021. Date of publication August 24, 2021; date of current version October 15, 2021. This work was supported in part by the University of Florida AI Catalyst under Grant P0195022 and in part by the University of Florida Informatics Institute Seed Funding. *(Corresponding author: Wanli Xing.)*

Wanli Xing and Hanxiang Du are with the School of Teaching and Learning, University of Florida, Gainesville, FL 32611 USA (e-mail: wanli.xing@coe.ufl.edu; h.du@ufl.edu).

Dongping Du, Ali Bakhshi, and Kuo-Chun Chiu are with the Department of Industrial, Manufacturing, and Systems Engineering, Texas Tech University, Lubbock, TX 79409 USA (e-mail: dongping.du@ttu.edu; ali.bakhshi@ttu.edu; kuchiu@ttu.edu).

Digital Object Identifier 10.1109/TLT.2021.3107349

When a model is developed and well-specified for capturing the relationship between various data features and student learning status, it is cost-effective to transfer such as model to future offerings of the same course and to other courses.

Research on building transferrable predictive models is very limited. Some studies have attempted to improve the transferability of a predictive model by building the model using a large amount of historical data from a variety of courses [14]. Others only used data from a live course to eliminate the need for transferability [15], which means the data for model training is limited. While there has been some initial effort to build transferrable predictive models from an algorithmic perspective [16], [17], most have focused on black box machine learning models such as neural networks and ensemble models. It is difficult for researchers and practitioners to interpret modeling results and further transform the results into actionable intelligence for teachers. There is a lack of research on the updating of predictive models to build historical data for future use.

In this study, we proposed a Bayesian updating approach relying on a white box model [logistic regression (LR)] to improve the transferability of a predictive model to a subsequent offering of the same course and to a different course. This study has two goals. First, it aims to generate concrete evidence to show whether we can directly apply a predictive model built using data from a historical course directly to the following offering of the same course or a different course. Second, we intend to employ a Bayesian updating approach to improve predictive model transferability over time and across courses. This study further conducted sensitivity analysis in order to test the robustness of the approach by examining the size of the new data used for updating the model. The following two research questions guided this study.

1) Can a predictive model built on historical data from a course directly apply to the following offering of the same course or to a different course?
2) How and to what extent can the Bayesian updating approach improve the predictive model transferability?

## II. Background

Predictive modeling, one of the most popular topics in LA, has been widely used in educational research and practice [18]. The typical application of predictive analytics is to develop early warning systems. That is, to identify students at risk of dropping out or failing a course as accurately and as soon as possible so that teachers can provide timely support and intervention to students. For example, a Harvard research group applied multinomial LR to detect students that are at risk of dropping out of a MOOC course and sent them an e-mail to ask about their lack of participation [19]. Foster and Siddle [20] used predictive models to send nonengagement alerts to first year at-risk undergraduate students. The effectiveness of the early warning systems heavily depends on the accuracy of the predictive models.

Most current studies have focused on training and testing a predictive model in the same course in a posthoc way. These studies split the data from a completed course into training and testing sets and then report the prediction performance from the testing results as an approximation for prediction results in the future offering of the same class. For instance, Robinson *et al.* [21] applied natural language processing to predict student achievement in MOOCs using completed MOOC course data by splitting the data into training and testing datasets. Similarly, Zhang *et al.* [22] applied optimized LR to process the various behavioral data traces to predict student performance using a historical dataset collected from a completed online course. Lemay and Doleck [23] built a predictive model based on student historical data in a completed online course to forecast students' grades. Whitehill and his colleagues [24] can be referred to for additional, similar studies.

Building and testing a predictive model using only the historical data from a course may not transfer well to future offerings of the same course. From a conceptual perspective, the course may evolve over time and student populations differ which may impair the transferability of the predictive model built simply on historical data [11], [25]. Thus, it is important to update the model over time. It is even more challenging to transfer a predictive model constructed for one course to a different course. While there are a few conceptual discussions of difficulty for transferring model between courses, e.g., different course design, content, and students [11], [25], limited empirical studies were found to directly examine the predictive model transferability to a different course. Wolff *et al.* [26] applied the built decision tree model for a course directly to a different course and found that these models had reasonable quality without providing further information.

A few studies have begun to explore the transferability of predictive models in online learning and other contexts. These studies roughly fall into four categories. First, few researchers have directly applied a predictive model built for one context to another. For example, Boyer and Veeramachaneni [10] evaluated the application of a predictive model built on a previous edition of an online course to the same following course and found a drop of at least 0.1 in AUC. Ocumpaugh *et al.* [27] also had a similar problem when they developed predictive analytics to detect the affective states of different populations and found that the detectors trained on one population could not be used on other populations. Gitinabard, *et al.* [28] applied the predictive model built for a course in the previous semester to the following semester, resulting in a drop in F1 score. However, He *et al.* [29] found that predictive models trained on a first edition of a MOOC performed well on a second edition of the same course.

Second, some researchers attempted to build a transferable model with a large amount of training data from different courses. For instance, Kidzinsk *et al.* [14] analyzed how to transfer models to other instances of the same course and to different courses. In order to generalize the predictive model across many courses, they trained their predictive models with data from 20 different completed online courses. They admitted that there was a tradeoff between specificity and generalizability because in order to achieve high performance in a course, it was better to use variables that depended on the specific course context. Olivé *et al.* [16] attempted to build a one-size-fits-all predictive model using data from 78 722 students in 5487 courses. Their research showed that adding specific contextual information from courses resulted in better prediction accuracy and F1 scores.

Third, others have proposed "*in situ*" models. That is, to use only the available data in an ongoing course (e.g., using data from the first two weeks for week 3 prediction). So, compared with transferring model built using the historical data from a finished course, these "*in situ*" models have the same course context. Bote-Lorenzo and Gómez-Sánchez [15] built an *in situ* model to predict the decrease of academic engagement in MOOCs over several weeks. A limitation of these *in situ* models is that they cannot be used when the dependent variable is only available at the end of the course, such as the pass and failure in a course analyzed in this current article. More importantly, the limited nature of the data quantity for this method may weaken model performance.

Fourth, additional research explored transferability from an algorithm perspective. These algorithms involved black box data mining models such as deep neural networks and ensemble models. For example, Ding *et al*. [11] applied a representational learning LSTM neural network to improve the transfer learning of predictive models to other open online courses. Wan *et al.* [17] applied TrAdaboost, an ensemble transfer learning algorithm, which can learn the different data distributions between source data and target data to improve the transferability of predictive models. Botelho *et al*. [30] employed deep neural networks to build student attrition predictive models to improve model transferability across assignments.

In sum, most of the current research on predictive modeling was built and tested on historical data from a single course term. There is a lack of direct and systematic evidence to show whether such models can be used in future offerings of the same course and further applied to different courses. Some initial explorations have been performed to build transferrable predictive models using large-scale data from historical courses (modeling results can be improved with course-specific information), and limited data from a live course for *in situ* models (cannot use when the dependent variable is available at the end of the course, limitation on data quantity can reduce the prediction performance), as well as black box data mining models (difficult to interpret the modeling results). As a result, more research is needed to concretely investigate the direct application of predictive models to the same course in the future and to other courses, as well as ways to update the prediction model building on historical data to improve model transferability.

TABLE I
FEATURES AND EXPLANATION

| Variable | Attribute | Description |
|---|---|---|
| gender | Categorical | gender of the student |
| highEd | Categorical | the highest education background of the student |
| imdBand | Categorical | the indices of the economic background |
| age | Numeric | age of the student |
| credits | Numeric | number of credits attained |
| assessment | Numeric | assessment scores |
| dataplusCnt | Numeric | sum clicks on the additional information such as videos, audios, sites, etc. |
| dualpaneCnt | Numeric | sum clicks on the information on site and activity related to that information |
| externalquizCnt | Numeric | sum clicks on the external quiz activity |
| folderCnt | Numeric | sum clicks on the files relevant to course |
| forumngCnt | Numeric | sum clicks on the discussion forum |
| glossaryCnt | Numeric | sum clicks on the basic glossary related to contents of course |
| homepageCnt | Numeric | sum clicks on the course homepage |
| htmlactivityCnt | Numeric | sum clicks on the interactive html activity |
| collaborateCnt | Numeric | sum clicks on the online video discussions |
| contentCnt | Numeric | sum clicks on the contents of the assignment |
| elluminateCnt | Numeric | sum clicks on the online tutorial sessions |
| wikiCnt | Numeric | sum clicks on the Wikipedia content |
| pageCnt | Numeric | sum clicks on the information related to course |
| questionnaireCnt | Numeric | sum clicks on the questionnaires related to the course |
| quizCnt | Numeric | sum clicks on the course quiz |
| repeatactivityCnt | Numeric | sum clicks on the course contents from previous weeks |
| resourceCnt | Numeric | sum clicks on pdf resources such as books |
| sharedsubpageCnt | Numeric | sum clicks on the shared information between courses and faculty |
| subpageCnt | Numeric | sum clicks on the other sites enabled in the course |
| urlCnt | Numeric | sum clicks on the links to audio/video contents |

## III. METHODOLOGY

### A. Research Context and Data

In this section, we present the research context and data. Data were collected from two online courses in a large western online university in the year 2013 and 2014, one social science course and one STEM course. Data from an additional offering of the social science course in the following year was also gathered. These three courses can be studied as a stand-alone course or as part of a program or major and no previous qualifications are required. Each course was presented multiple times during the year. We named these courses Social Science Prior (SSP), STEM, and Social Science After (SSA) and they enrolled 1767, 1614, and 1613 students, respectively. SSP and SSA are the same course offered during different years, and STEM is a different course. Course evolution, differences in student populations, and the different setup between courses (STEM vs. the social science course) are expected to influence the transferability of the predictive models.

In order to build the predictive model, we used data gathered until two-thirds of each course were complete in order to predict final pass and failure results of students. The data features used in this study are illustrated in Table I. We will rely on these data sources to examine 1) whether the models built on historical data in a course (SSP) can be directly used in the following identical course SSA (direct transfer SSP –> SSA) and whether a model built for the STEM course can be directly used in the SSA course (direct STEM –> SSA); 2) whether the Bayesian updating approach can improve the transferability of a predictive model, which builds on historical data from course SSP to a subsequent offering of the same course SSA (Bayesian updating transfer SSP –> SSA) and a predictive model built for one course STEM to a different course SSA (Bayesian updating transfer STEM –> SSA).

### B. Simple LR

In the regression model, the response variable takes two possible values, 0 and 1, which classifies students into fail and pass groups. Let the response variable $y_i$ be a Bernoulli random variable with the probability of pass $\pi_i$

$$y_i \sim \text{Bernoulli}(\pi_i). \tag{1}$$

We assume a linear relationship between the input variables $X_i = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{iK})$ and the log-odds of the response $y_i = 1$, which can be written as the following equation:

$$\log_b \frac{\pi_i}{1-\pi_i} = \beta_0 x_{i0} + \ldots + \beta_1 x_{ik} + \ldots + \beta_K x_{iK} \tag{2}$$

where $\pi_i$ is the probability that each student passes the course, $x_{i0} = 1$, $x_{ik}$ is the $i$th observation for the $k$th input, $\beta_0$ is the constant term, and $\beta_k$ is the coefficient for variable $k$. The natural logarithm is used in the log-odds function. Optimal model parameters can be obtained through maximum likelihood estimation (MLE), where the log-likelihood function of all observations can be calculated as follows:

$$\text{lnL}\ (y,\beta) = \sum_{i=1}^{n} y_i \sum_{k=0}^{K} \beta_k x_{ik} - \sum_{i=1}^{n} \ln\left(1 + \exp\left(\sum_{k=0}^{K} \beta_k x_{ik}\right)\right) \tag{3}$$

### C. Bayesian LR

The traditional MLE identifies point estimates of parameters to construct a regression model [31]. On the contrary, the Bayesian LR employs an adaptive formula to update the posterior distribution of model parameters when new data become available. With this characteristic, Bayesian LR can be utilized to enhance the performance of predictive models limited by their transferability among different data sources, such as courses over time and different courses. Moreover, Bayesian models have the merit of avoiding the overfitting issue caused by scarce training data with immense feature size [32], [33]. The Bayesian LR framework is described as follows. Similar to the simple LR, the response variable $y_i$ can be modeled as a Bernoulli random variable and the probability of a student pass is $\pi_i$. Solving (2) for $\pi_i$ becomes

$$\pi_i = \frac{e^{\sum_{k=0}^{K} \beta_k x_{ik}}}{\left(1 + e^{\sum_{k=0}^{K} \beta_k x_{ik}}\right)} \tag{4}$$

Then, the likelihood function of $Y = (y_1, \ldots, y_n)^T$ can be calculated as follows:

$$\begin{aligned} f\ (Y|\beta) &= \prod_{i=1}^{N} \pi_i \\ &= \prod_{i=1}^{N} \left[\left(\frac{e^{\sum_{k=0}^{K} \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^{K} \beta_k x_{ik}}}\right)^{y_i} \left(1 - \frac{e^{\sum_{k=0}^{K} \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^{K} \beta_k x_{ik}}}\right)^{(1-y_i)}\right]. \end{aligned} \tag{5}$$

Further, Bayesian inference is applied to obtain the posterior distribution of $\boldsymbol{\beta}$. According to Bayes' theorem, the posterior distribution of parameters can be estimated as follows:

$$f\ (\beta|Y) = \frac{f(Y,\beta)}{f(Y)} = \frac{f(Y|\beta)\pi(\beta)}{f(Y,\beta)d\beta} \propto f(Y|\beta)\pi(\beta) \tag{6}$$

where $f(\beta|Y)$ is the posterior joint probability distribution of model parameters $\boldsymbol{\beta}$ conditional on the data $Y$, $f(Y|\beta)$ is the likelihood function in (5), and $\pi(\beta)$ is the prior distribution of parameters $\boldsymbol{\beta}$. In our study, we assume an independent normal prior distribution for each model parameter

$$\beta \sim N(\boldsymbol{\mu}, \sigma \boldsymbol{I}_K) \tag{7}$$

where $\boldsymbol{\mu}$ is the mean vector of the prior distribution, and $\boldsymbol{I}_K$ is a $K$ dimensional identity matrix. Substitute (5) for $f(Y|\beta)$ and the normal distribution function for $\pi(\beta)$, the posterior joint probability distribution $f(\beta Y)$ can be written as follows:

$$\begin{aligned} \pi(\beta|Y) \propto \prod_{i=1}^{N} &\left(\frac{e^{\sum_{k=0}^{K} \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^{K} \beta_k x_{ik}}}\right)^{y_i} \\ &\left(1 - \frac{e^{\sum_{k=0}^{K} \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^{K} \beta_k x_{ik}}}\right)^{(1-y_i)} \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(\beta-\mu)^2}{2\sigma^2}\right\}. \end{aligned} \tag{8}$$

### D. Bayesian Updating Approach

Based on the abovementioned theory, we use Bayesian inference to update the predictive model of student pass and fail to achieve improved accuracy. The inference will improve the LR model developed from historical data of online courses using a new data source or transferring a model to a different course. Let us suppose that a model has been constructed by the data $Y_1$, and the new data $Y_2$ has become available from a new course. Hence, the posterior distribution of model parameters $\boldsymbol{\beta}$ conditional on data $Y_1$ and $Y_2$ can be derived based on Bayes' theorem as follows:

$$\begin{aligned} \pi(\beta|Y_1, Y_2) \propto f(Y_1, Y_2|\beta)\pi\ (\beta) = &\ f(Y_2|Y_1, \beta) f(Y_1|\beta)\pi(\beta) \\ &\propto f(Y_2|\beta)\pi(\beta|Y_1). \end{aligned} \tag{9}$$

As seen in the equation, the distribution of model parameter $\beta$ can be updated from the distribution of the model parameters

**Algorithm 1** Transfer learning

Initialization;
**Given** $Logit(X \mid \widehat{\beta})$ developed for a previous class
$(X, Y)$ from a new class
Define Prior$(\beta) \sim MVN(\widehat{\beta}, \mathbb{I})$
**procedure** PARTITION$(X, Y)$ ▷ Partitioning into training and testing datasets
Update model using training dataset
Compute likelihood of $Y$ as $f(Y \mid \widehat{\beta})$
Obtain posterior distribution as $f(\widehat{\beta} \mid Y) = f(Y \mid \widehat{\beta}) prior(\beta)$
Randomly generate $n$ samples from $f(\widehat{\beta} \mid Y)$
Update $\widehat{\beta}$ with the mean of $n$ samples
**procedure** EVALUATION ▷ Evaluating performance using testing dataset
Make prediction $\widehat{Y} = Logit(X \mid \widehat{\beta})$
Compute $FPR, TPR, AUC, ROC$ curves

Fig. 1. Pseudocode of transfer learning algorithm.

estimated using data $Y_1$ as well as the likelihood of observing $Y_2$ given the parameters from $Y_1$. The difference between LR models and Bayesian LR models is the prior distribution. When updating an existing model, the prior distributions are adjusted with the updated prior distribution $\pi(\boldsymbol{\beta} \mid Y_1)$ that are obtained from the existing model developed by data $Y_1$. With the new source of data $Y_2$, this method can be applied to an updated model for estimating parameters. In short, a Bayesian updating approach is a practical and effective way to update original predictive models. An overview of the algorithmic procedure is illustrated in Fig. 1.

### E. Receiver Operating Characteristic (ROC) Curve

The ROC curve is a common tool used to assess the predictive performance of models with binary outcomes [34]. Once the predicted probability of the observation exceeds a threshold, the binary outcome (event $= 1$ and nonevent $= 0$) of a model classifies an observation as either an event (with this grade, the student can pass the class) or a nonevent (with this grade, the student fails the class). The predictive performance of a model with binary outcome can be measured by two complementary indicators: 1) the true positive rate (sensitivity) that is the proportion of events predicted as an event; and 2) the true negative rate (specificity) that is the proportion of nonevents predicted as a nonevent. Both specificity and sensitivity depend on the same probability threshold falling between zero and one. The ROC curves also indicate the relationship between the sensitivity and the false alarm rate (one–specificity) for thresholds from zero to one. To draw a ROC curve, we need to calculate the false alarm rate and sensitivity for multiple thresholds varying between zero and one. The area under the ROC curve (AUC) can be regarded as an index for evaluating each model's predictive performance. The AUC ranges between zero and one. Furthermore, a larger AUC stands for greater predictive performance. With this method, we can judge each model objectively. We calculate the accuracy of our model to compare the model's predictive performance by using (10). In the calculation, we use optimum threshold from our ROC curve.

$$\text{Accu} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{10}$$

where *Accu* is the accuracy of predictive models, *TP* is number of true positive instances, *TN* is the number of true negative instances, *FP* is the number of false positive, and *FN* is the number of false negative instances of model. In sum, the final performance is evaluated using ROC, AUC, false alarm rate, and sensitivity. False alarm is defined as a student who passed the course was predicted to fail the course by the model. Sensitivity is the true positive rate measuring the proportions of students that are correctly predicted.

### F. Experimental Setup

This section describes the experimental process. First, after conducting feature selection in the original feature set, we used these datasets to build the Bayesian LR model for SSA, the target course for model to transfer to, using five-fold cross validation, where the data is partitioned evenly into five folds. Models were built for SSP and STEM as well. We further compared their model parameters to show the difference in these models as indirect evidence for indicating models building on historical data cannot directly be transferred to other courses. Second, we directly applied the models built for SSP and STEM courses to the SSA course data to show the application of predictive models built from historical data of the same course to a subsequent course (SSP –> SSA) and to a different course (STEM –> SSA). This step is to test the model built in prior time (SSP) or for a different course (STEM) in unseen course data (SSA). This step will produce direct evidence of the effectiveness of model transferability over time and across courses. Third, a Bayesian updating approach was introduced to update the SSP and STEM models first and then applied to the SSA course (updated SSP –> SSA; updated STEM –> SSA). We also further examined the sensitivity of the updated model by using different subsamples of the course (SSA) and compared the updated model performance with the predictive models based on the subsamples of the SSA data to show robustness of the proposed Bayesian updating approach. Each experiment was run ten times, and subsamples were randomly selected. All experiments were done in MATLAB 2020b. The built-in generalized linear model function was used to develop the LR models. The Bayesian code was developed by our research team.

## IV. Results of Data Analysis

### A. Comparison of the Predictive Models for Different Courses

Bayesian LR was used to develop the predictive models for SSP, SSA, and STEM independently to show differences in model parameters using five fold cross validation. To evaluate the possible correlations between these course variables in Table I, we take the following steps to pick the best combination of course features for the final models.

1) *Univariate Analysis*: Simple standard LR analyses (one variable at a time) were conducted to identify the variables (the collection of variables across courses) that are significantly correlated with the binary outcome (pass or fail).

TABLE II
COMPARISON OF EACH MODEL PARAMETERS WITH MEAN AND STANDARD ERROR (SE)

| Variables | SSP | | SSA | | STEM | |
|---|---|---|---|---|---|---|
| | Model 1 | | Model 2 | | Model 3 | |
| | Mean | SE | Mean | SE | Mean | SE |
| Constant | 5.858 | 0.581 | 2.453 | 0.338 | 5.162 | 0.542 |
| assessment | -0.029 | 0.003 | 0.059 | 0.027 | -0.016 | 0.007 |
| dataplusCnt | - | - | - | - | -0.239 | 0.105 |
| dualpaneCnt | - | - | - | - | 0.001 | 0.0005 |
| highEd | -0.356 | 0.112 | -0.394 | 0.103 | -0.049 | 0.006 |
| homepageCnt | - | - | - | - | -0.074 | 0.022 |
| imdBand | - | - | - | - | 0.056 | 0.025 |
| oucollaborateCnt | - | - | -0.021 | 0.003 | - | - |
| oucontentCnt | - | - | -0.029 | 0.014 | -0.003 | 0.0007 |
| ouwikiCnt | - | - | - | - | -0.0007 | 0.0003 |
| quizCnt | -0.049 | 0.007 | -0.003 | 0.0004 | - | - |
| urlCnt | - | - | -0.004 | 0.001 | 0.008 | 0.003 |

TABLE III
PREDICTIVE PERFORMANCE OF MODELS 1 AND 2 ON SSA

| False Alarm | Sensitivity | | Difference |
|---|---|---|---|
| | Model 2 | Model 1 | |
| 0.05 | 0.36 | 0.33 | 4% |
| 0.10 | 0.47 | 0.42 | 5% |
| 0.15 | 0.56 | 0.49 | 5% |
| 0.20 | 0.63 | 0.54 | 9% |
| 0.25 | 0.67 | 0.60 | 8% |
| 0.30 | 0.73 | 0.71 | 3% |
| AUC | 0.808 | 0.752 | .056 |

2) Stepwise logistic analysis was then applied to select predictive variables from the set of variables identified in univariate analysis. The final selected variables were used to develop the Bayesian LR models.
3) The Markov Chain Mont Carlo (MCMC) method [35] was then used to estimate the posterior probability distribution of the model parameters.

The estimates of the model parameters and their mean and standard error are given in Table II. These models are built directly based on their own course data. As expected, these models are quite different from each other. Some course variables that are significant in one model are insignificant in other models. To some extent, this reflected that models may not transfer well from a historical course to a subsequent one (Model 1 vs. Model 2 or SSP –> SSA) or transfer from one course to another (Model 3 vs. Model 2 or STEM –> SSA). Given the significant difference in terms of constructed models across courses, these results can serve as an initial and indirect evidence to demonstrate that direct application of a predictive model built on historical data to a future course or to a different course may not be feasible. The next section will directly apply the model built using historical data to other courses to examine whether direct transfer works or not.

## B. Predictive Model Direct Transferability Results

First, to test whether a predictive model built using historical data can be directly applied to a future offering of the same course, we applied the model built for SSP (Model 1) directly to course SSA. Table III summarized the predictive performance of Model 1 (as in Table II) on course SSA. For comparison, the predictive performance of Model 2 in Table II on course SSA is also given in Table III. The sensitivity of Model 1 is lower than that of Model 2 at different false alarm rates.

TABLE IV
PREDICTIVE PERFORMANCE OF MODELS 2 AND 3 ON SSA

| False Alarm | Sensitivity | | Difference |
|---|---|---|---|
| | Model 2 | Model 3 | |
| 0.05 | 0.36 | 0.36 | 1% |
| 0.10 | 0.47 | 0.40 | 7% |
| 0.15 | 0.56 | 0.43 | 12% |
| 0.20 | 0.63 | 0.48 | 15% |
| 0.25 | 0.67 | 0.53 | 15% |
| 0.30 | 0.73 | 0.56 | 18% |
| AUC | 0.808 | 0.664 | .144 |

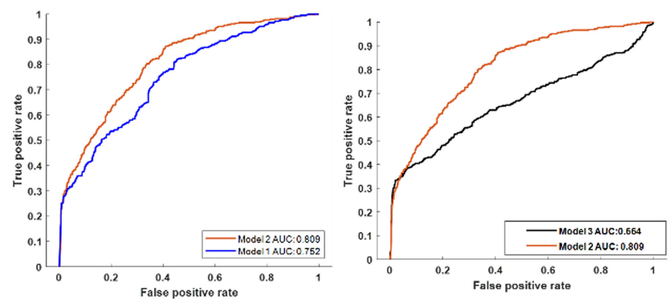



Fig. 2. ROC curves and area under curve of Model 1 and 3 on SSA.

The average difference in sensitivity between the two models is 6%. The result shows that Model 1 cannot accurately predict the pass/fail of course SSA. That is, directly applying the predictive model developed using historical data cannot perform well on the future offering of the same course.

Second, to investigate whether a predictive model for one course can directly transfer to another course, we applied the model built for STEM (Model 3) directly to SSA. Table IV summarized the prediction performance of Model 3 on course SSA. Similarly, we included the sensitivities of Model 2 (as in Table II) on course SSA in Table IV for comparison. The average difference between models 2 and 3 is 11%. This indicated that Model 3 was not effective in predicting student success in course SSA, which means that a predictive model built for one course is difficult to directly apply to a different course. Fig. 2 shows the results of ROC curves and area under curve of applying models 1 and 3 directly to SSA.

## C. Bayesian Updating Results

The Bayesian updating approach was used to improve the transferability of the predictive model over time and between courses. It can update an "old" or source model into an application context using only small amounts of new data in the target course. To investigate the effects of the sample size of the new data on updating performance, 2.5%, 5.0%, 10%, 30%, 50%, and 80% samples of the SSA course were randomly selected.

TABLE V
PREDICTIVE PERFORMANCE OF UPDATED MODEL 1 ON SSA

| False Alarm | Sensitivity of Model 2 | Sensitivity of Model 1 | Sensitivity of Model 1 updated by different sub-samples (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2.50% | 5% | 10% | 30% | 50% | 80% |
| 0.05 | 0.36±0.036 | 0.33 | 0.21±0.102 | 0.33±0.019 | 0.36±0.028 | 0.36±0.016 | 0.37±0.009 | 0.37±0.007 |
| 0.10 | 0.47±0.030 | 0.42 | 0.33±0.075 | 0.43±0.016 | 0.46±0.029 | 0.46±0.015 | 0.46±0.012 | 0.47±0.006 |
| 0.15 | 0.56±0.057 | 0.49 | 0.40±0.092 | 0.52±0.029 | 0.55±0.041 | 0.56±0.027 | 0.54±0.014 | 0.55±0.014 |
| 0.20 | 0.63±0.057 | 0.54 | 0.46±0.095 | 0.59±0.034 | 0.62±0.050 | 0.64±0.028 | 0.61±0.020 | 0.63±0.013 |
| 0.25 | 0.68±0.032 | 0.60 | 0.50±0.093 | 0.66±0.030 | 0.68±0.044 | 0.70±0.028 | 0.67±0.020 | 0.68±0.011 |
| 0.30 | 0.73±0.027 | 0.71 | 0.55±0.090 | 0.72±0.036 | 0.72±0.043 | 0.75±0.019 | 0.73±0.022 | 0.74±0.012 |
| AUC | 0.81±0.008 | 0.75 | 0.74±0.039 | 0.78±0.020 | 0.79±0.010 | 0.81±0.006 | 0.81±0.006 | 0.81±0.002 |

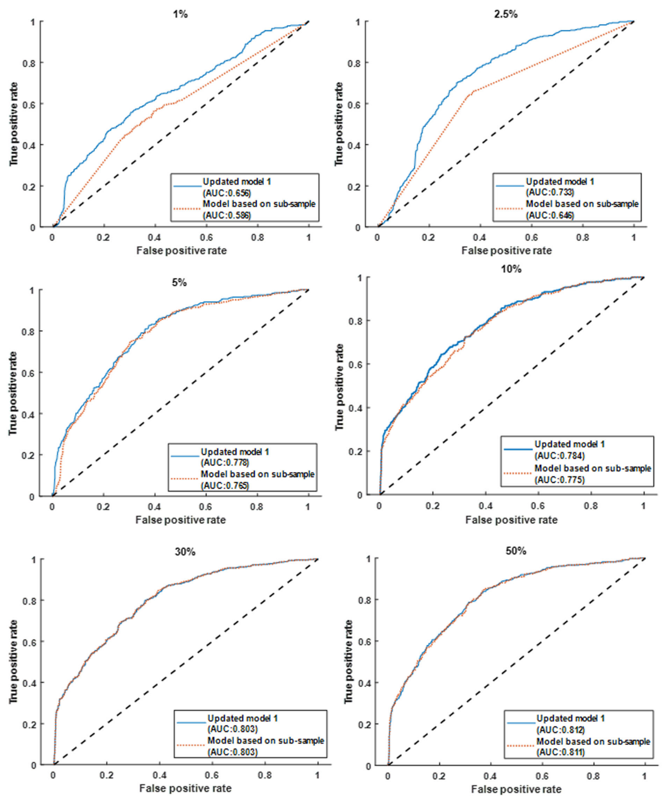



Fig. 3. ROC curves and area under curve of updated Model 1 and the model for each subsample.

To examine whether the Bayesian updating approach can improve the temporal transferability of the predictive model built on historical data (SSP, Model 1) to the future offering of the same course (SSA), the prediction performance of the updated Model 1 (as in Table III) was evaluated on SSA. Table V summarizes the prediction performance of Model 1, updated Model 1, and Model 2 (built from SSA data) on the SSA course. Updated Model 1 provides higher sensitivities than Model 1, even for the minimum subsample that only contains 2.5% of the samples from the SSA data at most false alarm rates. The performance of the updated Model 1 improves as the subsample size increases from 2.5% to 80%. As shown in Table V, when 30% of the samples is used to update Model 1, the prediction performance of updated Model 1 is very close to or even surpasses that of Model 2, showing that the Bayesian updating approach is effective in improving the temporal transferability of the predictive models with low data requirements.

To better illustrate the advantages of the Bayesian updating approach, we compared the ROC curves of the updated Model 1 and Model 2 developed using different subsamples. Specifically, 1%, 2.5%, 3%, 10%, 30%, and 50% of the samples in SSA data were used to update Model 1 and to derive Model 2, respectively. The ROC curves and the AUCs of the two models built from each subsample are illustrated in Fig. 3. The prediction performance of updated Model 1 is noticeably better than that of Model 2 for small subsamples, particularly for sample size less than 30%. When the subsample size is greater than 30%, the AUCs of Model 2 come close to the updated Model 1. This shows the robustness of the Bayesian updating approach for sample size; it has significant advantages over the traditional models especially when data are limited.

We then investigated whether the Bayesian updating approach was able to enhance the transferability of a predictive model built for one course (STEM, Model 3) to another course (SSA). Similarly, subsamples of the SSA course were applied to update Model 3 (as in Table III). The prediction performance of Model 3, updated Model 3, and Model 2 on the SSA course are provided in Table VI. As seen, subsamples of all percentages consistently improve the sensitivities of Model 3 at most false alarm rates, and the updated Model 3 can achieve comparable sensitivities to Model 2 using only 30% of the samples. The predictive performance of updated Model 3 increases as the amount of data increases. Therefore, the Bayesian updating approach is also effective in improving the transferability of predictive model between different courses.

We also further compared the ROC curves of the updated Model 3 and Model 2 developed using only different subsamples. Similarly, 1%, 2.5%, 3%, 10%, 30%, and 50% of the samples in SSA data were used to update Model 3 and to build Model 2, respectively. Fig. 4 demonstrates that the ROC curves of updated Model 3 generally outperform that of Model 2 for all subsamples. The AUCs of the updated Model 3 are significantly higher than those of Model 2 for sample size less than 30%. This indicates that the proposed Bayesian updating approach is robust and more effective for building a prediction model for a different course if the available data is limited.

We further compared the predictive performance between the Bayesian updating approach (BU) and other classical machine learning algorithms when supplemented with different ratios of subsamples. The benchmark models are LR, Support Vector Machine (SVM), and Random Forest (RF). Table VII and Fig. 5 examine model transferability within the same course during different terms, namely SSP –> SSA. In general,

TABLE VI
PREDICTIVE PERFORMANCE OF UPDATED MODEL 3 ON SSA

| False Alarm | Sensitivity of Model 2 | Sensitivity of Model 3 | Sensitivity of Model 3 updated by different sub-samples (%) 2.50% | 5% | 10% | 30% | 50% | 80% |
|---|---|---|---|---|---|---|---|---|
| 0.05 | 0.36±0.036 | 0.36 | 0.09±0.128 | 0.31±0.048 | 0.32±0.056 | 0.32±0.011 | 0.36±0.011 | 0.37±0.012 |
| 0.10 | 0.47±0.030 | 0.40 | 0.17±0.200 | 0.40±0.026 | 0.42±0.044 | 0.48±0.018 | 0.45±0.007 | 0.46±0.007 |
| 0.15 | 0.56±0.057 | 0.43 | 0.27±0.232 | 0.48±0.040 | 0.53±0.046 | 0.55±0.019 | 0.55±0.015 | 0.56±0.017 |
| 0.20 | 0.63±0.057 | 0.48 | 0.46±0.178 | 0.54±0.038 | 0.57±0.042 | 0.66±0.029 | 0.62±0.020 | 0.62±0.011 |
| 0.25 | 0.68±0.032 | 0.53 | 0.49±0.189 | 0.58±0.038 | 0.64±0.036 | 0.73±0.021 | 0.68±0.017 | 0.68±0.011 |
| 0.30 | 0.73±0.027 | 0.56 | 0.51±0.200 | 0.68±0.039 | 0.68±0.042 | 0.76±0.018 | 0.74±0.012 | 0.74±0.007 |
| AUC | 0.81±0.008 | 0.66 | 0.74±0.032 | 0.77±0.018 | 0.79±0.017 | 0.80±0.005 | 0.81±0.003 | 0.81±0.002 |

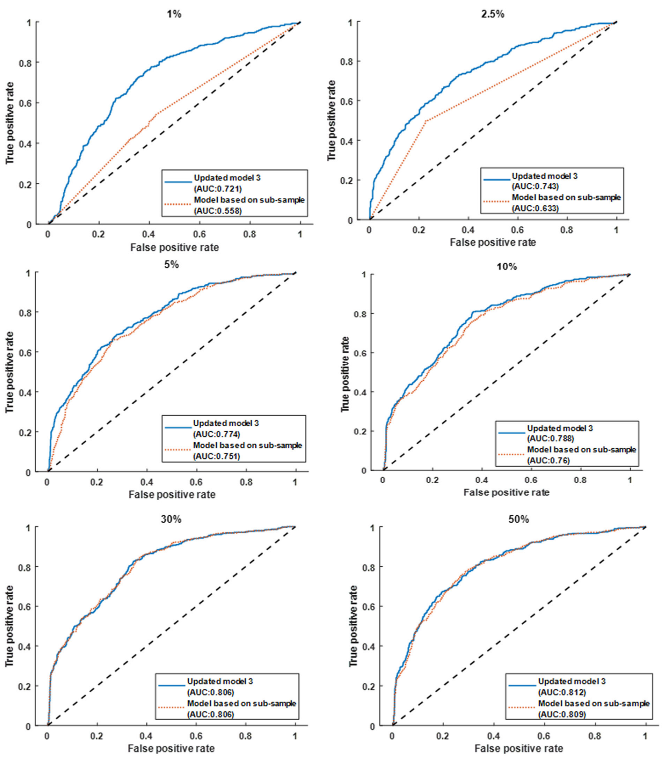



Fig. 4. ROC curves and area under curve of updated Model 3 and the model for each subsample.

the figure suggests that BU can outperform SVM and its direct counterpart LR when supplemented with a small-medium portion (0.05–0.4) of the data from the target context. Interestingly, BU's advantages over SVM and LR gradually decreases as the subsample ratio increases. Compared with RF, BU can outperform or show a comparable performance when the false alarm rate is set at a relatively low value (e.g., 0.05–0.15). However, RF tends to outperform BU when the false alarm rate increases. It is worth noting that BU tends to yield a more stable performance compared with the benchmarks given that the standard deviation of BU is smaller. The error band of BU starts wide, but gradually converges to a narrow width as the subsample ratio increases, while the error band of the benchmarks tend to stay wide or even show a tendency of increase.

Table VIII and Fig. 6 examine model transferability in different contexts, namely STEM –> SSA. Similar to Fig. 5, BU has shown a general advantage over SVM and LR when supplemented with data from SSA. Interestingly, what distinguishes from Fig. 5 is that BU has demonstrated a constant advantage over its counterpart LR when at least 5% of SSA subsamples are available. Likewise, BU shows a slightly better or comparable performance than RF when the false alarm rate is low, while the results reverse when the false alarm rate becomes greater. Moreover, BU still shows a more robust performance with a narrow error band, while the benchmark models tend to have wider error band when subsample ratios increase.

## V. DISCUSSION

Little research has been conducted to examine the transferability of predictive models built on historical datasets to future offerings of the same course. We also have a limited understanding of the transferring of a performance predictive model to a different course. A case to point is that there is no concrete and systematic knowledge of whether we can directly transfer the predictive model building on historical data to a future course or to a different one. Though some research has explored transferring predictive models [11], [15], [30], most have focused on black box data mining methods such as neural networks and ensemble models. These methods are often not transparent, and the effects of the predictors are less interpretable. This study advances our knowledge and understanding by investigating the feasibility and effectiveness of directly applying a predictive model to the same future course and to a completely different course. We also propose a Bayesian updating approach to improve the transferability of predictive models.

This study found that directly using the predictive model built on historical data with a subsequent offering of the same course showed a significant drop in terms of prediction performance. One possible reason for such a drop is because students have different characteristics, and course design may evolve over time. Compared with previous studies [10], [26], our research provided more evidence than merely the prediction accuracy as we thoroughly investigated the model properties and performance by building models independently for each course. The finding of prediction accuracy decrease is consistent with Boyer and Veeramachaneni's work [10] and Gitinabard's research [28] but disagreed with He's discovery [29]. Further, our study also found that directly applying a predictive model built for one course to a different course does not work effectively either. Apparently, different courses have significantly different instructional design and emphasis and result in tremendous differences in terms of data and predictive model parameters.

More importantly, this study proposed the Bayesian updating approach to improve the transferability of the predictive models.

TABLE VII
PREDICTIVE PERFORMANCE OF UPDATED MODEL 1 AND ITS BENCHMARKS ON SSA: LOGISTIC REGRESSION (LR), RANDOM FOREST (RF), AND SVM

| False Alarm | Sensitivity of Model 2 | Model | Sensitivity of model 3 updated by different sub-samples (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2.50% | 5% | 10% | 30% | 50% | 80% |
| | | LR | 0.32±0.047 | 0.32±0.034 | 0.33±0.026 | 0.33±0.034 | 0.34±0.030 | 0.32±0.085 |
| 0.05 | 0.36±0.036 | RF | 0.23±0.028 | 0.25±0.036 | 0.27±0.036 | 0.29±0.035 | 0.30±0.051 | 0.33±0.070 |
| | | SVM | 0.09±0.020 | 0.10±0.018 | 0.11±0.018 | 0.14±0.018 | 0.15±0.028 | 0.17±0.050 |
| | | LR | 0.44±0.037 | 0.43±0.034 | 0.43±0.028 | 0.44±0.023 | 0.47±0.034 | 0.47±0.071 |
| 0.10 | 0.47±0.030 | RF | 0.39±0.035 | 0.40±0.036 | 0.42±0.028 | 0.45±0.041 | 0.45±0.053 | 0.46±0.098 |
| | | SVM | 0.19±0.039 | 0.20±0.035 | 0.21±0.029 | 0.24±0.039 | 0.25±0.046 | 0.30±0.079 |
| | | LR | 0.54±0.018 | 0.55±0.016 | 0.55±0.020 | 0.54±0.024 | 0.57±0.033 | 0.56±0.048 |
| 0.15 | 0.56±0.057 | RF | 0.51±0.028 | 0.53±0.039 | 0.55±0.023 | 0.56±0.034 | 0.56±0.039 | 0.55±0.074 |
| | | SVM | 0.33±0.052 | 0.33±0.062 | 0.33±0.047 | 0.37±0.041 | 0.39±0.069 | 0.44±0.089 |
| | | LR | 0.60±0.014 | 0.61±0.015 | 0.61±0.011 | 0.61±0.017 | 0.62±0.021 | 0.61±0.042 |
| 0.20 | 0.63±0.057 | RF | 0.61±0.024 | 0.63±0.035 | 0.65±0.022 | 0.66±0.030 | 0.64±0.033 | 0.69±0.066 |
| | | SVM | 0.47±0.060 | 0.48±0.064 | 0.50±0.058 | 0.54±0.044 | 0.54±0.080 | 0.56±0.105 |
| | | LR | 0.66±0.014 | 0.66±0.009 | 0.66±0.009 | 0.65±0.015 | 0.65±0.019 | 0.65±0.039 |
| 0.25 | 0.68±0.032 | RF | 0.68±0.018 | 0.70±0.024 | 0.72±0.022 | 0.73±0.023 | 0.72±0.038 | 0.74±0.087 |
| | | SVM | 0.59±0.065 | 0.62±0.046 | 0.64±0.035 | 0.68±0.037 | 0.68±0.051 | 0.69±0.093 |
| | | LR | 0.71±0.018 | 0.70±0.015 | 0.70±0.015 | 0.70±0.018 | 0.69±0.024 | 0.70±0.051 |
| 0.30 | 0.73±0.027 | RF | 0.74±0.019 | 0.76±0.020 | 0.77±0.016 | 0.79±0.022 | 0.80±0.026 | 0.78±0.073 |
| | | SVM | 0.71±0.045 | 0.72±0.026 | 0.73±0.020 | 0.75±0.024 | 0.76±0.033 | 0.74±0.088 |
| | | LR | 0.78±0.012 | 0.78±0.010 | 0.78±0.007 | 0.78±0.008 | 0.79±0.013 | 0.79±0.026 |
| AUC | 0.81±0.008 | RF | 0.79±0.007 | 0.80±0.009 | 0.81±0.007 | 0.82±0.010 | 0.82±0.013 | 0.82±0.031 |
| | | SVM | 0.73±0.010 | 0.73±0.010 | 0.74±0.007 | 0.77±0.011 | 0.78±0.020 | 0.78±0.034 |

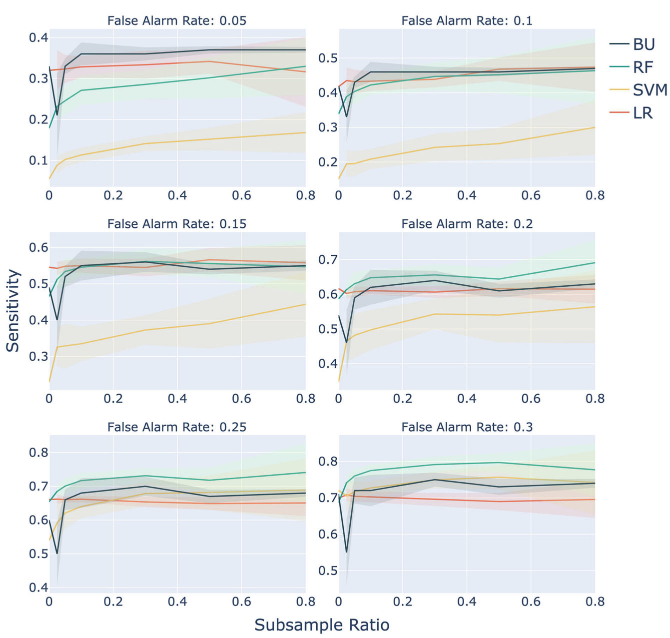



Fig. 5. Sensitivity for each false alarm rate on SSP –> SSA between Bayesian Updating Method (BU) and benchmark models: Logistic Regression (LR), SVM, and Random Forest (RF).

The results demonstrated that the updated model outperforms the direct application of the model to the future offering of the course or to a different course. Compared with previous attempts to build transferrable models [15], [30], this Bayesian updating approach is a white box model, which can help researchers and practitioners understand the significance of different features and how they contribute to the prediction of students passing or failing a course. For example, the odd ratios of the regression model can pinpoint significant predictors and reveal their negative or positive effects on student learning outcomes. We further showed the robustness of the updating approach by using different sizes of subsamples. The results indicated that the updated model can achieve superior performance even with small amounts of data.

The advantage of the Bayesian transfer-learning model or the contribution of this approach is that the model can be continuously updated with new collected data rather than building the model from scratch every time, which reduces both the effort of model development and computational cost of analyzing a large amount of historical data from multiple courses. In other words, model building using data from historical courses can be quickly updated using very recent collected data, making the model much closer to the live course. Also, a model building from other courses can be potentially transferred to a totally different course, which reduces the effort taken in model development, as well as computational cost. Overall, this approach proposes a new methodological framework to carry experience and knowledge from past and other courses forward to new courses.

These findings have important implications for practice. To implement an early warning system, it is not that effective to directly employ a predictive model built on historical data sources in a subsequent offering of the same course or using a model built for one course to a different one. It is also not cost-effective to build predictive models from scratch every time that new course data is added or build a model for every course. Predictive model can be easily updated using the data in the target course so that the model can perform better in identifying at-risk students. Essentially, this Bayesian updating approach provides a methodological framework to dynamically update models over time and across courses. Also, identifying at-risk students is only one part of the solution. Teachers can depend on the predictive modeling results using our white box approach to provide more specific advice to students.

While this Bayesian updating approach has its unique advantages, there are some potential limitations for this approach. First, the sample this approach needs is small, but it requires data from the "target" course to update the original model. Logistically, it

TABLE VIII
PREDICTIVE PERFORMANCE OF UPDATED MODEL 3 AND ITS BENCHMARKS ON SSA: LOGISTIC REGRESSION (LR), RANDOM FOREST (RF), AND SUPPORT VECTOR MACHINE (SVM)

| False Alarm | Sensitivity of Model 2 | Model | Sensitivity of model 3 updated by different sub-samples (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2.50% | 5% | 10% | 30% | 50% | 80% |
| | | LR | 0.32±0.010 | 0.22±0.008 | 0.22±0.015 | 0.24±0.025 | 0.26±0.032 | 0.26±0.095 |
| 0.05 | 0.36±0.036 | RF | 0.32±0.035 | 0.32±0.046 | 0.32±0.042 | 0.30±0.043 | 0.32±0.054 | 0.34±0.086 |
| | | SVM | 0.22±0.031 | 0.24±0.034 | 0.25±0.026 | 0.29±0.030 | 0.30±0.032 | 0.30±0.058 |
| | | LR | 0.32±0.049 | 0.35±0.024 | 0.37±0.013 | 0.37±0.030 | 0.42±0.036 | 0.42±0.065 |
| 0.10 | 0.47±0.030 | RF | 0.45±0.028 | 0.45±0.032 | 0.46±0.028 | 0.46±0.044 | 0.47±0.061 | 0.46±0.078 |
| | | SVM | 0.37±0.014 | 0.37±0.026 | 0.38±0.036 | 0.43±0.026 | 0.44±0.041 | 0.46±0.074 |
| | | LR | 0.37±0.056 | 0.41±0.034 | 0.045±0.015 | 0.48±0.025 | 0.50±0.031 | 0.50±0.054 |
| 0.15 | 0.56±0.057 | RF | 0.55±0.024 | 0.55±0.031 | 0.56±0.025 | 0.58±0.052 | 0.57±0.041 | 0.55±0.077 |
| | | SVM | 0.46±0.016 | 0.47±0.014 | 0.47±0.012 | 0.51±0.018 | 0.53±0.039 | 0.55±0.061 |
| | | LR | 0.43±0.028 | 0.47±0.029 | 0.50±0.016 | 0.54±0.024 | 0.56±0.027 | 0.56±0.053 |
| 0.20 | 0.63±0.057 | RF | 0.63±0.019 | 0.64±0.025 | 0.64±0.022 | 0.66±0.032 | 0.65±0.041 | 0.65±0.085 |
| | | SVM | 0.54±0.017 | 0.56±0.014 | 0.58±0.016 | 0.60±0.033 | 0.63±0.052 | 0.67±0.099 |
| | | LR | 0.51±0.025 | 0.52±0.014 | 0.55±0.017 | 0.60±0.024 | 0.61±0.031 | 0.63±0.065 |
| 0.25 | 0.68±0.032 | RF | 0.70±0.013 | 0.71±0.020 | 0.71±0.023 | 0.73±0.020 | 0.72±0.032 | 0.73±0.081 |
| | | SVM | 0.61±0.019 | 0.64±0.014 | 0.65±0.018 | 0.70±0.034 | 0.71±0.043 | 0.72±0.098 |
| | | LR | 0.60±0.016 | 0.61±0.017 | 0.63±0.017 | 0.65±0.025 | 0.66±0.032 | 0.64±0.057 |
| 0.30 | 0.73±0.027 | RF | 0.76±0.013 | 0.76±0.015 | 0.77±0.018 | 0.79±0.021 | 0.79±0.028 | 0.78±0.057 |
| | | SVM | 0.67±0.018 | 0.70±0.013 | 0.73±0.016 | 0.37±0.024 | 0.78±0.031 | 0.78±0.082 |
| | | LR | 0.74±0.011 | 0.75±0.007 | 0.76±0.006 | 0.77±0.010 | 0.77±0.015 | 0.78±0.030 |
| AUC | 0.81±0.008 | RF | 0.80±0.006 | 0.80±0.008 | 0.81±0.006 | 0.82±0.011 | 0.82±0.015 | 0.82±0.030 |
| | | SVM | 0.71±0.015 | 0.74±0.007 | 0.76±0.008 | 0.80±0.010 | 0.81±0.015 | 0.82±0.032 |

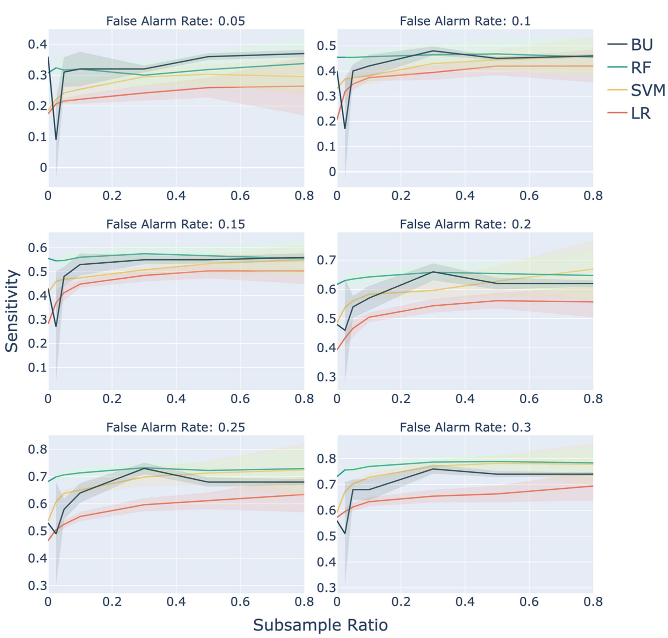



Fig. 6. Sensitivity for each false alarm rate on STEM –> SSA between Bayesian Updating Method (BU) and benchmark models: Logistic Regression (LR), SVM, and Random Forest (RF).

may still be challenging to implement this system when an institution has so many courses. One possible approach is to streamline this Bayesian updating process on the backend to improve the efficiency for building transfer models. Second, though this proposed method can transfer models over time, it can only update the model to the most recent one, not to a live course, because this method needs the student performance label from the target course to update model parameters. Still, it did not limit the contribution of this method to prevent building models from scratch whenever we need to update a historical model or transferring the model to a new course. Third, using the collected set of variables for the courses will increase model complexity and can potentially cause overfitting. To address this issue, $L_1$ Regularization can be used to penalize the regression. Fourth, this research only used three courses to conduct the experiment and obtain results. Caution should be taken in generalizing these findings.

## VI. CONCLUSION

This study systematically investigated the direct application of predictive models constructed using historical data sources to a future offering of the same course and to a different course. We further proposed the Bayesian updating approach to build the predictive model to improve model transferability. There are several future research directions. First, we plan to explore ways to update the model with several courses and analyze how to improve generalization of the predictive model in many courses. Second, this study used pass and fail as the performance indicators. It is worthwhile to explore more complex target variables such as engagement, learning trends, etc. Third, this study only examined the Bayesian updating approach using a basic white box model, logistic regression. Future studies can explore more sophisticated data models that can combine with the Bayesian updating approach.

## ACKNOWLEDGMENT

The opinions expressed are those of the authors and do not represent views of the Institution.

## REFERENCES

[1] I. E. Allen and J. Seaman, "Digital learning compass: Distance education enrollment report 2017," Babson Park, MA, USA: Babson Survey Research Group, May 2017. [Online]. Available: https://www.bayviewanalytics.com/reports/digtiallearningcompassenrollment2017.pdf

[2] D. Lederman, "Online education ascends," *Inside Higher Educ.*, Nov. 2018. [Online]. Available: https://www.insidehighered.com/digital-learning/article/2018/11/07/new-data-online-enrollments-grow-and-share-overall-enrollment

[3] I. E. Allen and J. Seaman, "Grade change: Tracking online education in the United States," Babson Survey Resarch Group, Babson Park, MA, USA, Jan. 2014. [Online]. Available: https://bayviewanalytics.com/reports/gradechange.pdf

[4] A. Becker, A. Freeman, C. Hall Giesinger, M. Cummins, and B. Yuhnke, "NMC/CoSN horizon report: 2016 K–12 edition," New Media Consortium, Austin, TX, USA, 2016. [Online]. Available: https://library.educause.edu/~/media/files/library/2017/11/2016hrk12EN.pdf

[5] M. Nichols, "Student perceptions of support services and the influence of targeted interventions on retention in distance education," *Distance Educ.*, vol. 31, no. 1, pp. 93–113, May 2010, doi: 10.1080/01587911003725048.

[6] W. Xing and D. Du, "Dropout prediction in MOOCs: Using deep learning for personalized intervention," *J. Educ. Comput. Res.*, vol. 57, no. 3, pp. 547–570, Jun. 2019, doi: 10.1177/0735633118757015.

[7] W. Xing, X. Chen, J. Stein, and M. Marcinkowski, "Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization," *Comput. Hum. Behav.*, vol. 58, pp. 119–129, May 2016, doi: 10.1016/j.chb.2015.12.007.

[8] P. Leitner, M. Khalil, and M. Ebner, "Learning analytics in higher education—A literature review," *Stud. Syst. Decis. Control*, vol. 94, pp. 1–23, 2017, doi: 10.1007/978-3-319-52977-6_1.

[9] Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systemic literature review of empirical evidence," *J. Educ. Technol. Soc.*, vol. 17, no. 4, pp. 49–64, Oct. 2014. [Online]. Available: https://www.ds.unipi.gr/et&s/journals/17_4/4.pdf

[10] S. Boyer and K. Veeramachaneni, "Transfer learning for predictive models in massive open online courses," in *Proc. 17th Int. Conf. Artif. Intell. Educ.*, Madrid, Spain, Jun. 2015, pp. 54–63, doi: 10.1007/978-3-319-19773-9_6.

[11] M. Ding, E. Hemberg, Y. Wang, and U. M. O'Reilly, "Transfer learning using representation learning in massive open online courses," in *Proc. 9th Int. Conf. Learn. Anal. Knowl*, Tempe, AZ, USA, Mar. 2019, pp. 145–154, doi: 10.1145/3303772.3303794.

[12] C. Xu, W. Wang, P. Liu, R. Guo, and Z. Li, "Using the bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models," *Transp. Res. Part C, Emerg. Technol.*, vol. 38, pp. 167–176, Jan. 2014, doi: 10.1016/j.trc.2013.11.020.

[13] C. Shew, A. Pande, and C. Nuworsoo, "Transferability and robustness of real-time freeway crash risk assessment," *J. Saf. Res.*, vol. 46, pp. 83–90, Sep. 2013, doi: 10.1016/j.jsr.2013.04.005.

[14] Ł. Kidziński, K. Sharma, M. S. Boroujeni, and P. Dillenbourg, "On generalizability of MOOC models," in *Proc. 9th Int. Conf. Educ. Data Mining*, Raleigh, NC, USA, Jun./Jul. 2016. [Online]. Available: https://www.educationaldatamining.org/EDM2016/proceedings/paper_85.pdf

[15] M. L. Bote-Lorenzo and E. Gómez-Sánchez, "An approach to build in situ models for the prediction of the decrease of academic engagement indicators in massive open online courses," *J. Universal Comput. Sci.*, vol. 24, no. 8, pp. 1052–1071, Aug. 2018. [Online]. Available: https://uvadoc.uva.es/bitstream/handle/10324/31416/2018_Bote-Lorenzo_etal_-JUCS_preprint.pdf?sequence=1&isAllowed=y

[16] D. M. Olivé, D. Q. Huynh, M. Reynolds, M. Dougiamas, and D. Wiese, "A quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 171–183, Apr. 2019.

[17] H. Wan, K. Liu, Q. Yu, and X. Gao, "Pedagogical intervention practices: Improving learning engagement based on early prediction," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 278–289, Apr. 2019.

[18] S. P. M. Choi, S. S. Lam, K. C. Li, and B. T. M. Wong, "Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions," *Educ. Technol. Soc.*, vol. 21, no. 2, pp. 273–290, Apr. 2018. [Online]. Available: https://www.j-ets.net/collection/published-issues/21_2

[19] J. Whitehill, J. J. Williams, G. Lopez, C. A. Coleman, and J. Reich, "Beyond prediction: First steps toward automatic intervention in MOOC student stopout," in *Proc. 8th Int. Conf. Educ. Data Mining*, Apr. 2015, pp. 171–178, doi: 10.2139/ssrn.2611750.

[20] E. Foster and R. Siddle, "The effectiveness of learning analytics for identifying at-risk students in higher education," *Assessment Eval. Higher Educ.*, vol. 45, no. 6, pp. 842–854, Aug. 2020, doi: 10.1080/02602938.2019.1682118.

[21] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach, "Forecasting student achievement in MOOCs with natural language processing," in *Proc. 6th Int. Conf. Learn. Anal. Knowl.*, Edinburgh, Scotland, Apr. 2016, pp. 383–387, doi: 10.1145/2883851.2883932.

[22] L. C. Zhang and J. Nouri, "A systematic review of learning computational thinking through scratch in K–9," *Comput. Educ.*, vol. 141, Nov. 2019, Art. no. 103607, doi: 10.1016/j.compedu.2019.103607.

[23] D. J. Lemay and T. Doleck, "Grade prediction of weekly assignments in MOOCS: Mining video-viewing behavior," *Educ. Inf. Technol.*, vol. 25, no. 2, pp. 1333–1342, Mar. 2020, doi: 10.1007/s10639-019-10022-4.

[24] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley, "MOOC dropout prediction: How to measure accuracy?," in *Proc. 4th ACM Conf. Learn. Scale*, Cambridge, MA, USA, Apr. 2017, pp. 161–164, doi: 10.1145/3051457.3053974.

[25] P. M. Moreno-Marcos *et al.*, "Generalizing predictive models of admission test success based on online interactions," *Sustainability*, vol. 11, no. 18, Jan. 2019, Art. no. 4940, doi: 10.3390/su11184940.

[26] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment," in *Proc. 3rd Int. Conf. Learn. Anal. Knowl.*, Leuven, Belgium, Apr. 2013, pp. 145–149, doi: 10.1145/2460296.2460324.

[27] J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan, and C. Heffernan, "Population validity for educational data mining models: A case study in affect detection," *Brit. J. Educ. Technol.*, vol. 45, no. 3, pp. 487–501, May 2014, doi: 10.1111/bjet.12156.

[28] N. Gitinabard, Y. Xu, S. Heckman, T. Barnes, and C. F. Lynch, "How widely can prediction models be generalized? Performance prediction in blended courses," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 184–197, Apr. 2019.

[29] J. He, J. Baileyt, B. I. P. Rubinstein, and R. Zhang, "Identifying at-risk students in massive open online courses," in *Proc. 29th AAAI Conf. Artif. Intell.*, Austin, TX, USA, Jan. 2015, vol. 29, no. 1. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/9471

[30] A. F. Botelho, A. Varatharaj, T. Patikorn, Di. Doherty, S. A. Adjei, and J. E. Beck, "Developing early detectors of student attrition and wheel spinning using deep learning," *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 158–170, Apr. 2019.

[31] I. J. Myung, "Tutorial on maximum likelihood estimation," *J. Math. Psychol.*, vol. 47, no. 1, pp. 90–100, Feb. 2003, doi: 10.1016/S0022-2496(02)00028-7.

[32] C. Lee, P. Y. Park, and M. Abdel-Aty, "Lane-by-lane analysis of crash occurrence based on driver's lane-changing and car-following behavior," *J. Transp. Saf. Secur.*, vol. 3, no. 2, pp. 108–122, Jun. 2011, doi: 10.1080/19439962.2010.550390.

[33] S. Mitra and S. Washington, "On the nature of over-dispersion in motor vehicle crash prediction models," *Accident Anal. Prevention*, vol. 39, no. 3, pp. 459–468, May 2007, doi: 10.1016/j.aap.2006.08.002.

[34] M. Ahmed, M. Abdel-Aty, and R. Yu, "Assessment of interaction of crash occurrence, mountainous freeway geometry, real-time weather, and traffic data," *Transp. Res. Rec.*, vol. 2280, no. 1, pp. 51–59, Jan. 2012, doi: 10.3141/2280-06.

[35] B. P. Carlin and S. Chib, "Bayesian model choice via Markov chain Monte Carlo methods," *J. Roy. Statist. Soc., Ser. B*, vol. 57, no. 3, pp. 473–484, Sep. 1995, doi: 10.1111/j.2517-6161.1995.tb02042.x.

**Wanli Xing** received the B.Ed. degree in educational technology from Jilin Normal University, Siping, China, in 2009, and the Ph.D. degree in information science and learning technologies from the University of Missouri, Columbia, MO, USA, in 2016.

He is currently an Assistant Professor of Educational Technology with the University of Florida, Gainesville, FL, USA. His research interest includes how technologies can deeply transform STEM education and online learning. He creates learning environments using cutting-edge technologies, such as Artificial Intelligence and augmented reality to support learning in diverse settings. He also designs and applies machine learning to understand, assess, and optimize learning and the environments in which it occurs.

**Dongping Du** (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the China University of Mining and Technology, Beijing, China, in 2008 and 2010, respectively, and the Ph.D. degree in industrial engineering from the University of South Florida, Tampa, FL, USA, in 2015.

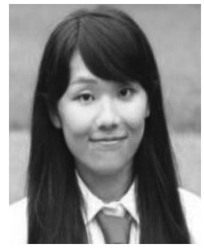
She is currently an Assistant Professor with the Department of Industrial, Manufacturing, and Systems Engineering, Texas Tech University, Lubbock, TX, USA. Her research interests include modeling and simulation, applied statistics, machine learning, and risk assessment with applications in healthcare, manufacturing, and education.

**Ali Bakhshi** received the B.S. degree from the Iran University of Science and Technology, Tehran, Iran, in 2009, and the M.S. degree from the Mazandaran University of Science and Technology, Mazandaran, Iran, in 2013, both in industrial engineerin. His master's research focused on fuzzy modeling of a multidepot vehicle routing problem. He is currently working toward the M.S. degree in systems and engineering management from the Department of Industrial, Manufacturing, and Systems Engineering, Texas Tech University, Lubbock, TX, USA.

His research interests include predictive modeling and statistical analysis of spatial-temporal datasets, and the application of Bayesian philosophy to improve statistical models.

**Kuo-Chun Chiu** received the B.E. degree from the National Taipei University of Technology, Taipei, Taiwan, in 2010, and the M.E. degree from National Tsing Hua University, Hsinchu, Taiwan, in 2012, both in chemical engineering. He is currently working toward the Ph.D. degree in industrial engineering from the Department of Industrial, Manufacturing, and Systems Engineering, Texas Tech University, Lubbock, TX, USA.

His research interests include applied statistics, predictive modeling, and computer modeling and optimization.



**Hanxiang Du** received the B.A. degree in Chinese language and literature, and the M.A. degree in literary theory from Sun Yat-sen University, Guangzhou, China, in 2011 and 2014, respectively. She is currently working toward the Ph.D. degree in educational technology from the University of Florida, Gainesville, FL, USA.

Her research interests include educational data mining, learning analytics, and K–12 computer science and engineering education.