Identifying At-Risk Online Learners by Psychological Variables Using Machine Learning Techniques

Hsiang-Yu Chien, Oi-Man Kwok, Yu-Chen Yeh, Noelle Wall Sweany,
Eunkyeng Baek, and William McIntosh
Texas A&M University

Abstract
The purpose of this study was to investigate a predictive model of online learners' learning outcomes through machine learning. To create a model, we observed students' motivation, learning tendencies, online learning-motivated attention, and supportive learning behaviors along with final test scores. A total of 225 college students who were taking online courses participated. Longitudinal data were collected over three semesters (T1, T2, and T3). T3 was used as training data given that it contained the largest sample size across all three data waves. To analyze the data, two approaches were applied: (a) stepwise logistic regression and (b) random forest (RF). Results showed that RF used fewer items and predicted final grades more accurately in a small sample. Furthermore, it selected four items that might potentially be used to identify at-risk learners even before they enroll in an online course.

Keywords: machine learning, random forest, online learning, at-risk online learners, stepwise regression, logistic regression

Chien, H.-Y., Kwok, O.-M., Yeh, Y.-C., Sweany, N.W., Baek, E., & McIntosh, W. (2020). Identifying at-risk online learners by psychological variables using machine learning techniques. *Online Learning, 24*(4), 131-146. https://doi.org/10.24059/olj.v24i4.2320

Identifying At-Risk Online Learners by Psychological Variables Using Machine Learning Techniques

More than six million students took at least one online course in 2015, according to the latest Distance Education Enrollment Report (Digital Learning Compass, 2017). One third of all students in the United States are earning credits and pursuing academic degrees through online or remote learning, and this type of learning is becoming increasingly more popular and even necessary under some circumstances (Allen & Seaman, 2013), such as during the recent coronavirus (COVID-19) pandemic. For instance, nearly 300 universities were forced to shut down all in-person, face-to-face courses, and online delivery became the major delivery model for most of schools across the United States due to the COVID-19 outbreak in March of 2020 (Foresman, 2020). However, not all students are prepared for this new teaching format and use of technology in an online learning environment and, as a result, many end up struggling.

To be successful in online courses, students need to demonstrate certain characteristics, including time management, motivation, active participation, independent learning, technology efficacy, communication, and integrity (Taormino, 2010). Given the increasing number of courses being taught online, whether partially or in some kind of hybrid form, institutions of higher learning should help students prepare for this new teaching format. To that end, identifying students who might not be prepared for online learning is a necessary first step.

Previous researchers have developed diagnostic systems for identifying at-risk learners in online learning environments based on large log data obtained from online learning platforms (Er, 2012; Jayaprakash et al., 2014; Keshtkar et al, 2016). Many online higher education providers have been targeting improvement in learning outcomes, such as course completion, final grades, and actual academic achievement (Moody, 2004).

In addition to students' online learning behaviors, psychological perspectives in terms of their needs and readiness to succeed in an online learning environment are also important and can provide a more complete picture through the use of different learning theories (Poulet et al., 2009). For example, Yeh et al. (2019) examined the effectiveness of psychological variables such as achievement goal motivation rather than typical behavioral frequency measures (e.g., log data) to predict students' learning outcomes. The authors constructed a three-path mediation model to test for the mediation effects of self-regulated learning strategies and supportive online learning behaviors on the relationship between online learners' motivation and their learning outcome (expected grade). Results showed that both self-regulated learning strategies and supportive learning behaviors played an important role in predicting the students' success in an online learning environment. Thus, these psychological variables help us understand why online learners succeed or fail. Besides, these psychological variables can be measured before students enroll in classes as a means of gaining a different perspective and useful hints about students' readiness for online courses compared with the log data collected during the learning process.

Yeh et al.'s (2019) model consisted of a 53-item questionnaire, which, despite the promising results, is too long for screening at-risk students (Kwok et al., 2019). The purpose of the present study, therefore, was to reduce the model to a more manageable size by determining key questions for identifying at-risk students. Similar to Kwok et al.' (2019), most of the studies in this area have adopted traditional statistical approaches (e.g., *t*-test, ANOVA, regression, logistic regression and structural equation modeling). As an alternative, it was decided to use machine learning techniques, which are relatively new to the online learning literature, for item selection and classification purposes.

The final goal of this study was to develop a predictive model for online learners by applying various psychological and behavioral aspects of online learners, such as motivation from the online learners to predict their actual learning outcome (i.e., students' final grade). While constructing the predictive models, we also had an intermediate goal to compare the prediction accuracy of stepwise logistic regression versus random-forest algorithm, two commonly used machine learning approaches. We then adopted the items from the predictive models to develop a screener that can effectively identify at-risk online learners.

Review of Relevant Literature

At-risk Online Learners

The characteristics of at-risk online learners are discussed in detail by Funk (2005). For example, at-risk learners are not expected to succeed, including dropping out early or failing the course. Those learners typically receive intermittent and inconsistent reinforcement for personal accomplishments and tend to demonstrate lower degrees of persistence compared to peers. In addition, procrastination is another major issue in online courses. Thus, students who are potentially at risk in distance education generally postpone and cram assignments at the last minute. Not surprisingly, these students show poorer achievement and long-term retention (Asarta and Schmidt, 2013; Tuckman, 2005).

Additional factors linked with at-risk online learners include dyslexia, low self-esteem, weak information and communications technology (ICT) skills, lack of tutor guidance and support, seldom log in or communicate online regularly, repetition of a module after failing, and lacking complete formative assessment (Hughes and Lewis, 2003; Hughes, 2007; Miller et al., 2000; Selwyn and Gorard, 2003; Wallace, 2003). Further, Osborn (2001) found that at-risk students frequently change their study environments, along with demonstrating lower motivation, less computer confidence, less encouragement to take the course, unexpected responsibilities, and an extreme internal locus of control. According to Osborn (2001), these factors are the major reason why at-risk online learners have a substantially higher rate of withdrawal from videoconferencing and web-based distance education.

Given these findings and the increasing growth in online education, the development of an effective screener that allows the instructors or educators to identify at-risk online learners based on psychological perspectives early on becomes very important. Moreover, early identification offers a sufficient time frame for potential assistance and even intervention to help at-risk online learners succeed in the online learning environment.

A Classic Data Exploratory Technique: Stepwise Logistic Regression

Funk (2005) defined at-risk online learners as those who are not expected to succeed (e.g., dropping out early or failing the course). Kwok et al. (2019) used stepwise regression to select eight items (out of 92) and then used these items to construct a structural equation model (SEM) for predicting online learners' expected grade and academic expectations. Both the learning strategy items and the failure-avoidant motivation items were found to be statistically significant in predicting expected grades. However, only the learning strategy items were statistically significant in predicting academic expectations.

Despite these important findings, Kwok et al. (2019) did not examine a predictive model for identifying which variable(s) could predict learning outcomes such as passing or failing a course. When testing models with classification purposes such as detecting at-risk students, use of logistic regression is an option. Logistic regression estimates a linear model for the log-odds ratio of the target event (Menard, 2002). Further, stepwise logistic regression selects items that may explain the variance of the log-odds ratio. For example, Bonny et al. (2000) used stepwise logistic regression to create a parsimonious classification model to detect at-risk adolescents. They selected 7 out of 12 variables that helped to build a logistic regression model. Since the stepwise logistic regression can be considered a classic data exploratory technique, we aimed decided to use is in the present study as the baseline method and compare it with another method, random forest, as detailed below.

Even though stepwise logistic regression is a classic data analysis procedure, its use is subject to several concerns (Mundry & Nunn, 2009; Steyerberg et al., 1999; Whittingham et al., 2006). First, it tends to inflate the false positive rate. Type I error rates accumulate during the iteration of a stepwise procedure, so the result tends to over-select the items. Second, it may lead to overfitting because the item selection process is not taken properly into account when fitting a logistic regression (Kuhn & Johnson, 2019)

A Machine Learning Technique: Random Forest

Because of the shortcomings of stepwise logistic regression, we considered another method of item selection and classification, random forest. Lakkaraju et al. (2015) examined multiple machine learning approaches to detect at-risk students who might not graduate from high school. The authors found that random forest (RF) outperformed other methods, including logistic regression, support vector machine (SVM), C4.5 (an algorithm for decision trees), and AdaBoost. A similar pattern was also found by Mahboob et al. (2016). A common characteristic of these studies is that they trained their model using relatively small samples. This is important in the current context, as it is sometimes difficult to collect large sets of data in educational settings. Mahboob et al. (2016) only had 60 participants, yet RF still outperformed the C4.5 algorithms and the Naive Bayes algorithm.

RF is an algorithm developed from decision trees. Decision tree is an intuitive classification algorithm that picks the criteria that can separate the target samples as tree nodes iteratively. However, it is not as predictive as other regression and classification approaches. RF can improve prediction performance by randomly selecting variable sets in each node iteration, thereby decorrelation the tree and reducing the variance. RF keeps the easy interpretation characteristic and improves the prediction accuracy; besides, it may be used with any sample size (Fassnacht et al., 2014; Mahboob et al., 2016). With some fine-tuning, RF has proven more accurate than other algorithms when the independent variables were either demographic variables (Lakkaraju et al., 2015) or system log data. However, to date, self-report questionnaires have rarely been used as training data in RF analyses. Considering its characteristics, RF with psychological variables and small sample size became one of the approaches deemed applicable to constructing a predictive model.

In summary, the present study compared stepwise followed by logistic regression and RF to investigate which approach would lead to a more accurate predictive model under a small sample condition. Eventually, the items selected in the final predictive model were to be used to create a screener that could identify potentially at-risk online learners.

Method

Participants and Procedures

Participants were recruited from a large public university in Texas. A total of three waves of data were collected across three semesters: spring 2018 (T1), fall 2018 (T2), and spring 2019 (T3). At the beginning of each semester, students who had registered for at least one online course from the College of Education and Human Development (CEHD) were invited to participate through a recruitment email that contained a description of the research purpose and an online survey created by using Qualtrics. Students who completed the survey would receive an Amazon gift card as compensation for their efforts. The survey collected basic demographic information along with various measures, including psychological needs and readiness to succeed in an online learning environment. In T1, 64 students (8 males and 56 females) completed the full survey. Of these, 54 students (84.4 %) studied within the CEHD whereas 10 (15.6 % of the total sample) were enrolled in departments outside of the CEHD. In T2, one male and 45 female students completed the full survey. Among these students in T2, 38 students (82.6%) studied within the CEHD whereas 8 students (17.4%) studied in departments outside of CEHD. Finally, in T3, 115 students (12 males and 103 females) completed the full survey, including 85 students (73.9%) studied within the CEHD and 30 students (26.1%) studied in departments outside the CEHD.

Instruments

During the data collection process, the total number of items in the survey were reduced from 196 at T1 to 65 items in T3 based on commonly used item reduction procedures such as exploratory factor analysis, and the maintained items were all theoretically sounded. For example, the 18 items of a 3 x 2 achievement goal model used in T1 were dropped from T2 given that the exploratory factor analysis could not extract the proposed 6 constructs. Thus, we kept a shorter, 2 x 2 version of the achievement goal model in T2. Similarity, potentially duplicate items with similar meaning were excluded in T2. We also deleted items that contained no variation (i.e., students responded in the exact same pattern). Given that T3 had the largest sample size and included all the items from both T1 and T2, we decided to use the T3 data as the training data, whereas T1 and T2 data were used to test the predictive model. Excluding demographics related variables such as gender and university major, we ran a stepwise regression and retained 32 items that explained 74.2% variance of the variance in our outcome variable in the final training dataset. The details of these 32 items were described below.

Reasons for Taking the Online Course (10 items)

Students' reasons for taking the online course were measured by 10 items that were adopted and modified from Pintrich et al.'s (1991) Motivated Strategies for Learning Questionnaire Manual, including "content seems interesting," "is required for all students at college," "will be useful to me in other courses," "is an easy elective," "will help improve my academic skills," "is required for major (program)," "was recommended by a friend," "was recommended by a counselor," "will improve career prospects," and "fit into my schedule." Participants were asked whether they were taking the online course for each of these reasons with two response options (Yes/ No).

Revised Online-Learning Motivated Attention and Regulation Scale (4 items; OL-MARS v.2; Wu, 2017)

The OL-MARS v.2 is based on the theory of meta-attention. Four items from the OL-MARS v.2 scale were used in the present study to measure participants' perceived attention state and use of regulatory strategies: "I turn on the computer in order to do my homework, but I still visit Facebook first," "If I postpone what I should be doing because of using the Internet, I feel guilty," "When using the computer for studying, I think of what I want to eat later or what I have just eaten," and "When I see or hear notifications from social media (e.g., Twitter, Instagram, Facebook), I cannot wait to check them." Answers were given using a 5-point Likert scale ranging from 1 ("not at all like me") to 5 ("very much like me").

The Test of Online Learning Success (5 items; TOOLS; Kerr et al., 2006)

Five items from TOOLS were adopted to assess participants' readiness for online learning, including "I require help to understand written instructions," "I can learn by working independently," "I need faculty to remind me of assignment due dates," "I am capable of solving problems alone," and "I am self-disciplined when it comes to my studies." Responses were measured on a 6-point Likert scale with 0 indicating "not applicable" and 5 "strongly agree".

Self-Regulated Online Learning Questionnaire (3 items; SOL-Q; Jansen et al., 2017)

The SOL-Q measures self-regulated learning for fully online courses with a focus on individual learning strategies. In the present study, we selected the following three items from the SOL-Q: "I find it hard to stick to a study schedule for this online course," "I choose the location where I study for this online course to avoid too much distraction," and "I know what the instructor expects me to learn in this online course." Answers were given along a 7-point scale ranging from 1 ("not at all true for me") to 7 ("very true for me").

Achievement Goal Questionnaire (3 items; AGQ; Elliot & McGregor, 2001)

We adopted the AGQ to assess four types of achievement goals in participants' online courses: mastery approach, performance approach, mastery avoidance, and performance avoidance goals from the traditional 2 × 2 achievement goal framework. The mastery/performance dimension was defined as refers to the learners themselves or to peers, respectively. That is, if students compared their goal to their previous achievement, it was considered a mastery goal. On the contrary, if their goal was to compare with peers, it was considered a performance goal. The approach/avoidance dimension was defined by learners' attitude towards to the goal. Thus, the approach goal was defined as learners pursuing mastery/performance goals aggressively, like performing better than peers. Conversely, if learners treated their goal passively, like just wanting not to perform worse than peers, this goal was considered performance avoidance.

Each achievement goal was described by three items (total of 12 items), with responses given along a 7-point Likert scale (1 = "not at all true of me" to 7 = "extremely true of me"). A sample item for measuring a mastery-approach goal is, "I want to learn as much as possible from this online class"; for measuring a performance-approach goal, "It is important for me to do well compared to others in this online class"; for a mastery-avoidance goal, "I worry that I may not learn all that I possibly could in this online class;" and for a performance-avoidance goal, "My goal in this online class is to avoid performing poorly." In this study, only three items were selected, those are "I desire to completely master the material presented in this online class," "Sometimes I am afraid that I may not understand the content of this online class as thoroughly as I would like," and "I just want to avoid doing poorly compare with others in this online class."

Supportive Online Learning Behaviors (7 items; SOLB; Yeh et al., 2019)

Students' supportive online behaviors were measured by five items from the SOLB ("communicate effectively with faculty and classmates," "create a schedule," "have a dedicated study space," "know your resources," and "manage your time") with two response options ("Yes" and "No"). "Communicate effectively with faculty and classmates" meant that since in-person communication is sometimes not an option in online learning situations, students can make use of email, chats, forums, and other formats to communicate with fellow students and professors if they have questions and need clarification. For the item "create a schedule," because procrastination is the greatest enemy of online learners, students could make a to-do list of the tasks they need to complete to make sure that they stay organized and do not fall behind in their online class. The idea behind "have a dedicated study space" is that with an online course, all of the time is spent outside of the classroom; therefore, students should find a quiet place with a good internet connection, access to power, distraction-free, and available for use at any time to take their online course. For the "know your resources" item, students should ensure their computer is working well, install any needed software, and verify their browser is up-to-date. That is, it is important to ensure course related technologies work properly, so students can focus their attention on course materials and not be distracted by technology problems. Finally, for "manage your time," students need to schedule time (and enough of it) in their personal calendar to study the materials in their online course and complete assignments, just as they might attend a face-to-face lecture at a regular time each week, for example.

Analysis

Rstudio (RStudio Team, 2015) with R-3.6.1 (R Core Team, 2019) and the R package of "randomForest" (Liaw & Wiener, 2002) were used to build the predictive model. To predict whether students would obtain an "A" for their final grade, two approaches were employed. The first adopted stepwise followed by logistic regression. First, items were selected by backward stepwise regression. Then only the selected items were used for fitting a logistic regression to estimate the probability of receiving an "A" at the end of the semester. The second approach used the RF method (Breiman, 2001). As mentioned, RF grows a large number of decision trees with randomly selected variables in each node. In the present study, we set out to grow 500 trees and randomly selected two thirds of the total variables (32 variables). After pruning, we compared the receiver operating characteristics (ROC) and the prediction accuracy of the two approaches. ROC is a probability curve that is based on the model's true positive and false positive rate. Based on the area under the curve, we gain insight into which model is better. Prediction accuracy was measured by the percentage of correct predictions in all responses.

Results

Selected Items

In the stepwise regression with backward selection, we selected 13 out of 32 items (see Table 1). The Akaike Information Criterion (AIC) of the model containing these items was 125.75 lower than the original model with all the items. The selected items were as follows: two items related to the reasons for taking the course, one item on social media notification, two items on the test of online learning success, two items on self-regulated online learning, one item on the mastery approach goal, and five items on learning strategies. The final logistic regression model with these 13 items was:

$$l\left(\frac{P(get\ A)}{1-P(get\ A)}\right) = -380.79 - 451.78 * Q3 - 694.77 * Q6 - 14.91 * Q11 + 202.89 * Q17 - 43.45 * Q18 + 131.63 * Q21 + 59.51 * Q22 - 162.61 * Q23 + 537.66 * Q26 + 1131.07 * Q29 - 198.36 * Q30 - 478.3 * Q31 + 426.43 * Q32$$

Using the same training data set, RF gave us a decision tree (see Figure 1) that was built by four items (see Table 1). The first node was from an item of OL_MARS. The second and third nodes were from two items of SOL_Q, and the fourth node was from an item of TOOLS. Among these four items, three (Q21, Q22, and Q17) were also selected by the stepwise followed by logistic regression approach.

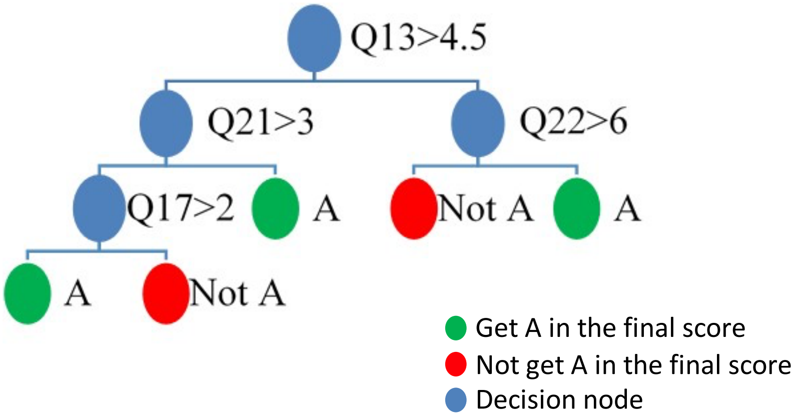
Table 1

Selection Results from Two Approaches

Item	**Construct**	**Description**	**Logistic**	**RF**
Q3	Reason	This course will be useful to me in other courses.	✓	
Q6	Reason	This course is required for major (program).	✓	
Q11	OL-MARS	I turn on the computer in order to do my homework, but I still visit Facebook first.	✓	
Q13	OL-MARS	When using the computer for studying, I think of what I want to eat later or what I have just eaten.		✓
Q17	TOOLS	I need faculty to remind me of assignment due dates.	✓	✓
Q18	TOOLS	I am capable of solving problems alone.	✓	
Q21	SOL-Q	I choose the location where I study for this online course to avoid too much distraction.	✓	✓
Q22	SOL-Q	I know what the instructor expects me to learn in this online course.	✓	✓
Q23	AGQ	I desire to completely master the material presented in this online class.	✓	
Q26	SOLB	Communicate effectively with faculty and classmates	✓	
Q29	SOLB	Stay organized.	✓	
Q30	SOLB	Have a dedicated study space.	✓	
Q31	SOLB	Know your resources.	✓	
Q32	SOLB	Manage your time.	✓	

Note. Reason = Reasons for taking the online course; OL-MARS = Revised Online-Learning Motivated Attention and Regulation Scale; TOOLS = Test of Online Learning Success; SOL-Q = Self-Regulated Online Learning Questionnaire; AGQ = Achievement Goal Questionnaire; SOLB = Supportive Online Learning Behaviors; Logistic = Logistic regression approach; RF = Random forest approach.

Figure 1. RF results.



Comparison of the Two Approaches

Compared with stepwise followed by logistic regression, RF produced a larger area under the curve (AUC) in both sets of test data (see Figures 2 and 3). RF also outperformed the stepwise followed by logistic regression in terms of prediction accuracy. Specifically, in T1, the correct prediction rate of logistic regression was only 82.81% (53 correct out of 64), whereas RF had 61 correct predictions out of 64, a 95.31% correct prediction rate. Similarly, in T2, logistic regression had only an 80.43% correct prediction rate (37 correct out of 46) whereas RF had an 89.13% (41 correct out of 46) prediction rate.

Figure 2. ROC curves in T1.

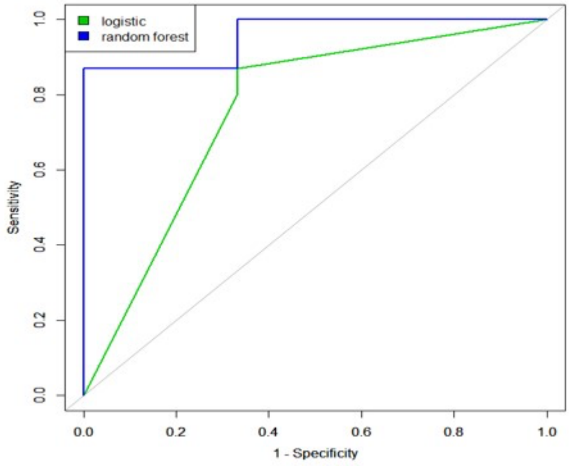
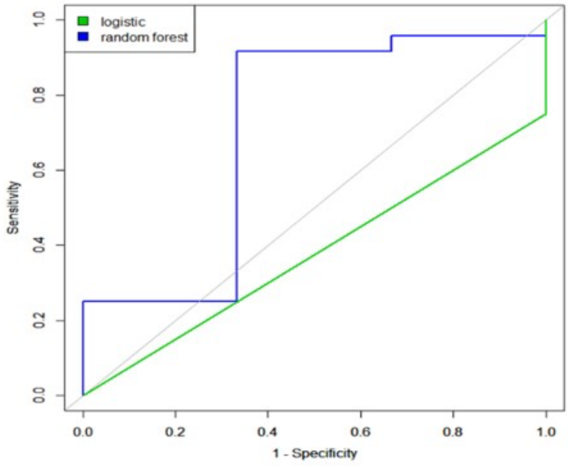


Figure 3. ROC curves in T2.



Discussion

Overall, RF selected fewer and more accurate items than the stepwise followed by logistic regression approach. That is, RF only selected 4 items for the final predictive model, compared with 13 selected by the stepwise followed by logistic regression. Further, RF was approximately 10% more accurate than the stepwise followed by logistic regression approach.

Three items were selected by both approaches. The first item was "I choose the location where I study for this online course to avoid too much distraction." Since online learners are not "stuck" in a traditional classroom setting, structuring their physical learning environment (e.g., finding and setting up a comfortable and regular place to study that fits their unique learning style) is crucial to reduce disturbances during the learning process (Barnard et al., 2008; Du, 2016). In the same vein, DeCandia (2019) mentioned that online students need to create structure and get more organized to manage technological distractions and ensure they stay on track to achieve their academic goals.

The second item was "I know what the instructor expects me to learn in this online course." Some research has shown that teacher expectations can significantly affect student achievement (Brophy, 1983; Cooper, 2000; Monhardt, 1995; Rubie-Davies, 2010). For example, Monhardt (1995) found that if students knew what they were expected to do and how they were expected to act, they behaved accordingly. Compared with traditional face-to-face settings, online environments provide learners with more control over their learning materials. Therefore, if online learners have a better understanding of teacher expectations from the outset, they may be able to choose a more effective order and pace along with a more individualized approach to fulfill course requirements and, therefore, be more likely to achieve a better learning outcome.

The third item was "I need faculty to remind me of assignment due dates." This item is directly related to whether or not a student is an independent learner. Compared with traditional face-to-face instruction, online courses require students to plan for their own self-development and self-management, allowing them to develop a more responsive attitude towards performing learning tasks. Hence, if a student needs others to remind him/her of assignment due dates, for example, he or she is less likely to receive an "A" at the end of the semester (Gorbunovs et al., 2016; Lin & Hsieh, 2001; Rappel, 2017).

Besides the three items that were selected by both approaches, RF selected the following item, which was used as the first branch: "When using the computer for studying, I think of what I want to eat later or what I have just eaten." Those who responded high on this item were less likely to receive an "A" for their final grade. While this item at first glance seems surprising, indeed, a strong relation has been found between food-intake patterns and academic performance in previous studies (Kleinman et al., 1998; Murphy et al., 1998; No Kid Hungry's Center for Best Practices, 2019). The process of online learning can be boring at times, which can lead to emotional eating (Koball et al., 2012). Although having flexible study hours is viewed as an advantage of online learning, picking the "right time" to study is crucial for students to be successful in online learning environments. For example, selecting study time around mealtimes may lead to unintended distractions and, therefore, should be avoided, if possible.

Conclusions and Limitations

As illustrated, stepwise logistic regression retained more questions from the original questionnaires, helping us get a more complete picture of participants' responses. Thus, the retention of more items offered an opportunity (with more information) to explain the reasons behind learning outcome. On the other hand, random forest retained fewer items while having a better prediction outcome. It is easy to follow the decision tree to detect at-risk students, but sometimes it is hard to explain why. Comparing the ROCs of two sets of testing data, RF clearly outperformed stepwise logistic regression, a finding that is consistent with the findings by those of Lakkaraju et al. (2015). Therefore, we can conclude that RF is a better option for psychological variables and data with a small sample size. A side benefit of using psychological variables is that they can be measured before a student takes a given course and, in the current context, help us understand students' online learning readiness from a different perspective, which is hardly achieved by using the predictive model with log data.

RF selected four questions to detect at-risk students. The first was "When using the computer for studying, I think of what I want to eat later or what I have just eaten." This suggests students might be taking online courses around dinner time, which might cause distractions. Based on the regression tree results, it may still be okay for students to think about food during the course as long as they understand the instructor's expectations for the online course. However, students may achieve a poorer learning outcome if they think about food while taking the course, do not have a proper place dedicated to online learning, and do not remember the deadline for assignments. However, these items were selected from a limited question pool since the T3 has smaller question set. In future research, a larger question set might improve the reliability of training results. The sample size is another limitation of the RF results. Although the RF may be used with small sample sizes, a larger sample size would help this method become more reliable.

There are a few limitations in our current study. As in other survey studies, our participants voluntarily answered all the survey questions and revealed their final grade., Hence there may be some selection bias. For example, students might be more willing to report their final grade if they had received a good grade. Such potential bias might also be the cause of the low variability in the final grade variable. Another issue related to final grades is that since our participants were enrolled in different online courses, it is difficult to control the grades-confounding variables. Additionally, the majority of the participants were female due to the source of our data (mostly collected within the college of education, which enrolls more female students than male students). To improve the predictive model, a more diversified sample (e.g., with more male students from other colleges) will be needed in future studies.

Author Note

Hsiang-Yu Chien, Department of Educational Psychology, Texas A&M University

Oi-Man Kwok, Department of Educational Psychology, Texas A&M University

Yu-Chen Yeh, Department of Educational Psychology, Texas A&M University

Noelle Wall Sweany, Department of Educational Psychology, Texas A&M University

Eunkyeng Baek, Department of Educational Psychology, Texas A&M University

William McIntosh, Department of Sociology, Texas A&M University

This research was supported by a grant from the T3: Texas A&M Triads for Transformation project.

Correspondence concerning this article should be addressed to Oi-Man Kwok, Department of Educational Psychology, Texas A&M University, 4225 TAMU, College Station, TX 77843-4225. Contact: omkwok@tamu.edu

References

Allen, I. E., & Seaman, J. (2013). *Changing course: Ten years of tracking online education in the United States.* Babson Survey Research Group and Quahog Research Group, LLC. http://www.onlinelearningsurvey.com/reports/changingcourse.pdf

Asarta C. J., & Schmidt J. R. (2013). Access patterns of online materials in a blended course. *Decision Sciences Journal of Innovative Education, 11*(1), 107–123.

Barnard, L., Paton, V., & Lan, W. (2008). Online self-regulatory learning behaviors as a mediator in the relationship between online course perceptions with achievement. *The International Review of Research in Open and Distance Learning, 9*(2), 1–11.

Bonny, A. E., Britto, M. T., Klostermann, B. K., Hornung, R. W., & Slap, G. B. (2000). School disconnectedness: Identifying adolescents at risk. *Pediatrics*, *106*(5), 1017–1021.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Brophy, J. E. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75*, 631–661.

Cooper, H. M. (2000). Pygmalion grows up. In P. K. Smith & A. D. Pellegrini (Eds.), *Psychology of education: Major themes* (pp. 338–364). RoutledgeFalmer.

DeCandia, C. (2019). *Managing distractions as an online student*. Affordable Colleges Online. https://www.affordablecollegesonline.org/college-resource-center/managing-distractions-for-online-students/

Digital Learning Compass. (2017). *Digital Learning Compass: Distance education enrollment report 2017.* http://digitallearningcompass.org/

Du, J. (2016). Predictors for Chinese students’ management of study environment in online groupwork. *International Journal of Experimental Educational Psychology, 36*(9), 1614–1630.

Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*, 501–519.

Er, E. (2012). Identifying at-risk students using machine learning techniques: A case study with IS 100. *International Journal of Machine Learning and Computing, 2*(4), 476.

Fassnacht, F. E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., & Koch, B. (2014). Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment*, *154*, 102–114.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874.

Foresman, B. (2020, March 24). Here are the U.S. universities that have closed due to coronavirus. EdScoop. https://edscoop.com/universities-closed-due-coronavirus-2020/

Funk, J. T. (2005). At-risk online learners: Reducing barriers to success. *eLearn*, *2005*(8), 3.

Gorbunovs, A., Kapenieks, A., & Cakula, S. (2016). Self-discipline as a key indicator to improve learning outcomes in e-learning environment. *Procedia - Social and Behavioral Sciences, 231,* 256–262.

Hughes, G., & Lewis, L. (2003). Who are successful online learners? Exploring the different learner identities produced in virtual learning environments. In J. Cook & D. McConnell (Eds.), *Communities of practice. Research Proceedings of the 10th Association for Learning Technology Conference*. Association for Learning Technology.

Jansen, R. S., Van Leeuwen, A., Janssen, J., Kester, L., & Kalz, M. (2017). Validation of the self-regulated online learning questionnaire. *Journal of Computing in Higher Education, 29*(1), 6–27.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open-source analytics initiative. *Journal of Learning Analytics*, *1*(1), 6–47.

Kerr, M. S., Rynearson, K., & Kerr, M. C. (2006). Student characteristics for online learning success. *Internet and Higher Education, 9*, 91–105.

Keshtkar, F., Cowart, J., & Crutcher, A. (2018). *Predicting risk of failure in online learning platforms using machine learning algorithms for modeling students' academic performance*. http://medianetlab.ee.ucla.edu/papers/ICMLWS1. pdf

Kleinman, R. E., Murphy, J. M., Little, M., Pagano, M., Wehler, C. A., Regal, K., & Jellinek, M. S. (1998). Hunger in children in the United States: Potential behavioral and emotional correlates. *Pediatrics, 101*(1), 100–111.

Koball, A. M., Meers, M. R., Storfer-Isser, A., Domoff, S. E., & Musher-Eizenman, D. R. (2012). Eating when bored: Revision of the Emotional Eating Scale with a focus on boredom. *Health Psychology, 31*(4), 521.

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

Kwok, O. M., Yeh, Y. C., Chien, H. Y., Sweany, N. W., Baek, E., & McIntosh, W. (2019). Finding the at-risk online learners: Development of the Online Readiness Screener (ORES). In the *Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge (LAK19)* (pp. 159–160). ACM.

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015, August). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1909–1918). ACM.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News 2*(3), 18–22.

Lin, B., & Hsieh, C. T. (2001). Web-based teaching and learner control: A research review. *Computers & Education, 37*, 377–386.

Mahboob, T., Irfan, S., & Karamat, A. (2016). *A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms* [Conference presentation]. 19th International Multi-Topic Conference, Islamabad, Pakistan.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

Miller, N., Kennedy, H., & Leung, L. (2000). Tending to the tamagotchi: Rhetoric and reality in the use of new technologies for distance learning. In S. Wyatt, F. Henwood, N. Miller, & P. Senker (Eds.), *Technology and in/equality: Questioning the information society* (pp. 129–146). Routledge.

Monhardt, B. M. (1995). Safe by definition. *American School Board Journal, 182*(2), 32–34.

Moody, J. (2004). Distance education: Why are the attrition rates so high? *The Quarterly Review of Distance Education, 5*(3), 205–210.

Mundry, R., & Nunn, C. (2009). Stepwise model fitting and statistical inference: Turning noise into signal pollution. *The American Naturalist*, *173*(1), 119–123.

Murphy, J. M., Wehler, C. A., Pagano, M. E., Little, M., Kleinman, R. E., & Jellinek, M. S. (1998). The relationship between hunger and psychosocial functioning in low income American children. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*(2), 163–170.

No Kid Hungry's Center for Best Practices. (2019). *Learn how hunger affects your school.* http://bestpractices.nokidhungry.org/playbook/schools/learn-how-hunger#learn-how-hunger-affects-your-school.

Osborn, V. (2001). Identifying at-risk students in videoconferencing and web-based distance education. *American Journal of Distance Education, 15*(1), 41–54.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning.

Poulet, C., Veale, D., Arnol, N., Levy, P., Pepin, J. L., & Tyrrell, J. (2009). Psychological variables as predictors of adherence to treatment by continuous positive airway pressure. *Sleep Medicine*, *10*(9), 993–999.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

Rappel, L. (2017). Self-direction in on-line learning. *Journal of Educational Systems, 1*(1), 6–14.

RStudio Team. (2015). *RStudio: Integrated development for R.* RStudio, Inc. http://www.rstudio.com/.

Rubie-Davies, C. M. (2010). Teacher expectations and perceptions of student attributes: Is there a relationship? *British Journal of Educational Psychology, 80*, 121–135.

Selwyn, N., & Gorard, S. (2003) Reality bytes: Examining the rhetoric or widening educational participation via ICT. *British Journal of Educational Technology, 34*(2), 169–181.

Steyerberg, E., Eijkemans, M., & Habbema, D. (1999). Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology, 52*(10), 935–942.

Taormino, M. (2010). Student preparation for distance education. *Distance Learning, 7*(3), 55.

Tuckman B. W. (2005). Relations of academic procrastination, rationalizations, and performance in a web course with deadlines. *Psychological Reports, 96*, 1015–1021.

Wallace, R. (2003) Online learning in higher education: A review of research on interactions among teachers and students. *Education, Communication & Information, 3*(2), 241–280.

Whittingham, M., Stephens, P., Bradbury, R., & Freckleton, R. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*(5), 1182–1189.

Wu, J. Y. (2017). The indirect relationship of media multitasking self-efficacy on learning performance within the personal learning environment: Implications from the mechanism of perceived attention problems and self-regulation strategies. *Computers & Education, 106*, 56–72.

Yeh, Y. C., Kwok, O. M., Chien, H. Y., Sweany, N. W., Baek, E., & McIntosh, W. A. (2019). How college students' achievement goal orientations predict their expected online learning outcome: The mediation roles of self-regulated learning strategies and supportive online learning behaviors. *Online Learning, 23*(4), 23–41.