Kastamonu Education Journal, 2023, Vol. 31, No:1, 143-154
doi: 10.24106/kefdergi.1246458

| Research Article / Araştırma Makalesi |

An Explainable Machine Learning Approach to Predicting and Understanding Dropouts in MOOCs

Açıklanabilir Makine Öğrenmesi Yöntemi ile Kitlesel Açık Çevrimiçi Derslerde Başarısızlığı Tahmin Etme ve Anlama

Erkan Er[1]

Keywords

1. Dropout prediction
2. Explainable machine learning
3. Ai in education
4. Learning analytics
5. Moocs

Anahtar Kelimeler

1. Başarısızlık tahmini
2. Açıklanabilir makine öğrenme
3. Eğitimde yapay zekâ
4. Öğrenme analitikleri
5. Kitlesel açık çevrimiçi dersler

Received/Başvuru Tarihi
23.11.2022

Accepted / Kabul Tarihi
13.01.2023

Abstract

Purpose: The purpose of this study is to predict dropouts in two runs of the same MOOC using an explainable machine learning approach. With the explainable approach, we aim to enable the interpretation of the black-box predictive models from a pedagogical perspective and to produce actionable insights for related educational interventions. The similarity and the differences in feature importance between the predictive models were also examined.

Design/Methodology/Approach: This is a quantitative study performed on a large public dataset containing activity logs in a MOOC. In total, 21 features were generated and standardized before the analysis. Multi-layer perceptron neural network was used as the black-box machine learning algorithm to build the predictive models. The model performances were evaluated using the accuracy and AUC metrics. SHAP was used to obtain explainable results about the effects of different features on students' success or failure.

Findings: According to the results, the predictive models were quite accurate, showing the capacity of the features generated in capturing student engagement. With the SHAP approach, reasons for dropouts for the whole class, as well as for specific students were identified. While mostly disengagement in assignments and course wares caused dropouts in both course runs, interaction with video (the main teaching component) showed a limited predictive power. In total six features were common strong predictors in both runs, and the remaining four features belonged to only one run. Moreover, using waterfall plots, the reasons for predictions pertaining to two randomly chosen students were explored. The results showed that dropouts might be explained by different predictions for each student, and the variables associated with dropouts might be different than the predictions conducted for the whole course.

Highlights: This study illustrated the use of an explainable machine learning approach called SHAP to interpret the underlying reasons for dropout predictions. Such explainable approaches offer a promising direction for creating timely class-wide interventions as well as for providing personalized support for tailored to specific students. Moreover, this study provides strong evidence that transferring predictive models between different contexts is less like to be successful.

Öz

Çalışmanın amacı: Bu çalışmanın amacı, açıklanabilir bir makine öğrenmesi yaklaşımı kullanarak aynı kitlesel açık çevrimiçi dersin (KAÇD) iki farklı öğretiminde öğrencilerin başarısızlık nedenleri tahmin edilmiştir. Açıklanabilir yaklaşımla, kara kutu tahmin modellerinin pedagojik bir bakış açısıyla yorumlanmasını ve ilgili eğitim müdahaleleri için eyleme dönüştürülebilir içgörüler üretilmesini amaçlanmıştır. Öngörü modelleri arasındaki özellik önemindeki benzerlik ve farklılıklar da incelenmiştir.

Materyal ve Yöntem: Bu araştırma, bir KAÇD'indeki öğrencilerin etkinlik günlüklerini içeren büyük bir genel veri kümesi üzerinde gerçekleştirilen nicel bir çalışmadır. Analizden önce toplam 21 özellik oluşturulmuş ve standardize edilmiştir. Tahmin modelleri oluşturmak için kara kutu makine öğrenme algoritması olarak çok katmanlı algılayıcı sinir ağı kullanılmıştır. Model performansları, doğruluk ve AUC metrikleri kullanılarak değerlendirilmiştir. SHAP, farklı özelliklerin öğrencilerin başarı veya başarısızlıkları üzerindeki etkileri hakkında açıklanabilir sonuçlar elde etmek amacıyla kullanılmıştır.

Bulgular: Sonuçlara göre, tahmine dayalı modeller oldukça yüksek bir performans göstermiştir. Bu da oluşturulan özelliklerin öğrenci katılımını yüksek bir doğruluk payıyla ölçtüğünü göstermektedir. SHAP yaklaşımıyla, tüm sınıfın yanı sıra belirli öğrenciler için dersi neden bıraktıkları otomatik olarak belirlenebilmiştir. Çoğunlukla ödevlere ve ders materyallerine olan ilgisizlik her iki derste de başarısızlığa neden olurken, ana içeriklerin sunulduğu video ile etkileşim sınırlı bir tahmin gücü göstermiştir. Toplamda altı özellik, dersin her iki öğretiminde de belirleyici olmuştur ve geri kalan dört özellik, yalnızca tek öğretimde öne çıkmıştır. Ayrıca, şelale grafikleri kullanılarak, rastgele seçilen iki öğrenciye ilişkin tahminlerin nedenleri araştırılmıştır. Sonuçlar öğrencilerin farklı nedenlerle başarısız olabileceğini göstermiş ve bu nedenlerin tüm sınıf için yapılan analizlere göre farklı olabileceğini ortaya koymuştur.

Önemli Vurgular: Bu çalışma, dersi bırakma tahminlerinin altında yatan nedenleri yorumlamak için SHAP adlı açıklanabilir bir makine öğrenmesi yaklaşımının kullanımını göstermektedir. Bu tür açıklanabilir yaklaşımlar, zamanında sınıf çapında müdahaleler oluşturmak ve ayrıca belirli öğrencilere göre uyarlanmış kişiselleştirilmiş destek sağlamak için önemli bir potansiyele sahiptir. sunar. Ayrıca bu çalışma, tahmine dayalı modellerin farklı bağlamlar arasında aktarılmasının pek mümkün olmadığına dair güçlü kanıtlar sunmaktadır.

[1] **Corresponding Author**, Computer Education and Instructional Technology, Faculty of Education, Middle East Technical University, Ankara, TURKEY. https://orcid.org/ 0000-0002-9624-4055

Citation/Alıntı: Er, E. (2023). An Explainable Machine Learning Approach to Predicting and Understanding Dropouts in MOOCs, *Kastamonu Education Journal, 31*(1), 153-164. doi: 10.24106/kefdergi.1246458

INTRODUCTION

The automatic prediction of at-risk students has been one of the prominent research topics in the learning analytics field (Tseng et al., 2016). Although they existed long before (Roblyer & Marshall, 2002), studies on the use of machine learning to predict at-risk learners have exponentially grown after the emergence of Massive Open Online Courses (MOOCs) around 2011 (Greene et al., 2015; Tseng et al., 2016; Bote-Lorenzo & Gómez-Sánchez, 2018). In MOOC contexts, at-risk learners are usually named as dropouts (Halawa et al., 2014). MOOCs are unique in that thousands of individuals from around the world with different background and motivation participate them online at no cost, which has not been feasible in formal learning settings of higher education. The interactions of MOOC participants during their learning journey create unprecedented amounts of trace data (or digital footprints). The trace data holds huge potential to derive valid engagement metrics and indicators for the prediction of dropouts. These datasets have been very valuable as the accuracy in dropout prediction requires a good quality of indicators (aka, features in machine learning) about a high number of students. Thus, MOOCS have provided, in the last 10 years, a very convenient and competent context for advancing the research on the automatic identification of dropouts (or at-risk learners).

The literature has provided sufficient evidence that with students' online interaction data, machine learning algorithms can accurately predict students who may dropout (Gardner & Brooks, 2018). The primary intervention from dropout predictions has been the early warning systems (Arnold & Pistilli, 2012; Akcapinar et al., 2019). These systems usually contain a dashboard-like interface to somehow visualize students based on their probability to fail (Arnold & Pistilli, 2012). Although the early warning systems provide a viable approach to increasing instructors' awareness about at-risk learners, they are limited in providing explainable and interpretable insights due to the black-box nature of some powerful machine learning algorithms (such as, deep learning and neural networks) (Khosravi et al., 2022). The power of these predictive models in making accurate predictions comes from the advanced mathematical computation required by the behind-the-scenes algorithms. However, this needed complexity makes the interpretation of the results troublesome from the end-user perspective. For example, in the case of dropout prediction, instructors might be provided with a very trustable report about the students with dropout risk; however, such a report without the underlying reasons for the failure limits instructors' ability to intervene properly (Holmes et al., 2021).

Although explainable machine learning and artificial intelligence has been around for several years already (Samek & Müller, 2019), the need for explainable models has been very lately highlighted in the learning analytics literature (Jang et al., 2022; Khosravi et al., 2022). Although there are some initial efforts (Khosravi et al., 2022; Shabaninejad et al., 2022), much more research is needed to explore the potentials of different approaches to creating explainable models in learning analytics. To make a timely contribution to the literature, this study explores the use of an explainable machine learning approach in the context of MOOC dropout prediction. This approach is called SHAP (SHapley Additive exPlanations), which uses Shapley values from game theory (Shapley, 1953) to compute the feature importance in black-box models. The explainable models were applied in two different runs of the same MOOC to explore why students may dropout and also to test the transferability of the findings across contexts. We aim to identify the reasons for predictions provided by a black-box algorithm, and to provide new insights into the ongoing discussion on the transferability of machine learning models across contexts for dropout prediction (Gašević et al., 2016; Er et al., 2020). In this regard, this study addresses the following research questions:

- To what extend the SHAP approach can help explain the predictive power of engagement indicators in the black-box dropout prediction models?
- How do the results regarding the features' contributions to the dropout prediction converge or diverge across two runs of the same course?

BACKGROUND

Dropout prediction in MOOCs

There has been an abundance of research regarding the automatic identification of dropouts in MOOCs (He et al., 2015; Jiang et al., 2014; Crossley et al., 2016; Nagrecha et al., 2017; Bote-Lorenzo & Gómez-Sánchez, 2018; Er et al., 2019). The predictions are usually performed through machine learning models that classify students as at-risks or not based on the features (or indicators) generated from the course engagement data (Sinha et al., 2014). Several factors can affect the performance of machine learning models (i.e., their ability to accurately identify at-risk students). One factor is the quality of the data. A large number of students along with rich data about them is necessary for the effective training of a machine learning model. Poorly trained models are unlikely to produce accurate and reliable results about the target variable or outcome (e.g., dropout or not) (Ye & Biswas, 2014). Associated with this, another important factor is the quality of the predictors and their association with the outcome to be predicted. For example, features that capture relevant aspects of engagement can be more potent to predict if a student is likely to fail or succeed (Veeramachaneni et al., 2014).

The literature has been informative in demonstrating the predictive power of distinct engagement indicators (or features) computed from the clickstream logs. Clickstream logs contain all data about learners' interaction with the MOOC platform (e.g., visiting a page, playing or pausing a video, downloading a file, attempting a quiz, viewing the discussions) (Whitehill et al., 2017). These raw logs offer a rich space to explore for generating features that somehow represent some aspects of student engagement in MOOCs. While the commonly used features in the literature are mostly based on the number of actions performed by learners

such as the number of times a content is visited or a video is watched (Bote-Lorenzo & Gómez-Sánchez, 2018; Er et al., 2020), some other temporal features such as when a lecture was viewed or how early an assignment was submitted were also used as predictors (Ye & Biswas, 2014).

The machine learning algorithms can be categorised into open-box and black-box models. In open-box models, a machine learning algorithm (such as logistic regression) produces results about the predictive power of each feature (in the form of a coefficient); therefore, their results are more interpretable by humans. On the other hand, close-box models uses computationally expensive algorithms (such as, neural network, deep learning) and rather fed with a big sparse data. Although they are found to produce more accurate results (Waheed et al., 2020), the results are less interpretable as they do not generate any information about the predictive power of features. Yet, such information is key to the interpretability of these models, and thus, to their capacity to inform educational decision making. Intervening in an educational context would be more possible and effective if it can be correctly analyzed and interpreted why a student may fail.

Explainable machine learning

An explainable machine learning model differs from a traditional model in that a human can interpret and judge why predictions were made in a certain way (Carvalho et al., 2019). Many models known as black box are optimized for performance rather than interpretability, and it is a challenge to gain insights into how certain features influence the prediction outcome. Some methods have been proposed to extract useful information from these models to increase their interpretability, and these methods promise new opportunities and research directions. Education is one of these domains where explainable black-box models can offer much to unknown underlying reasons for student dropout, while producing very accurate results. A promising approach is to translate black-box machine learning predictions into a format that is interpretable and explainable by the instructors (and any other stakeholders involved) to inform decision-making processes in learning and teaching (Laet et al., 2020).

Some recent research focused on the importance of explainability in learning analytics (Hasib et al., 2022; Shabaninejad et al., 2022; Holmes et al., 2022). While some studies explored the approaches to increasing explainability with dashboards (Shabaninejad et al., 2022), the use of explainable machine learning approaches is still in its infancy in the learning analytics literature (Jang et al., 2022) and more research is needed to explore the potentials of different approaches and methods. Increasing the adoption of such approaches will help open the black box of algorithms for greater interpretability from a pedagogical perspective while still harnessing the computational power of complex algorithms for higher accuracy. In this regard, this study attempts to make a timely contribution to the literature by illustrating the use of SHAP approach for understanding the underlying reasons for dropouts in two runs of the same MOOC.

METHOD

In this section, the methods, data collection tools, data collection process and analysis used for the purposes of your article should be written together with the reasons why they were used. In this section, the methods, data collection tools, data collection process and analysis used for the purposes of your article should be written together with the reasons why they were used. In this section, the methods, data collection tools, data collection process and analysis used for the purposes of your article should be written together with the reasons why they were used.

Data and the feature set

In this study, research data was extracted from a big public dataset[2]. This dataset is composed of millions of logs about student activities in many MOOCs taught via XuetangX (a popular MOOC platform in China) (Feng et al., 2019). For this study, the MOOC titled Introduction to Mao Zedong Thought, was selected (course id is 10610224X) as it contained the largest enrolment. The course lasted four weeks, and each week a new chapter was introduced. Each chapter concluded with an assignment and students were encouraged to share ideas and questions in discussion forums. In the public repository, no other information was provided about the course. In this research, the data pertaining to its two runs were analyzed.

In these MOOCs, students were considered dropout if they did not complete all mandatory activities and stopped visiting the course before all modules were taught. In the test dataset of the first run of the MOOC, there were 351 dropouts from 1087 participants, and in the second run, there were 360 dropouts among 902 participants. The features generated from this data are provided in Table 1. Before the data analysis, all features were standardized.

SHAP: An explainable learning analytics technique

With complex algorithms and a fair amount of data, it is possible to build accurate predictive models. Predictions are more powerful if accompanied with affordances to explain the underlying reasons for a particular outcome (e.g., why a student is expected to drop out). SHAP (SHapley Additive exPlanations) is a technique that adopts a game-theoretic approach (Shapley, 1953) to explain the outcome of predictive machine learning models (Lundberg & Lee, 2017). To increase the model transparency, SHAP

[2] http://moocdata.cn/data/user-activity

computes Shapley values as a measure of the predictive power of the features in a model. In this study, SHAP is used to explain the black-box machine learning models built for predicting dropouts in two runs of the same MOOC.

Python module SHAP is used to provide both global and local interpretability[3]. Here, while global level indicates the predictive analysis involving all participants, local level focuses on specific students as distinct cases when explaining the particular reasons for a selected dropout. SHAP values at global level can illustrate how each predictor feature contributes to the final outcome (i.e., dropout or not), either positively or negatively. At local level, SHAP values are also produced for unique observations, which refer to the individual student records. In this way, while the importance of different engagement indicators can be interpreted at global level, the reasons why a specific student drops out or not can be explained at local level.

Model evaluation and feature selection

Predictive models were built using Multi-layer Perceptron, a neural network algorithm used for the classification tasks (Mitra & Pal, 1995). Particularly, the Scikit-Learn implementation of this algorithm (called MLPClassifer) was used in this study (Pedregosa et al., 2012). Multi-layer perceptron is composed of an input layer, a hidden layer, and an output layer and considered a black-box algorithm since it involves a very complex association between the weights and the function being approximated.

The performance of the predictive models was evaluated using the accuracy and area under the curve (AUC) metrics. These metrics have been the widely used in the literature (Kennedy et al., 2015; Waheed et al., 2020).

Table 1. Features generated from the course log data

Feature category	Feature name	Feature description
Sessions and access to course pages	click_about	Number of times students visited the about page
	click_info	Number of times students visited the info page
	click_progress	Number of times students checked their overall progress
	click_courseware	Number of times students accessed to the system (LMS)
	close_courseware	Number of times students left the system (LMS)
	unique_session_count	Total count of unique sessions per student
	avg_nActions_per_session	Average number of actions per session
Students' engagement in assignments	problem_get	Number of times students open a problem
	problem_check*	Number of times students clicked to check the correctness of their answer
	problem_save	Number of times students saved their current progress on a problem
	problem_check_correct	Number of times the answer was correct for the problem checked
	problem_check_incorrect	Number of times the answer was incorrect for the problem checked
Students' interactions with course vides	load_video	Number of times students loaded a video
	play_video	Number of times students (re)played button in a video
	seek_video	Number of times students sought through a video
	pause_video	Number of times students paused a video
	stop_video	Number of times students stopped a video
Students' discussion form activities	click_forum	Number of times students viewed a discussion forum
	create_thread	Number of times students created a discussion thread
	create_comment	Number of times students posted a discussion comment
	delete_comment	Number of times students deleted a comment

FINDINGS

Model Performances

The neural network models were built and tested for two runs of the MOOC, named as Course #1 and Course #2. According to the results presented in Table 2, the performance of the models was fair, suggesting that the predictive features were considerably associated with the outcome variable. In Course #1, the model performed slightly better.

Table 2. Performance of the prediction models.

Course	Accuracy	AUC
#1	0.80	0.85
#2	0.74	0.81

[3] https://shap.readthedocs.io/en/latest/

According to the models' performances, the features (i.e., engagement indicators) included in the model could effectively estimate student success or failure. To analyse and interpret how these features contributed to the accurate predictions, SHAP values were analysed at global (considering all predictions) and local levels (considering a prediction for a selected student).

Global Interpretability of the Models

At global level, SHAP values were used to identify the important features and explain how they contribute to the final outcome. The results are presented in two sub-sections as follows.

Overall feature importance

The feature importance for Course #1 and Course #2 is visualised in the box plots shown in Figure 1 and Figure 2, respectively. In these plots, the most important 10 features are displayed in a descending order based on the magnitude of their contributions to the prediction accuracy, as determined by the average of the SHAP values (in x-axis).

In Course #1, unique_session_count and click_courseware (which represent the general course engagement) and problem_get and problem_check_correct (which are associated with assignment engagement), have been the features that contributed the most to the prediction. The rest of the features had relatively less importance, which includes two video-related indicators, load_video and pause_video, as well as some other general course engagement (such as click_about, close_courseware). Thus, the number of times the students opened a session in the MOOC platform and their assignment-related activities had a higher predictive capacity than the students' engagement with videos (which are the main teaching content in MOOCs).

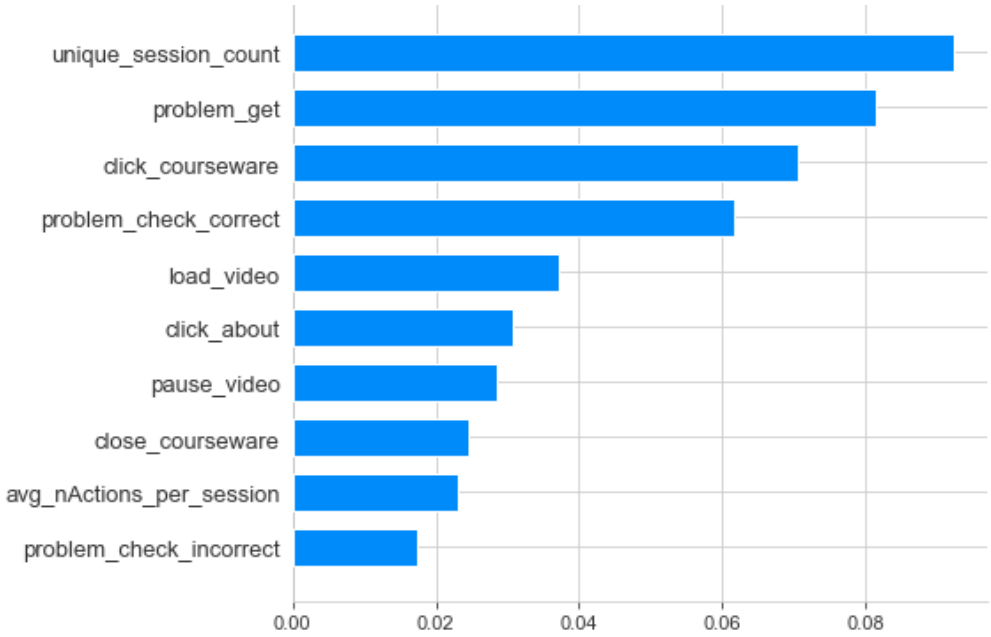

Figure 1. Feature importance in Course #1

The predictive power of the features in Course #2 differs from those in Course #1 as shown in Figure 2, while some similarities are also noted. To begin with, in Course #2, click_courseware played a significant role in predicting students' course outcomes, with a large margin against the rest of the features. Following click_courseware, five other features were influential, containing two features about the assignments (problem_check_correct, and problem_check), one feature about the video engagement (pause_video), and other two features about the general course engagement (click_about, and unique_session_count).

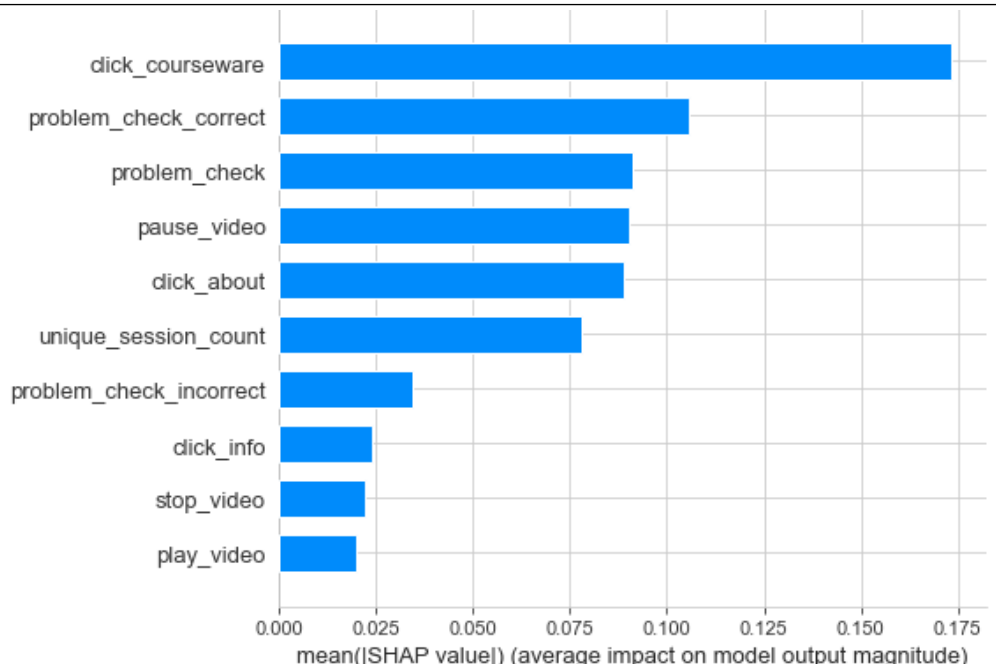

Figure 2. Feature importance in Course #2

Six features were common in both courses, which include unique_session_count, click_courseware, problem_check_correct, click_about, pause_video, and problem_check_incorrect. The remaining four features were found to be strong predictors in only one of the courses. Moreover, in neither of the courses, features about students' discussion form activities were found to be predictive.

The relationships of the features with the predicted outcome

Although the box plots above give information about the importance of all features, they are limited in illustrating their specific associations with the target outcome. For example, solely based on the feature importance, it cannot be known if a greater value in click_courseware is associated with a dropout or success, or vice versa.

SHAP values can be visualized through a summary plot to communicate the positive or negative relationship of the engagement indicators (i.e., predictors) with students' final course outcome (i.e., target variable). The summary plots for Course #1 and Course #2 are depicted in Figures 3 and 4, respectively. In these summary plots, every dot refers to a single student record in the data set, and the colour of a dot differentiates between the succeeded (red) or failed (blue) students. Moreover, in the x-axis, the features are ranked in descending order based on their importance, and in the y-axis, the horizontal location shows the magnitude of the effect of a feature on each student in a positive or negative direction. In the dataset, students who dropped out are labelled as 1 and the others are labelled as 0. For this reason, a positive SHAP value indicates a feature's correlation with dropout, whereas a negative SHAP value is associated with success.

According to Figure 3, in Course #1, all features except click_about hold a negative relationship with the predicted outcome, dropout. That is, when these variables have higher values (denoted by the red dots), students are less likely to drop out (as the blue dots placed on the left hand-side where the SHAP values are negative); whereas as their values increase (denoted by the red dots), the final outcome tends to be a dropout (since the red dots are placed on the right hand-side where the SHAP values are positive). The only feature with a negative relationship with student success is click_about. In other words, students with higher click_about value tended to be those who dropped out the course. In overall in Course #1, higher student engagement, as indicated by almost all important features, contributes to student success.

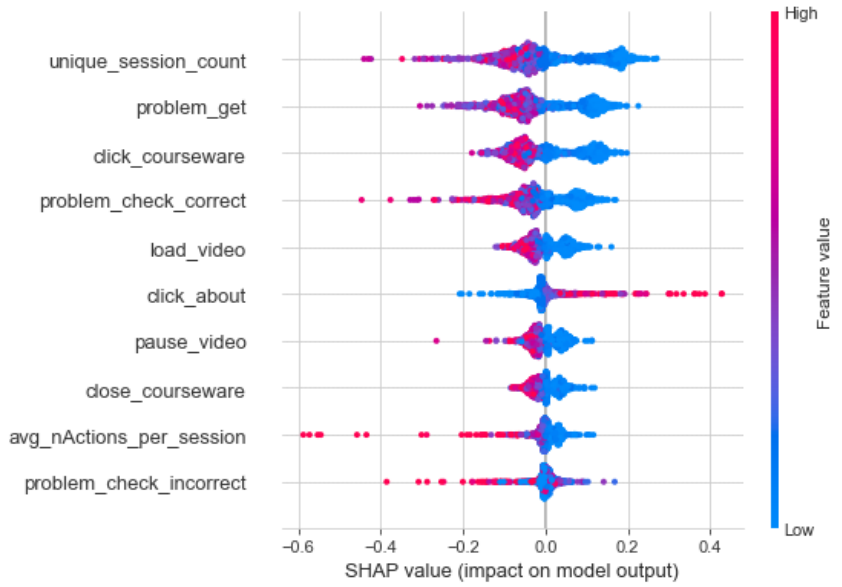

Figure 3. Relationships of the predictors with the target variable in Course #1

As shown in the summary plot in Figure 4, similar results were obtained in Course #2 in terms of the negative relationship between many features and student dropout. In other words, as in Course #1, student disengagement was considerably associated with failure. However, in Course #2, the positive indicators of dropout were different. In particular, high values of problem_check, followed by click_info and play_video, contributed to the prediction of student dropout. None of these features were found to be significant predictors in Course #1. problem_check was the only assignment-related feature with negative relationship with student success. However, its constituting variables (problem_check is the sum of problem_check_correct and problem_check_incorrect) demonstrated a positive relationship with student success. This finding shows that decomposing some indicators can help capture student engagement at deeper level and correctly identify the effects of specific student behavior.

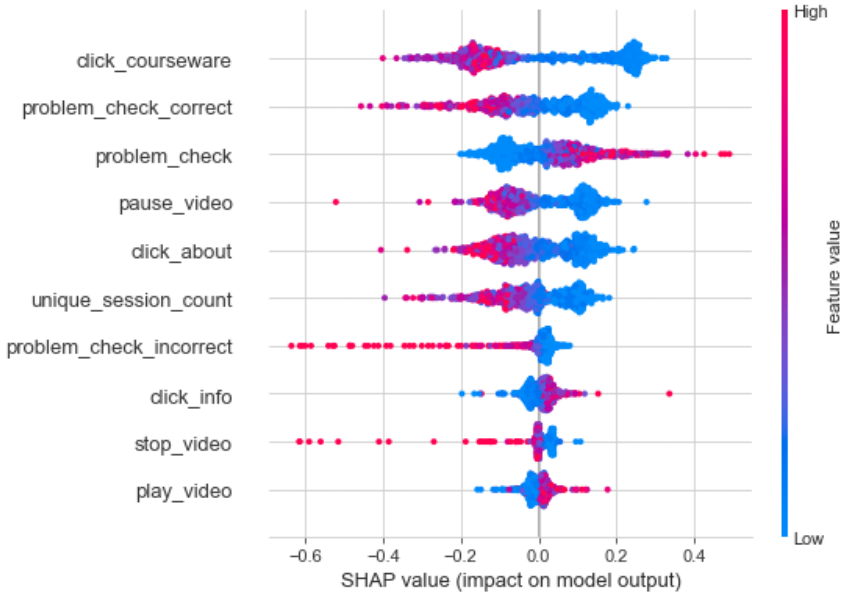

Figure 4. Relationships of the predictors with the target variable in Course #2

Local interpretability of the models

SHAP values can be visualized in several ways to produce explanations for individual predictions. In this way, the possible reasons why a specific student drops out or succeeds can be identified and interpreted. The explanations for specific cases can be different than the prediction trends at the global level.

Waterfall plots are designed to display explanations for an individual prediction based on the SHAP values of all the input features corresponding to that specific prediction. In other words, in this study, a waterfall plot can show why a specific student succeeded or failed in the course given the SHAP values of the features for that student. A waterfall plot starts with a base value at the bottom (which is the average of all observations) and moves toward right with features that pushes the prediction higher (dropout, as shown in red color) or left with the features that pushes the prediction lower (success, as shown in blue color).

In this study, to illustrate how waterfall plots can be effective in explaining reasons for individual predictions, one student was randomly chosen from Course #1 and Course #2, as shown in Figure 5 and Figure 6, respectively. Regarding the first case, most features contributed positively to the success of the student except for two features, namely problem_check_correct and problem_get, which increased students' dropout possibility. In the second case, which is depicted in Figure 6, the high values of problem_check_correct and unique_session_count contributed the most strongly to the student's success, whereas the high value of problem_check was an important indicator of failure for the same student. These results show that local interpretability enables us to focus each student as a different case and understand the possible factors behind a specific student's success or failure.

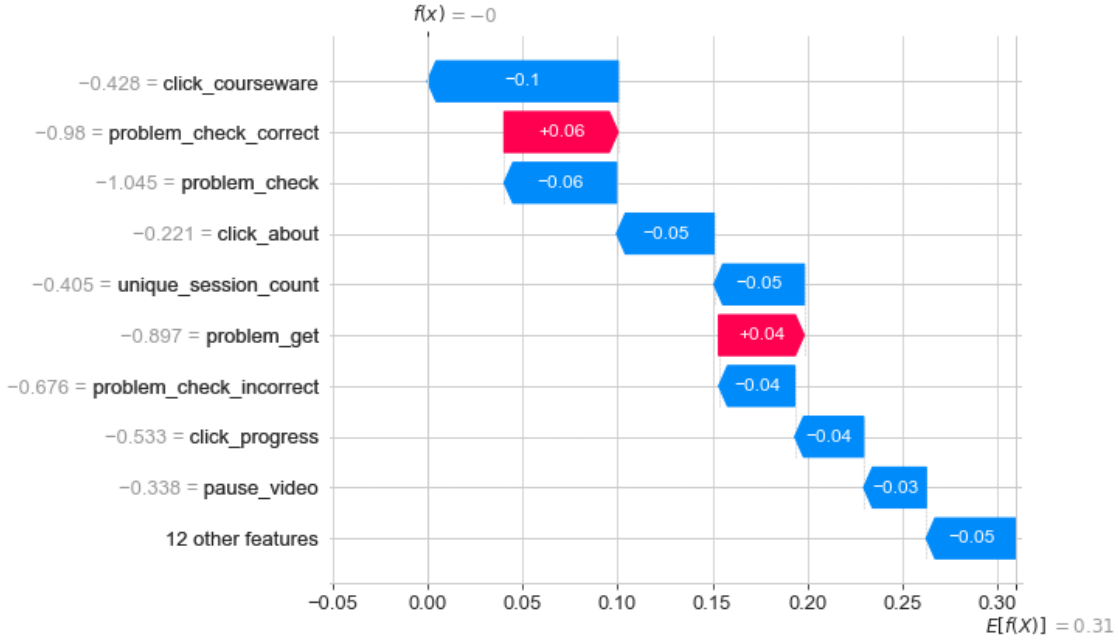

Figure 5. Waterfall plot about a specific student in Course #1

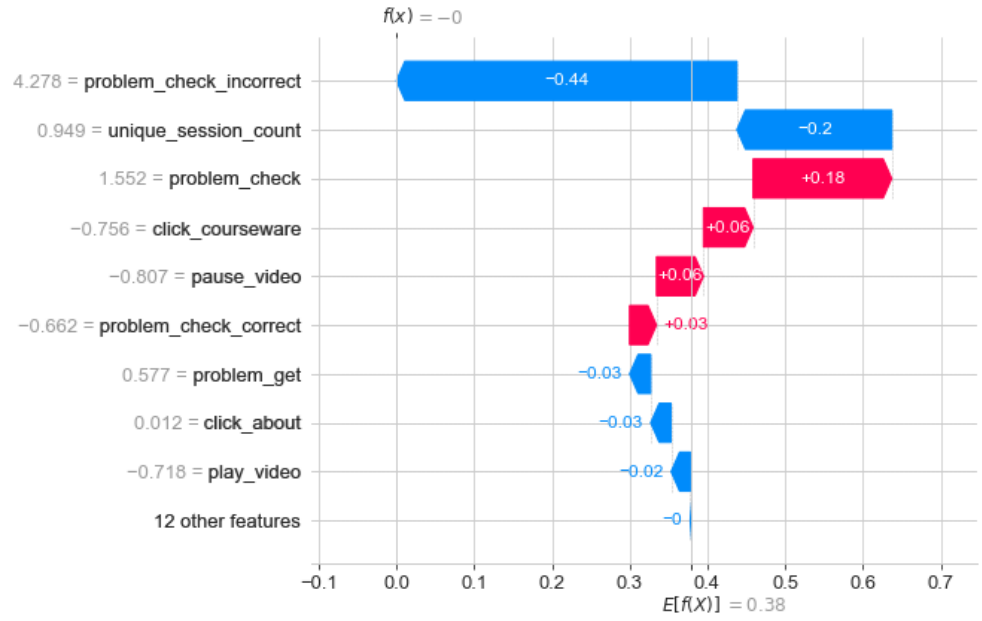

Figure 6. Waterfall plot about a specific student in Course #2

DISCUSSION

The results of this study demonstrate the importance of the use of explainable machine learning approaches in making rich sense of prediction results and provide evidence about the degree to which the same feature set converges or diverges in the two runs of the same MOOC. The discussion of the results and the main findings are presented per each research question as follows.

RQ1. To what extend the SHAP approach can help explain the predictive power of engagement indicators in the black-box dropout prediction models?

The results of this study show that SHAP approach can be effective in opening the black-box prediction models on student dropout. This approach first enabled the analysis of features at global level, averaging the predictions across all students. This analysis reveals the important engagement indicators associated with success, which can provide relevant information for designing interventions to prevent dropout. For example, in the first run of the course, students could be guided to access the MOOC platform regularly (unique_session_count), study the course wares more often (click_courseware), and work on the problems in the system (problem_get). Similarly, in the second run of the course, accessing the course wares was a very strong predictor of student success: students mostly dropped out when the number of times they accessed the course wares was very low. In this case, the instructor could send some reminders to help students regulate their use of the educational materials.

Moreover, in both runs, the frequency of checking answers for the problems (problem_check) solved was negatively associated with success whereas when these frequencies were computed separately for correctly (problem_check_correct) and incorrectly (problem_check_incorrect) solved problems they contributed positively to student success (the contribution of with the largest enrolment was higher). This result may indicate that when students who dropped out tended to quickly try different answers and check its correctness until they find the correct answer, which results in high problem_check value and inconsistencies in problem_check_incorrect and problem_check_incorrect values. On the other hand, other students seemed to reach to the correct answers in a smaller number of trials (i.e., lower problem_check) without inflating the problem_check_incorrect value. An important implication of this finding is that the MOOC platform should limit the number of trials on a problem and make relevant recommendations to the students about the learning materials that they should study (Fazeli et al., 2017). Also, possibly for the next runs of the course, the course instructor could introduce solved example problems as a supplementary guided material to solidify students' learning in difficult topics.

Furthermore, the SHAP approach provided rich intuitions about the reasons why a specific student may fail or succeed, which is not possible with other open-box algorithms such as logistic regression. Although the overall prediction results and the feature interpretations can inform the design class-wide interventions, students' can be provided with more individualized interventions if the particular reasons for their failure could be known. This requires a local-level analysis of the predictions, which is afforded by the SHAP approach. The result of this study revealed that features' SHAP values could change between different students, and they may offer student-specific interpretations of the predictions. Such interpretations could allow instructors to identify why specific students drops out and provide rather personalized support tailored to specific learner needs. That is, explainable machine learning approaches can promote the implementation of personalized education.

RQ2. How do the results regarding the features' contributions to the dropout prediction converge or diverge across two runs of the same course?

Regarding the importance of the features across the two runs of the same course, although a considerably high overlap was noted in the list of the most predictive features, the individual predictive power of each feature varied. For example, in both runs of the MOOC, students' interactions with the problems (e.g., problem_check_correct) and course wares (e.g., click_courseware) played a determining role in their success; however, accessing the course wares was the strongest predictor in the second run while playing a moderate role in the first run of the course. Similarly, the number of sessions in the MOOC platform (unique_session_count), a moderate predictor in the second run, was the feature that contributed the most to the predictions in the first run of the MOOC.

In overall in both courses, indicators about engagement in problems were more powerful predictors of dropout than the indicators regarding students' video activities. This finding indicates that in MOOCs, compared to interacting with lecture videos, student engagement in active learning processes such as working on problems or assignments can be more effective in promoting achievement. Among the video-related features, pause_video was more prominent, suggesting that pausing a video was a good indicator of students' cognitive engagement in the video content, as noted in previous studies (Yoon et al., 2021).

The limited feature overlap between the different runs of the same course highlights the dynamic and complex nature of learning that is affected by many uncontrolled factors. Although the course content might remain untouched, a new learning cohort with different motivations, attitudes, and skills is likely to create a different learning atmosphere, unique peer and teach interactions. All these factors can influence how students engage in a course, which in turn determines their final performance. Therefore, the findings of this study underlines the fact that transferring predictive models between courses (i.e., training a machine learning model in one course and using the trained model in a different course) is quite difficult even for different runs of the same course, and therefore are unlikely to produce very accurate results when different courses are involved. This finding is consistent with some previous research (Gašević et al., 2016; Er et al., 2020). Nevertheless, building predictive models on a big data set involving multiple courses can be useful in determining predictors that can generalize to different courses (Feng et al., 2019).

CONCLUSION AND RECOMMENDATIONS

As we collect more educational data with an increasing depth (rich data about a single learner) and breadth (data about more learners), advance machine learning algorithms comes handy in processing big educational data to obtain accurate results about students who may dropout. However, such algorithms tend to be black box as they do not tell much about the reasons for a specific prediction. This study illustrated the use of an explainable machine learning approach called SHAP to interpret the underlying reasons for dropout predictions at global and local levels. Such explainable approaches offer a promising direction for creating timely class-wide interventions as well as personalised support for specific students.

The major limitation of this study is the lack of detailed description of the context. The data was obtained from a public repository and no detailed information about the MOOC was provided (such as the connection between chapters, the number of videos in each chapter, external resources). A more comprehensive understanding of the context could help better explain why certain features contributed more or less in different runs of the course, and therefore could strengthen the discussion of the findings. A future study can replicate this research in a well-known context to obtain further insights about the effectiveness of explainable machine learning approaches. Moreover, a more comprehensive future research can make a comparison between a black-box and open-box algorithms to investigate the agreement between the explainable approach and the results obtained from an open-box algorithm such as logistic regression or decision tree.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, author-ship, and/or publication of this article.

Statements of publication ethics

I hereby declare that the study has not unethical issues and that research and publication ethics have been observed carefully.

Researchers' contribution rate

The study was conducted and reported by the corresponding author, who is the only author of this article.

Ethics Committee Approval Information

Since this study was conducted on a publicly available big data accessible via this url: http://moocdata.cn/data/user-activity, ethics committee approval was not required. Besides this available dataset, no additional data was collected.

REFERENCES

Akcapinar, G., Altun, A., & Askar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, *16*(40).

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: using learning analytics to increase student success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 267–270.

Bote-Lorenzo, M. L., & Gómez-Sánchez, E. (2018). An approach to build in situ models for the prediction of the decrease of academic engagement indicators in massive open online courses. *Journal of Universal Computer Science*, *24*(8), 1052–1071.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics (Switzerland)*, *8*(8), 1–34. https://doi.org/10.3390/electronics8080832

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge- LAK '16*, 6–14. https://doi.org/10.1145/2883851.2883931

Er, Erkan., Gómez-Sánchez, E., Bote-Lorenzo, M. L., Dimitriadis, Y., & Asensio-Pérez, J. I. (2020). Generating actionable predictions regarding MOOC learners' engagement in peer reviews. *Behaviour and Information Technology*, *39*(12). https://doi.org/10.1080/0144929X.2019.1669222

Er, Erkan, Gómez-Sánchez, E., Bote-Lorenzo, M.L. Dimitriadis, Y., & Asensio-Pérez, J. I. (2019). Generating actionable predictions regarding MOOC learners' engagement in peer reviews. *Behaviour and Information Technologies*, *39*(12), 1107–1121.

Fazeli, S., Drachsler, H., & Sloep, P. (2017). Applying Recommender Systems for Learning Analytics: A Tutorial. *Handbook of Learning Analytics*, 235–240. https://doi.org/10.18608/hla17.020

Feng, W., Tang, J., & Liu, T. X. (2019). Understanding dropouts in MOOCs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 517–524.

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, *28*(2), 127–203. https://doi.org/10.1007/s11257-018-9203-z

Gašević, D., Dawson, S., Rogers, T., & Gašević, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet and Higher Education*, *28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Greene, J. a., Oswald, C. a., & Pomerantz, J. (2015). Predictors of Retention and Achievement in a Massive Open Online Course. *American Educational Research Journal*, *XX*(X), 1–31. https://doi.org/10.3102/0002831215584621

Halawa, S., Greene, D., & Mitchell, J. (2014). Dropout prediction in MOOCs using learner activity features. *Proceedings of the Second European MOOC Stakeholder Summit*, 58–65.

Hasib, K. M., Rahman, F., Hasnat, R., & Alam, M. G. R. (2022). A machine learning and explainable AI approach for predicting secondary school student performance. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference, CCWC 2022*, 399–405. https://doi.org/10.1109/CCWC54503.2022.9720806

He, J., Bailey, J., & Rubinstein, B. I. P. (2015). Identifying at-risk students in massive open online courses. *Proceedings of the Twenty Ninth AAAI Conference on Artificial Intelligence*, 1749–1755.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2021). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00239-1

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, *32*(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. In *Education and Information Technologies* (Issue Early Access). https://doi.org/10.1007/s10639-022-11120-6

Jiang, S., Williams, A. E., Schenke, K., Warschauer, M., & Dowd, D. O. (2014). Predicting MOOC performance with week 1 behavior. *Proceedings of the 7th International Conference on Educational Data Mining*, 273–275.

Kennedy, G., Coffrin, C., de Barba, P., & Corrin, L. (2015). Predicting success: How learners' prior knowledge, skills and activities predict MOOC performance. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 136–140. https://doi.org/10.1145/2723576.2723593

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, *3*(2022). https://doi.org/10.1016/j.caeai.2022.100074

Laet, T. De, Millecamp, M., Broos, T., Croon, R. De, Verbert, K., & Duorado, R. (2020). Explainable Learning Analytics: Challenges and opportunities. *Proceedings of Learning Analytics and Knowledge Conference (LAK20)*.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774.

Mitra, S., & Pal, S. K. (1995). Fuzzy multi-layer perceptron, inferencing and rule generation. *IEEE Transactions on Neural Networks*, *6*(1), 51–63.

Nagrecha, S., Dillon, J. Z., & Chawla, N. V. (2017). MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable. *26th International Conference on World Wide Web Companion*, 351–359. https://doi.org/10.1145/3041021.3054162

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). *Scikit-learn: Machine learning in Python*. *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

Roblyer, M. D., & Marshall, J. C. (2002). Predicting success of virtual high school students. *Journal of Research on Technology in Education*, *35*(2), 241–255.

Samek, W., & Müller, K.-R. (2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Towards explainable artificial intelligence* (pp. 5–22). Springer.

Shabaninejad, S., Khosravi, H., Abdi, S., Indulska, M., & Sadiq, S. (2022). Incorporating Explainable Learning Analytics to Assist Educators with Identifying Students in Need of Attention. In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22), June 1â•fi3, 2022, New York City, NY, USA* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3491140.3528292

Shapley, L. (1953). A Value for n-person Games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games. Annals of Mathematical Studies* (pp. 307–317). Princeton University Press. https://doi.org/10.1515/9781400881970-018

Sinha, T., Jermann, P., Li, N., & Dillenbourg, P. (2014). Your click decides your fate : Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3–14.

Tseng, S.-F., Tsao, Y.-W., Yu, L.-C., Chan, C.-L., & Lai, K. R. (2016). Who will pass? Analyzing learner behaviors in MOOCs. *Research & Practice in Technology Enhanced Learning*, *11*(8), 1–11. https://doi.org/10.1186/s41039-016-0033-5

Veeramachaneni, K., O'Reilly, U.-M., & Taylor, C. (2014). Towards feature engineering at scale for data from massive open online courses. *ArXiv*.

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*. https://doi.org/10.1016/j.chb.2019.106189

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). MOOC dropout prediction: How to measure accuracy? *Proceedings of the Fourth ACM Conference on Learning@Scale*, 161–164. https://doi.org/10.1145/3051457.3053974

Ye, C., & Biswas, G. (2014). Early prediction of student dropout and performance in MOOCs using higher granularity temporal information. *Journal of Learning Analytics*, *1*(3), 169–172.

Yoon, M., Lee, J., & Jo, I. H. (2021). Video learning analytics: Investigating behavioral patterns and learner clusters in video-based online learning. *Internet and Higher Education*, *50*(October 2020), 100806. https://doi.org/10.1016/j.iheduc.2021.100806