International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-021-00246-2

ARTICLE

Towards Designing Profitable Courses: Predicting Student Purchasing Behaviour in MOOCs

Mohammad Alshehri[1,2] · Ahmed Alamri[1,3] · Alexandra I. Cristea[1] · Craig D. Stewart[1]

Accepted: 1 March 2021/Published online: 23 March 2021

Abstract

Since their 'official' emergence in 2012 (Gardner and Brooks 2018), massive open online courses (MOOCs) have been growing rapidly. They offer low-cost education for both students and content providers; however, currently there is a very low level of *course purchasing* (less than 1% of the total number of enrolled students on a given online course opt to purchase its certificate). The most recent literature on MOOCs focuses on identifying factors that contribute to student success, completion level and engagement. One of the MOOC platforms' ultimate targets is to become self-sustaining, enabling partners to create revenues and offset operating costs. Nevertheless, analysing learners' purchasing behaviour on MOOCs remains limited. Thus, this study aims to *predict students purchasing behaviour and therefore a MOOCs revenue*, based on the rich array of activity clickstream and demographic data from learners. Specifically, we compare how several machine learning algorithms, namely *RandomForest*, *GradientBoosting*, *AdaBoost* and *XGBoost* can predict course purchasability using a large-scale data collection of 23 runs spread over 5 courses delivered by The University

FutureLearn: a MOOC provider, https://www.futurelearn.com/

✉ Mohammad Alshehri
mohammad.a.alshehri@durham.ac.uk; malshehri@uj.edu.sa

Ahmed Alamri
ahmed.a.alamri@durham.ac.uk; asalamri4@uj.edu.sa

Alexandra I. Cristea
alexandra.i.cristea@durham.ac.uk

Craig D. Stewart
craig.d.stewart@durham.ac.uk

[1] Department of Computer Science, Durham University, South Road, Durham DH1 3LE, UK

[2] Department of Management Information Systems, College of Business, University of Jeddah, Jeddah, Saudi Arabia

[3] Department of Computer Science. College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

of Warwick between 2013 and 2017 via FutureLearn. We further identify the common representative predictive attributes that influence a learner's certificate purchasing decisions. Our proposed model achieved promising accuracies, between 0.82 and 0.91, using only the time spent on each step. We further reached higher accuracy of 0.83 to 0.95, adding learner demographics (e.g. *gender*, *age group*, *level of education*, and *country*) which showed a considerable impact on the model's performance. The outcomes of this study are expected to help design future courses and predict the profitability of future runs; it may also help determine what personalisation features could be provided to increase MOOC revenue.

Keywords Machine learning · MOOCs · Purchasing prediction · Learner analytics

Introduction

Online courses have been around for decades, however they have generally catered to a limited audience. To address this limitation, along with other e-learning challenges, massive open online courses (MOOCs) were developed, specifically to reach an unlimited number of potential learners from around the world. Tracing their history from MIT's 2001 OpenCourseWare initiative, MOOCs entered the modern age of successful *commercialisation* with Stanford's Coursera in 2011 (Ng and Widom 2014), with 2012, coined "*the year of the MOOCs*" (Gardner and Brooks 2018) highlighting the rapid growth of audience and market. Now MOOCs have become increasingly popular and their scale and availability makes it possible to offer a diverse set of learning content to students from all over the world in an accessible and engaging manner. Thus, many MOOC providers such as *FutureLearn,*[1] *edX,*[2] *Udemy*[3] and *Coursera*[4] have started offering scalable online courses to the public. By the end of 2019, the number of MOOCs has reached 13.5 thousand, being delivered via more than 900 university partners around the world, and the total number of MOOC students has reached more than one hundred million (Shah 2018). Nevertheless, there are some indications that the number of registered learners and course populations are declining, due to the platforms' transition from semi-free to paid courses (Chuang and Ho 2016). Despite the unparalleled success of MOOCs, especially in terms of the thriving student enrolment, one of the more concerning aspects to date is the staggeringly low *completion* and *certification* rates, a funnel with students "leaking out" at various points along the learning pathway (Clow 2013; Breslow et al. 2013). While various studies have been conducted to investigate the links between learner behaviours and the course completion (Castaño-Muñoz et al. 2017; Pursel et al. 2016; Hansen and Reich 2015), the race towards identifying precise predictors of completion as well as the *predictors of course purchasing*, continues. Importantly, although MOOCs have started being analysed more thoroughly in the literature, few studies have looked into the characteristics and temporal activities for the purpose of *predicting learners' certification*

[1] www.futurelearn.com
[2] www.edx.org
[3] www.udemy.com
[4] www.coursera.org

decision behaviours. At the same time, the literature shows that user purchasing behaviour has been widely studied on pure e-commerce platforms (Zhang et al. 2018). To date, this kind of behaviour has not been extensively considered in the educational domain, even though MOOC providers have been struggling to build their own sustainable revenues (Dellarocas and Van Alstyne 2013). Considering the recent MOOCs' transition towards paid macro-programmes and online degrees with affiliate university partners, this paper presents a promising model to predict MOOCs purchasers using only the time spent by learners on each step along with their system-logged and manually entered characteristics. Specifically, this article examines the following research questions:

RQ1: can MOOC data be used to predict course purchase based on time spent on each step and learner characteristics?

RQ2: what are the most representative features that contribute towards predicting a learners purchase of MOOCs?

MOOCs and Futurelearn

MOOCs are online education platforms that normally provide course sessions on a weekly basis. Each weekly learning unit consists of several *steps* (tasks), this can be an article, discussion, video or a quiz (Cristea et al. 2018a). This study was applied on a large and diverse corpus (a total of 23 runs[5] spread over 5 courses) from different disciplines, delivered via *FutureLearn,* a UK-based MOOC platform with a wide range of partners of over 80 UK and international universities, 40 specialist organisations and government bodies (FuturLearn 2019). Whilst this platform has a large number of partners, it has not been analysed to the same degree when compared to other counterparts e.g., *Coursera*, *edX* and *Udemy*. FutureLearn is employed by partners to deliver multi-disciplinary educational and professional courses, predominantly free of charge (FutureLearn 2012). However, an option to upgrade for unlimited access and certification at the end of each course is in place, to at least balance the increasing cost of content production and maintenance, which varies from £10,000 to £50,000 per MOOC, with an average of £29,356, to both FutureLearn and its partners (O'Grady and Niamh 2019). Additionally, FutureLearn's ultimate goal is to become a self-sustaining platform, enabling partners to create revenues to offset operating costs (FutureLearn 2015). A Learner has two choices concerning how to learn on the platform. The first is free enrolment, where the access to course content is allowed for the duration of the course only (e.g., 6 weeks only on a 6-weeks course), including an unlimited number of accesses to the course content and materials (quizzes, comments, articles, videos, etc.). At the end of the course, the learner can benefit from one chance to attempt the whole final test, with a limited number of attempts for each question of this test. The second choice is paid enrolment, where the learner can obtain unlimited access to the course content, along with a certificate of achievement, either printed or online, after marking

[5] In FutureLearn, a 'run' refers to an instance of delivery of a given course. These 'runs' are normally delivered pseudo-synchronously (i.e., weeks in a given calendar year are allocated for the learning; however, learners are given flexibility as to when they are actually performing the allocated work).

90% of the course steps as completed, attempting all test questions and achieving a final score of over 70% (Pearson 2015). The course content, including comments and progress, are accessible for learners, as long as the course continues to exist on the platform.

Related Works

Overview

The area of analysing 'big data' and predicting relationships based on the inferenced results is one of the hottest topics of web-related research (Bello-Orgaz et al. 2016), this encompasses several research methods. Whilst, traditionally, educational research does not involve such large datasets, the interest in analysing 'big data' in education has increased with the advent of MOOCs, generating the emerging fields of learner analytics and educational data mining (Atenas 2015). The following subsections shed light on the current literature concerning the prediction of purchase behaviours generally on e-commerce platforms. Next, we present key previous models developed for predicting MOOC learner behaviours based on different features.

Purchasing Behaviour Prediction

Predicting users' attention for purchasing online has been the core of many studies in the field of e-commerce. (Van den Poel and Buckinx 2005) used different types of predictors to explore what contributes to the online shopper's purchasing decision during their next visit. These predictors were several variables derived from four different data sources: general clickstream behaviour at the level of the visit; user characteristics; historical purchase behaviour and some detailed clickstream information. This empirical study showed that predictors from all four categories were retained in the final (best subset) solution. It also indicated that clickstream was the top predictor, due to the role it plays when determining the user tendency to purchase a product. (Joshi et al. 2018) explored the key factors influencing the online purchase behaviour of shoppers across several geographical locations in India. The study proposed and validated a Random Forest-based model for each product category, in order to forecast the online market readability for each category. Retailers were encouraged to use their model for the purpose of identifying online shoppers' preferences at each given location within the state. Similarly, (Chong et al. 2016) identified the most representative factors that could predict online product purchases on Amazon,[6] but with a different method. The data used in the study included the customers' reviews, sentiment derived from the latter and online promotional strategies. The study found that combining the variables above can provide better predictors of sales than using each feature on its own.

These studies have some overlap with ours (in predicting purchasing behaviour), but at the same time are applied in a completely different domain (online retail, as opposed to education). Importantly, purchases tend to be very brief (one stop or a limited

[6] www.amazon.com

number of potentially non-related stops) encounters, whilst education online considers the narrative flow and perquisite structure of the steps. Nevertheless, we based our methods selection on the success of some of these studies, to the extent it was applicable.

Behaviour-Based Prediction

Several works based on statistics, Machine Learning (ML) and visualisation have focused on analyses and predictions of MOOC learner behaviour. (Lu et al. 2017) extracted a large number of features (19) to predict dropout, based on ML methods and support vector machines (SVM), from five courses (one run each only) on Coursera. (Robinson et al. 2016) used NLP techniques to predict dropout on only one HarvardX course; language features were selected via the lasso logistic regression model, and performance was evaluated with Area Under Curve (AUC). (Crossley et al. 2017) conducted a relatively small-scale study (320 students from only one course on Coursera), predicting completion (defined as an overall grade of >= 70%) via a cohesion analysis with the Touchstone Applied Science Associates (TASA)[7] corpus on the forum discussions. Completers and non-completers were compared in their study, based on a wide set of parameters, evaluating performance with recall, precision, and the F1 score. The study indicated that collaborating students have a higher chance of completion.

Activities-Based Prediction

Activity-based prediction focuses mainly on understanding learners on MOOCs, by identifying groups of learners with similar behavioural patterns (Liu et al. 2016a), such as clustering learners based on engagement factors or the number of quizzes attempted (Ferguson and Clow 2015; Kizilcec et al. 2013). Comments, as an important behaviour predictor, have been studied in many setups, including MOOCs. (Liu et al. 2016b) emphasised the importance of using ML methods to analyse MOOC comments, to detect the emotions of learners and predict the popularity of each course. (Zhang et al. 2016) focused on MOOC students self-grouping based on their preferences through conducting an online pre-course survey. This study explores the preferences of these learner groups using different types of communication (asynchronous text posts, synchronous text chats, or synchronous video and audio). The study showed that learners, based on their demographics, have different preferences towards the type of communication used. For example, female learners preferred asynchronous text-based posts more than their male peers. (Dmoshinskaia 2016) investigated the dropout rate via analysing two MOOC courses with 176 learners' comments on different objects (video, articles, exercises etc.). While several studies found an inverse relationship between the low sentiment score and the dropout rate, this study surprisingly indicated that learners with no negative comments posted are likely to drop the course very soon. Using conditional inference trees, students were clustered into different groups based on the correlation between a set of variables and the completion rate. The claim that learners with no negative comments may leave the course very soon, indicated that a

[7] http://lsa.colorado.edu/spaces.html

critical cognitive involvement is vital for succeeding. Their key finding was that students' negative expressions can be a potentially positive signal for more involvement and potential completion. (Wen et al. 2014) explored the relationship between the sentiment ratio, measured based on daily forum posts, and the number of learners who drop out each day. The study recommended using sentiment analysis with caution, while analysing noisy and quantity-limited comments. In addition, Chua et al. (Chua et al. 2017) and Tubman et al. (Tubman et al. 2016) analysed learner commenting behaviours, having explored patterns of discussion that occur in MOOCs. More recently, (Alshehri et al. 2018; Cristea et al. 2018b) examined how basic learner characteristics, such as gender, can influence learning behaviours, such as the patterns of making comments on different learning steps. More recently, (Alamri et al. 2019; Cristea et al. 2018c) addressed early dropout issue using the student's first registration date on the course and two learner activity-based features: time spent by learner and number of access from the first week data only. The study targeted the traditional challenge of MOOCs of high dropout rate - as only 3% of the registered students completed the course. Additionally, the study surprisingly showed that non-completers registered on the course way earlier compared to completers which proves that late enrolment is not necessarily a sign of later dropout.

Demographics-Based Prediction

Some previous explanatory data models attempted to analyse and predict learner success based on demographics. Their ultimate goal was to identify which group of MOOCs can be effective as tools for life-long learning. Although the majority of research on MOOC demographics investigated the experience of higher education students, the literature shows that most MOOC learners are professional, highly educated and from developed nations (de Waard 2017). Concerning gender, as one of the most important predictors of learner behaviour, there is a significant variance among MOOC platforms, with more male learners on EdX and Coursera and more females on FutureLearn (Liyanagunawardena et al. 2015). The differences in learners' demographics play an effective role in exploring learning outcomes. Closer to our own current study, (Morris et al. 2015) predicted learning outcomes (i.e., certificate earners) using learners' demographics, collected by pre-course questionnaires from five FutureLearn courses. The study suggests that some demographics, such as age and employment status, can strongly predict completion. The highest median age among completers is 43 years, whereas the first week's dropout median age is 34 years, as shown in the study. Similar studies on learner demographic-based navigation in MOOCs found that the older the learner, the greater the chance for certificate earning and repeated viewing of lecture sequences. However, it was noted that, as MOOCs often target non-predefined audiences from different backgrounds to reach the same defined learning goal, further examination at the level of learners' context should be conducted, as MOOC audience demographics are more complicated than the headline data denotes (Morris 2014). In addition, the authors recommended that true certification-leading engagement and motivation analysis cannot be measured by analysing the log data only, and representative learner demographic features should be involved to appropriately address learner behaviour (Guo and Reinecke 2014). (Shi and Cristea 2018) investigated two FutureLearn courses with six runs in total, noticing that gender and education may influence students' behaviours in terms of comments posted, questions attempted and steps (pages) completed. With regard to age versus success, (Packham et al. 2004) found that successful learners are female and aged between 31 and 50 years, regardless of their educational level and employment status, whereas (Ke and Kwak 2013) reported that older learners invest more time in online participation. (González-Gómez et al. 2012) suggested that males have more positive attitudes towards online learning, due to their higher computer self-efficiency, whilst (Vail et al. 2015) showed that females and male students benefit differently from adaptive support.

Certification Prediction

Several studies investigated MOOC certification; nevertheless, they have not explained precisely whether certification involves paying for the certificate at the end of the course or it is just a consequence resulting from the course completion. Table 1 shows all the predictive models that were formerly proposed for certification prediction since the rapid growth of MOOCs in 2012. Our survey selection criteria included works that: proposed predictive models; applied on MOOCs only datasets and aimed to predict certification. (Reich 2014) studied the relationship between intention of completion, and actual completion & certificate earning. The study was applied on 9 HarvardX MOOCs and showed that the correlation between the two variables was a stronger predictor of certification rather than any demographic traits. (Howarth et al. 2017) studied MOOC learners' subsequent educational goals after taking the course, by using consumer goal theory. They examined how completing a MOOC can motivate a learner to join the providing university to take a degree course. The study shows that MOOC completers with a certain deal of satisfaction with the course delivery, were more likely to progress to the course-host institution, than those who did not complete the course. This study showed that having a similar pedagogical and delivery approach in a university for both conventional and online courses can encourage learners to join further academic online study. This study became a roadmap for tertiary institutes on how to design an effective MOOC that can target potential future students. (Jiang et al. 2014) predicted MOOC certification using the first week behaviour (average quiz score, number of completed peer assessments, social network degree and being either a current or prospective student at the university offering the course). This logistic regression classifier model was trained and tested on one run of the MOOCs under certain conditions and incentives, by the provider; therefore, it might need to be replicated, for the results to be generalisable. Qiu et al. (Qiu et al. 2016) extracted factors of engagement in XuetangX (China, partner of edX) on 11 courses, to predict grades and certificate earning with different methods (LRC, SVM, FM, LadFG); their performance was evaluated using the area under the curve (AUC), precision, recall, and F1 score. However, the number of used features for demographics (gender, age, education), from forums (number of new posts and replies), learning behaviour (chapters browsed, deadlines completed, time spent on videos, doing assignments, etc.), courses delivery windows (delivered within 8 months only) and study learners (around 88,000) are relatively low. (Ruipérez-Valiente et al. 2017) used four different algorithms (RF, GB, k-NN and LR) to predict student certification on one edX-delivered course. They used a total of eleven independent variables to build the model and predict

Table 1 Certification Prediction Models versus our Model, see table 2 for key to abbreviations

Study	**Data Source**	**#Courses**	**#Students**	**Data Type**	**Approach**	**Metrics**	**PA**
Ramesh at al. (2013) (Ramesh et al. 2013)	Coursera	1	826	CS; FP	PSL	ACC; AUC	CV
Jiang et al. (2014) (Jiang et al. 2014)	Coursera	1	37,933	ASSGN; FP; SIS	LR	ACC; AUC; PREC; REC; F1	CV
Reich (2014) (Reich 2014)	HarvardX	9	79,525	DEM; SURV	LR; SM	ACC	n/a
Coleman et al. (2015) (Coleman et al. 2015)	edX	1	43,758	CS	LDA	ACC; REC	CV
Joksimovic et al. (2016) (Joksimović et al. 2016)	Coursera	1	84,786	FP	ERGM; LR	AIC	n/a
Qiu et al. (2016) (Qiu et al. 2016)	XuetangX	11	88,112	CS	LadFG; LR; SVM; FM	AUC; PREC; REC; F1	n/a
Xu and Yang (2016) (Xu and Yang 2016)	HarvardX- MITx	10	n/a	CS; FP	SVM	ACC	T&T
Ruipérez-Valiente at al. (2017)	edX	1	n/a	CS	RF; GB; kNN; LR	ACC; AUC; F1; Sen.; Spec.; Cohen's kappa coefficient	CV
Gitinabard et al. (2018) (Gitinabard et al. 2018)	Coursera; edX	1	65,203	CS; FP	LR	AUC; F1	CV
Our Model	FutureLearn	**9**	245,255	DEM; CS	RF, GB, AdaB, XGB	ACC; PREC; REC; F1	T&T; CV

Table 2 List of Abbreviations and Acronyms

Column	**Abbreviation / Acronym**	**Description**
Data Type	CS	Clickstream
Data Type	FP	Forum Posts
Data Type	DEM	Demographics
Data Type	ASSGN	Assignments
Data Type	SIS	Student Information System
Data Type	SURV	Survey
Approach	PSL	Probabilistic Soft Logic
Approach	LR	Logistic Regression
Approach	SM	Survival Model
Approach	LDA	Latent Dirichlet Allocation
Approach	ERGM	Exponential Random Graph Model
Approach	LadFG	Laten Dynamic Factor Graph
Approach	FM	Factorisation Machine
Approach	RF	Random Forest
Approach	GB	Gradient Boosting
Approach	KNN	K-Nearest Neighbour
Approach	AdaB	Adaptive Boosting
Approach	XGB	XGBoostnig
Metrics	ACC	Accuracy
Metrics	AUC	Area Under Curve
Metrics	PREC	Precision
Metrics	REC	Recall
Metrics	F1	F1-score
Metrics	Sen	CV
Metrics	Spec	CV
PA	CV	Cross Validation
PA	T&T	Train/Test Split Validation

the dependent variable – the acquisition of a certificate (true or false). (Gitinabard et al. 2018) more recently used behavioural and social features of one course “Big Data in Education”, which was first offered on Coursera and later on edX, to predict dropout and certification. In Table 1 we summarise the list of the surveyed certification prediction models followed by a list of the abbreviations and acronyms and used (Table 2).

Unlike the previously conducted studies concerning certification, our proposed model aims to predict the *financial decisions* of learners on whether or not to purchase the course after completion. Also, our work is applied to a less frequently studied platform, that of FutureLearn, as Table 1 clearly shows. Another concern is the low number of courses analysed, six out of the total nine studies were conducted with one course only. As students may behave differently based on the course being attended, the proposed models may not be suitable for broader generalisation. Our proposed model is based on a variety of courses from different disciplines: Literature, Psychology, Computer Science and Business. Another novelty in this study is predicting the learner's real financial decision on buying the course and gaining a certificate. The majority of the course purchase prediction models identify certification as an automatic consecutive step to the completion making them not different from completion predictors. This study identifies the most representative factors helpful for certification purchase prediction. It also proposes a four algorithm-based model for predicting MOOC purchasability using relatively few input features.

Methodology

Data Collection

When a learner joins FutureLearn for the first time, they are directly prompted to complete an *optional* survey about their characteristics. Existing learners are also prompted to complete this data if missing (Alshehri et al. 2018). All questions on the survey are optional and they aim to extract certain information about a learner's *gender, age group, education level, country, employment status and employment area*. In parallel, the system generates logs to correlate unique IDs and time stamps to learners, recording learner activities, such as weekly-based steps visited, completed, comments added, or question attempts (Alshehri et al. 2018). The current study is analysing data extracted from a total of 23 runs spread over 5 MOOC courses, on 4 distinct topic areas, all delivered through FutureLearn, by the University of Warwick, these topic areas are:

- *literature* (with course Shakespeare and his World [SP]; with course duration 10 weeks)
- *psychology* (with courses The Mind is Flat [TMF]: six weeks, and Babies in Mind [BIM]: four weeks)
- *computer science* (with course Big Data [BD]: nine weeks)
- *business* (with course Supply Chains [SC]: six weeks)

These courses were delivered repeatedly in consecutive years (2013–2017), thus we have data on several '*runs'* for each course. Table 3 below shows the number of enrolled, unenrolled, completed students, as well as those having purchased a

Table 3 The number of enrolled, unenrolled, completed and purchased students on 5 FutureLearn courses

Course	**#Runs**	**#Weeks**	**#Enrollers**	**#Un-enrollers**	**#Completers**	**#Purchasers**	**#SR**
BIM	6	4	48,777	3385	6963	676	3585
BD	3	9	33,430	4006	2359	268	2236
SP	5	10	63,630	9139	6304	750	6135
SC	2	6	5810	463	320	71	617
TMF	7	6	93,608	15,936	5315	321	5571
Total	23	35	245,255	32,929	21,261	2086	18,144

certificate. It also shows how many survey respondents (SR) of the total number of students in each course. Students *accessed 3,007,789 materials* and declared *2,794,578 steps completed*. Regarding these massive numbers, Table 3 clearly illustrates the low completion rate (almost 10% of the total number of registered students) and, more importantly, the even lower rate of purchasing students (less than 1% of the enrolled students).

Data Pre-Processing

The obtained dataset went through several processing steps, in order to be prepared and fed into the learning model. Due to the fact that some students were found to be enrolled on more than one run of the same course, the run number was attached to the student's ID, in order to avoid any mismatch during joining student activities "of several runs" with their characteristics. The step-based *Time Spent* feature, *tspent*, which we used here for prediction and proved to be a highly representative factor helpful for students' purchasing prediction, represents a computed value (rather than being provided as a log variable within the obtained dataset). This feature was defined as the difference between the first time a given student accesses a step *(first_visited time stamp)*, *tvisit*, and when that step is fully completed *(last_completed time stamp)*, *tcompleted*, as per student's declaration (by pressing the 'Mark as Completed' button), as shown in Fig. 1.

$$tspent(u,s) = tcompleted(u,s) - tvisit(u,s) \tag{1}$$

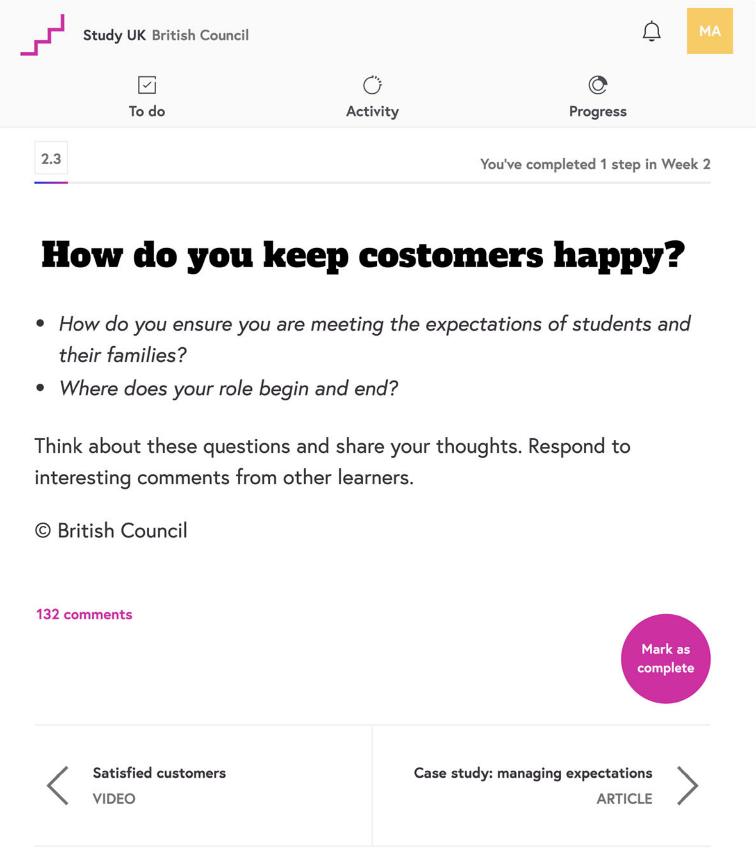
where u represents the current user, s the current step.

As our dataset has several logged dates and respective times for various activities of the students in the system, the *pd.to_datetime* function was applied, to convert these variables into a set of strings (year, month, day, hour, minute) to enrich our input features and allow for an as high performance as possible with the few features available, as well as taking into account the aim to use features available early in each run to allow for early predictability. The latter aspect is critical as it means that course providers could use our prediction model to create early interventions and thus guide more of the students towards paying behaviour, potentially increasing their revenue.

We trained and tested our model on different subsets of features, as well as different versions of the same feature (raw data or derived), such as: *time spent* only, and *time spent and characteristics*, in order to identify the potential representability of learner characteristics on the model performance. The pre-processing further contained some standard data manipulations, such as processing (replacing) missing values with zeros using the *fillna(0)* function and converting category-shaped characteristics into integers with *astype('category')*. We also used *apply(lambda x: x.cat.codes)* and *pd.factorize* along with Pandas (McKinney 2010) and NumPy (Oliphant 2006) to render the data format as machine-feedable. The pre-processing further contained eliminating irrelevant data generated by organisational administrators (455 admins across the 23 runs analysed).

Learning Algorithms

The purchasability problem is formulated as a binary classification problem. The current study applied four different classification and boosting algorithms as follows:



You've completed 1 step in Week 2

How do you keep costomers happy?

- *How do you ensure you are meeting the expectations of students and their families?*
- *Where does your role begin and end?*

Think about these questions and share your thoughts. Respond to interesting comments from other learners.

© British Council

132 comments

Fig. 1 The Interaction Component on a Weekly *step* Page

- Random Forest: An ensemble learning method that uses a combination of tree predictors where each tree depends on the value of a random vector. This algorithm samples these vectors independently and with the same distribution for all decision trees in the forest (Breiman 2001).
- Adaptive Boosting (AdaBoost): One of the boosting algorithm which was introduced by Freund and Schapire (Freund and Schapire 1997) to solve many complex problems, such as gambling, multiple-outcome prediction and repeated games. AdaBoost can be used in conjunction with other learning algorithms to improve performance via combining the output of other learning algorithms (weak learners) (Freund and Schapire 1997).
- Gradient Boosting: This algorithm utilises ensemble weak prediction models to generate a robust a prediction model for solving regression and classification problems. The optimisation of differentiable loss function here is employed for the purpose of generalisation (Friedman 2001).
- XGBoost: An optimised distributed gradient boosting based library with more efficiency, flexibility and portability. As a scalable end-to-end tree boosting system, XGBoost achieves better results with sparse data and weighted quantile sketch for approximate tree learning (Chen and Guestrin 2016).

The above algorithms were chosen due to the fact that they were deemed to be able to predict course purchasability well, by using only very few features: *time spent* on each step, *run number*[8] and learner *demographic characteristics*, i.e., gender, age group, education level, country. These features are believed to be timely logged, such as *time stamped activities*, or collected early on, such as *learner's demographics*, at registration time by any standard MOOC system, which would promote our model as generalisable. There are some further features that can be utilised for learner behaviour prediction, e.g., quizzes or leaving surveys; these features are either not generated by every MOOC platform, or logged later in the run, making early prediction of purchasing behaviour challenging. We have, in addition, identified the most representative features for MOOCs purchasing prediction, as shown in the results section. The results were validated using a couple of architectures: *Train/Test Split Validation* and *10-fold Cross Validation* as shown in the results section below.

Results and Discussion

In order to forecast the learner's intention for course purchasing, we have applied 4 predictive algorithms which reported 4 performance metrics using our proposed model. Tables 4 and 5 show the 4 metrics (*Accuracy*, *Precision*, *Recall* and *F1-score*) measurements based on dual sets of input data: time spent (tspent) only and tspent and characteristics. In Table 4, the results were obtained using learner's activities only (left) and learner's activities and characteristics (right). The results in general showed promising accuracy ranges of 0.82–0.89 for the *tspent* only with T/T split validation, 0.81–0.94 for *tspent and characteristics* with 10-fold split cross validation "being the highest range", 0.82–0.91 for *tspent* only with T/T split validation and 0.83–0.94 for *tspent and characteristics* with 10-fold split cross validation. Adding learner characteristics has slightly increased the model performance compared to using *tspent* features only, with exception to Shakespeare with T/T split validation. Thus increasing time spent on the MOOC platform is a desirable target for course designers; they can further refine this by targeting different adaptive strategies based on the learners' characteristics. Please note that time as a prediction variable is controversial in online platforms, due to it being potentially influenced by other external variables (e.g., by transmission rate, etc.). However, the fact that it is such a powerful predictor here speaks for it being an eminently usable indicator in our case.

Figure 2 shows the most representative features utilised to predict learners' purchasing intention. As it can be clearly seen, the 'Run' played a significant role in predicting Purchasability for many of the courses: it is the most representative factor in 'The Mind is Flat' course, the second one in 'Shakespeare' and 'Babies in Mind', the fourth and seventh factors in 'Supply Chains' and 'Big Data', respectively. Thus, the current number of the run (1 for first time a course is delivered, 2 for the second, etc.) is relevant here. It is noteworthy that, as runs usually are delivered at 1-year interval on FutureLearn, consecutive runs of the same course do not necessarily contain the same steps or step types (video, quiz, article, discussion). Therefore, adding the 'Run' number as an input feature can result in a high representation score and improve the

[8] Thus the first time a course is run, the number is 1, the second time 2, etc.

Table 4 Train/Test Split Validation

Course/Algorithm	**Time Spent only**							**Time Spent & Characteristics**						
	Acc.	***Precision***		***Recall***		***F1-score***		***Acc.***	***Precision***		***Recall***		***F1-score***	
		0	**1**	**0**	**1**	**0**	**1**		**0**	**1**	**0**	**1**	**0**	**1**
Big Data														
RandomForest	0.88	0.92	0.84	0.83	0.92	0.87	0.88	0.89	0.92	0.88	0.86	0.92	0.89	0.90
G.Boosting	0.87	0.92	0.82	0.81	0.92	0.86	0.87	0.90	0.90	0.91	0.90	0.91	0.90	0.91
AdaBoost	0.88	0.90	0.85	0.88	0.85	0.88	0.87	0.89	0.90	0.89	0.88	0.91	0.89	0.90
XGBoost	0.88	0.90	0.85	0.85	0.90	0.88	0.87	0.90	0.90	0.91	0.90	0.91	0.90	0.91
Babies in Mind														
RandomForest	0.82	0.88	0.78	0.71	0.91	0.79	0.84	0.81	0.89	0.76	0.72	0.91	0.79	0.83
G.Boosting	0.86	0.83	0.89	0.89	0.83	0.86	0.86	0.90	0.91	0.90	0.90	0.91	0.90	0.91
AdaBoost	0.87	0.81	0.93	0.94	0.80	0.87	0.86	0.89	0.89	0.89	0.89	0.89	0.89	0.89
XGBoost	0.88	0.85	0.91	0.90	0.85	0.88	0.88	0.92	0.93	0.91	0.91	0.93	0.92	0.92
Shakespeare														
RandomForest	0.85	0.97	0.78	0.74	0.98	0.84	0.87	0.87	0.96	0.81	0.76	0.97	0.85	0.88
G.Boosting	0.89	0.93	0.85	0.85	0.93	0.88	0.89	0.85	0.83	0.88	0.88	0.83	0.85	0.85
AdaBoost	0.91	0.93	0.88	0.88	0.93	0.91	0.90	0.87	0.83	0.90	0.91	0.83	0.87	0.86
XGBoost	0.89	0.92	0.85	0.86	0.92	0.89	0.88	0.84	0.82	0.86	0.86	0.83	0.84	0.84
Supply Chains														
RandomForest	0.86	0.81	0.92	0.93	0.79	0.87	0.85	0.93	1.00	0.83	0.89	1.00	0.94	0.91
G.Boosting	0.82	0.76	0.91	0.93	0.71	0.84	0.80	0.89	1.00	0.77	0.83	1.00	0.91	0.87
AdaBoost	0.82	0.80	0.85	0.86	0.79	0.83	0.81	0.93	0.94	0.90	0.94	0.90	0.94	0.90
XGBoost	0.82	0.80	0.85	0.86	0.79	0.83	0.81	0.93	1.00	0.83	0.89	1.00	0.94	0.91
The Mind is Flat														
RandomForest	0.86	0.98	0.78	0.75	0.98	0.85	0.87	0.85	0.96	0.77	0.75	0.96	0.84	0.86
G.Boosting	0.89	0.91	0.88	0.88	0.91	0.90	0.89	0.91	0.92	0.91	0.92	0.91	0.92	0.91
AdaBoost	0.87	0.87	0.87	0.88	0.85	0.87	0.86	0.94	0.95	0.93	0.93	0.94	0.94	0.93
XGBoost	0.87	0.88	0.85	0.86	0.87	0.87	0.86	0.89	0.90	0.89	0.90	0.89	0.90	0.89
Average	0.86	0.88	0.86	0.85	0.87	0.86	0.86	0.89	0.92	0.87	0.87	0.92	0.89	0.89

model performance. This shows potentially that refining a course might lead to better appreciation by the students and ultimately to certificate purchase. The results also show that the purchasing behaviour can be predicted fairly accurately from a set of relatively easy to obtain and process features.

Conclusion

MOOCs have been around for few years and been growing rapidly, offering an affordable education. Nevertheless, the currently very low level of course purchasing is a sustainability challenge (less than 1% of the total number of enrolled students).

Table 5 K-fold Cross Validation

Course/Algorithm	**Time Spent only**							**Time Spent & Characteristics**						
	Acc.	*Precision*		*Recall*		*F1-score*		Acc.	*Precision*		*Recall*		*F1-score*	
		0	1	0	1	0	1		0	1	0	1	0	1
Big Data														
RandomForest	0.86	0.87	0.86	0.85	0.87	0.86	0.86	0.87	0.88	0.86	0.85	0.88	0.86	0.87
G.Boosting	0.86	0.87	0.85	0.85	0.87	0.86	0.86	0.92	0.94	0.90	0.90	0.94	0.91	0.92
AdaBoost	0.85	0.85	0.86	0.86	0.84	0.85	0.85	0.90	0.90	0.90	0.90	0.90	0.90	0.90
XGBoost	0.87	0.87	0.88	0.88	0.86	0.87	0.87	0.92	0.94	0.90	0.90	0.94	0.92	0.92
Babies in Mind														
RandomForest	0.82	0.88	0.78	0.74	0.90	0.80	0.83	0.83	0.93	0.76	0.71	0.95	0.80	0.85
G.Boosting	0.86	0.85	0.88	0.88	0.85	0.87	0.86	0.90	0.90	0.91	0.91	0.90	0.91	0.90
AdaBoost	0.85	0.83	0.89	0.90	0.81	0.86	0.85	0.90	0.89	0.91	0.91	0.89	0.90	0.90
XGBoost	0.86	0.85	0.88	0.89	0.84	0.87	0.86	0.90	0.90	0.91	0.91	0.90	0.91	0.90
Shakespeare														
RandomForest	0.85	0.94	0.79	0.74	0.95	0.83	0.86	0.85	0.96	0.76	0.70	0.97	0.81	0.85
G.Boosting	0.87	0.88	0.85	0.85	0.88	0.86	0.87	0.89	0.89	0.88	0.88	0.89	0.88	0.89
AdaBoost	0.88	0.90	0.87	0.86	0.90	0.88	0.88	0.88	0.89	0.87	0.87	0.90	0.88	0.88
XGBoost	0.88	0.91	0.86	0.85	0.91	0.88	0.89	0.89	0.90	0.88	0.87	0.90	0.88	0.89
Supply Chains														
RandomForest	0.91	0.90	0.92	0.93	0.90	0.91	0.91	0.89	0.88	0.90	0.90	0.88	0.89	0.89
G.Boosting	0.88	0.87	0.89	0.90	0.87	0.88	0.88	0.93	0.93	0.93	0.93	0.93	0.93	0.93
AdaBoost	0.87	0.85	0.89	0.90	0.84	0.87	0.86	0.91	0.90	0.92	0.93	0.90	0.91	0.91
XGBoost	0.90	0.89	0.91	0.91	0.88	0.90	0.90	0.93	0.93	0.94	0.94	0.93	0.94	0.94
The Mind is Flat														
RandomForest	0.84	0.93	0.78	0.73	0.94	0.82	0.85	0.86	0.95	0.79	0.75	0.96	0.84	0.87
G.Boosting	0.87	0.88	0.87	0.87	0.88	0.87	0.87	0.92	0.91	0.94	0.94	0.90	0.92	0.92
AdaBoost	0.87	0.87	0.87	0.88	0.87	0.87	0.87	0.93	0.92	0.93	0.94	0.92	0.93	0.93
XGBoost	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.92	0.91	0.92	0.93	0.91	0.92	0.92
Average	0.87	0.88	0.86	0.86	0.88	0.86	0.87	0.9	0.91	0.89	0.88	0.91	0.89	0.9

Although one of a MOOC platforms' ultimate goals is to become self-sustaining, enabling partners to create revenues and offset operating costs, the analysis of learners' purchasing behaviour on MOOCs remains extremely limited. This study predicts students' purchasing behaviour and MOOCs revenues, based on a learner's activity clickstream and demographic data derived from the MOOC platform of futurelearn.com. We used and compared how several machine learning algorithms "RandomForest, GradientBoosting, AdaBoost and XGBoost" can predict course purchasability. We further identified the common representative predictive attributes that influence learners' certificate purchasing decisions. Our proposed model achieved promising accuracies, between 0.82 and 0.91, using the time spent on each step only. We further reached higher accuracies of 0.83 to 0.95, by adding learner demographics,

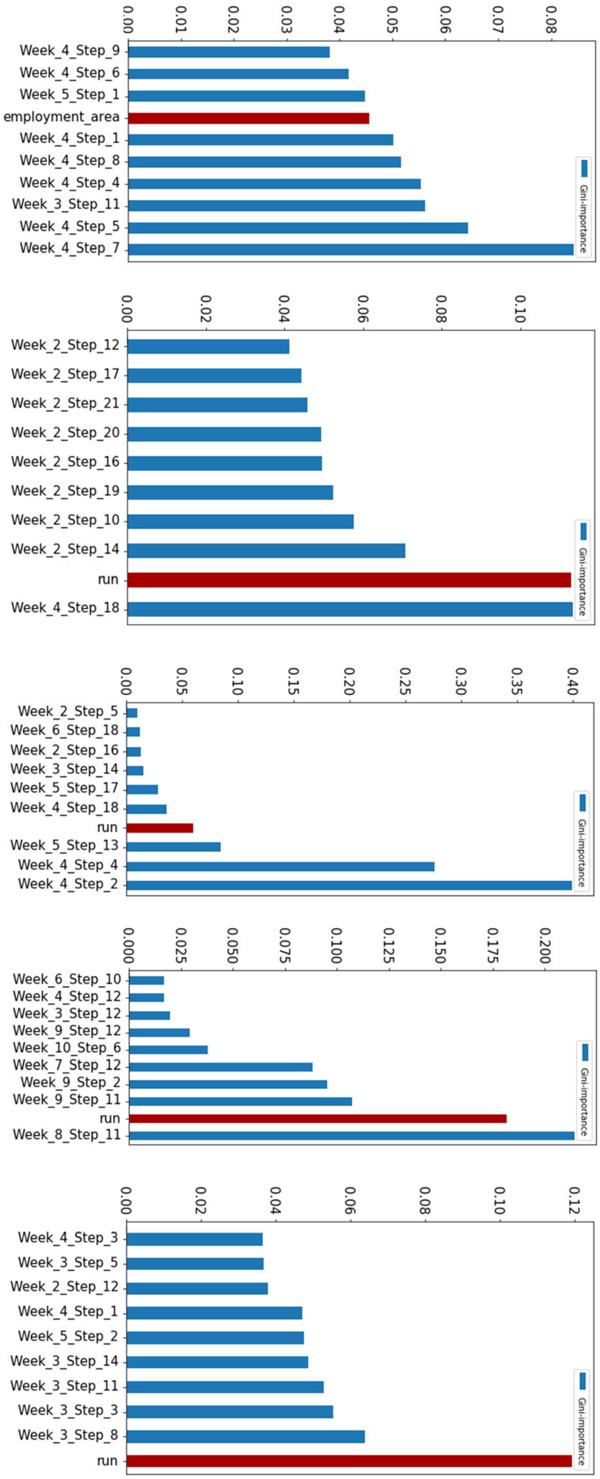


Fig. 2 The Most Representative Features for Predicting Course Purchasing Behaviour

e.g., gender, age group, level of education, and country, which showed a considerable impact on the model performance. Importantly, we show that a fairly straightforward data collection can be performed to identify the likelihood of purchases of courses. This can further lead to course design changes that take these parameters into account, in order to influence the learner, and thus potentially increase the desired purchasing behaviour frequency. However it is possible that other factors may be influencing the purchasing behaviour, which were not analysed in this study. Our immediate future work will be including analysing data on "comments and quizzes" and identifying whether they can be employed for greater identification of purchasing behaviours and further optimising the results. Moreover, our work is limited by analysing a single platform, FutureLearn; further work will look into performing similar analyses on different platforms. Finally, time, as discussed, is a potentially controversial variable to use on online platforms, and a time series prediction approach for purchasability (where actions/steps are recorded as ordered transitions) could also be pursued – although the Occam's razor principle further supports our current approach.

Acknowledgments The authors wish to thank The University of Warwick for providing access to their FutureLearn courses.

References

Alamri, A. et al. (2019). Predicting MOOCs dropout using only two easily obtainable features from the first week's activities. In International Conference on Intelligent Tutoring Systems. Springer.

Alshehri, M. et al. (2018). On the need for fine-grained analysis of Gender versus Commenting Behaviour in MOOCs. In Proceedings of the 2018 The 3rd International Conference on Information and Education Innovations. ACM.

Atenas, J. (2015). Model for democratisation of the contents hosted in MOOCs. *International Journal of Educational Technology in Higher Education, 12*(1), 3–14.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59.

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

Breslow, L., et al. (2013). Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment, 8*, 13–25.

Castaño-Muñoz, J., Kreijns, K., Kalz, M., & Punie, Y. (2017). Does digital competence and occupational setting influence MOOC participation? Evidence from a cross-course survey. *Journal of Computing in Higher Education, 29*(1), 28–46.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Chong, A. Y. L., Li, B., Ngai, E. W. T., Ch'ng, E., & Lee, F. (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies. *International Journal of Operations & Production Management, 36*, 358–383.

Chua, S.M., et al. (2017). Discussion Analytics: Identifying Conversations and Social Learners in FutureLearn MOOCs.

Chuang, I. and Ho, A. (2016). HarvardX and MITx: Four years of open online courses–fall 2012-summer 2016.

Clow, D. (2013). MOOCs and the funnel of participation. In Proceedings of the third international conference on learning analytics and knowledge. ACM.

Coleman, C.A., Seaton, D.T., and Chuang, I. (2015). Probabilistic use cases: Discovering behavioral patterns for predicting certification. in Proceedings of the second (2015) acm conference on learning@ scale.

Cristea, A.I., et al. (2018a). How is learning fluctuating? FutureLearn MOOCs fine-grained temporal analysis and feedback to teachers.

Cristea, A.I. et al. (2018b). Can learner characteristics predict their behaviour on MOOCs? In Proceedings of the 10th International Conference on Education Technology and Computers.

Cristea, A.I. et al. (2018c). Earliest predictor of dropout in MOOCs: a longitudinal study of FutureLearn courses. *Association for Information Systems*.

Crossley, S., et al. (2017). *Predicting success in massive open online courses (MOOCs) using cohesion network analysis*. Philadelphia: International Society of the Learning Sciences.

de Waard, I. (2017). Self-directed learning of experienced adult online learners enrolled in FutureLearn MOOCs. *The Open University*.

Dellarocas, C., & Van Alstyne, M. W. (2013). Money models for MOOCs. *Communications of the ACM, 56*(8), 25–28.

Dmoshinskaia, N. (2016). Dropout prediction in MOOCs: using sentiment analysis of users' comments to predict engagement. University of Twente.

Ferguson, R. and Clow, D. (2015). Examining engagement: analysing learner subpopulations in massive open online courses (MOOCs). In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge. ACM.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.

Friedman, J.H., Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 2001: p. 1189–1232, 29.

FutureLearn *FutureLearn Launch*. 2012.

FutureLearn. *5 ways to prepare for the workplace of 2026, starting right now*. 2015 18/03/2019].

FuturLearn. *Course Providers, Current Partners*. 18/03/2019].

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction, 28*(2), 127–203.

Gitinabard, N., et al., (2018). Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features. arXiv preprint arXiv:1809.00052.

González-Gómez, F., Guardiola, J., Martín Rodríguez, Ó., & Montero Alonso, M. Á. (2012). Gender differences in e-learning satisfaction. *Computers & Education, 58*(1), 283–290.

Guo, P.J. and Reinecke, K. (2014) Demographic differences in how students navigate through MOOCs. In Proceedings of the first ACM conference on Learning@ scale conference. ACM.

Hansen, J.D. and Reich, J. (2015). Socioeconomic status and MOOC enrollment: enriching demographic information with external datasets. In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. ACM.

Howarth, J., D'Alessandro, S., Johnson, L., & White, L. (2017). MOOCs to university: A consumer goal and marketing perspective. *Journal of Marketing for Higher Education, 27*(1), 144–158.

Jiang, S., et al. (2014). Predicting MOOC performance with week 1 behavior. in Educational data mining 2014.

Joksimović, S. et al. (2016) Translating network position into performance: importance of centrality in different network configurations. in Proceedings of the sixth international conference on learning analytics & knowledge.

Joshi, R., Gupte, R., & Saravanan, P. (2018). A random forest approach for predicting online buying behavior of Indian customers. *Theoretical Economics Letters, 8*(03), 448–475.

Ke, F., & Kwak, D. (2013). Online learning across ethnicity and age: A study on learning interaction participation, perception, and learning satisfaction. *Computers & Education, 61*, 43–51.

Kizilcec, R.F., Piech, C., and Schneider, E. (2013). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In Proceedings of the third international conference on learning analytics and knowledge. ACM.

Liu, Z., et al. (2016a). MOOC learner behaviors by country and culture; an exploratory analysis. *EDM, 16*, 127–134.

Liu, Z. et al. (2016b). Emotion and associated topic detection for course comments in a MOOC platform. In Educational Innovation through Technology (EITT), 2016 International Conference on. IEEE.

Liyanagunawardena, T. R., Lundqvist, K. Ø., & Williams, S. A. (2015). Who are with us: MOOC learners on a F uture L earn course. *British Journal of Educational Technology, 46*(3), 557–569.

Lu, X. et al. (2017). What Decides the Dropout in MOOCs? In International Conference on Database Systems for Advanced Applications. Springer.

McKinney, W. (2010). Data structures for statistical computing in python. in Proceedings of the 9th Python in Science Conference. Austin, TX.

Morris, N.P. (2014) How Digital Technologies, Blended Learning and MOOCs Will Impact the Future of Higher Education. ERIC.

Morris, N.P., Hotchkiss, S., and Swinnerton, B. (2015). Can demographic information predict MOOC learner outcomes. *Proceedings of the European MOOC Stakeholder Summit*, 199–207.

Ng, A. and Widom, J. (2014). Origins of the Modern MOOC (xMOOC). Hrsg. Fiona M. Hollands, Devayani Tirthali: MOOCs: Expectations and Reality: Full Report: p. 34–47.

O'Grady and Niamh. Do Moocs generate return on investment? [cited 2019 18/03/2019]; Available from: https://about.futurelearn.com/research-insights/do-moocs-generate-return-on-investment.

Oliphant, T.E. (2006). A guide to NumPy. Vol. 1. Trelgol Publishing USA.

Packham, G., et al. (2004). E-learning and retention: Key factors influencing student withdrawal. *Education+ Training,* ***46***(6/7), 335–342.

Pearson, S. (2015). Introducing more detailed, more rigorous certificates of achievement.

Pursel, B. K., Zhang, L., Jablokow, K. W., Choi, G. W., & Velegol, D. (2016). Understanding MOOC students: Motivations and behaviours indicative of MOOC completion. *Journal of Computer Assisted Learning, 32*(3), 202–217.

Qiu, J. et al. (2016). Modeling and predicting learning behavior in MOOCs. In Proceedings of the ninth ACM international conference on web search and data mining. ACM.

Ramesh, A., et al. (2013). Modeling learner engagement in MOOCs using probabilistic soft logic. in NIPS workshop on data driven education.

Reich, J. (2014). MOOC completion and retention in the context of student intent. EDUCAUSE Review Online, 8.

Robinson, C. et al. (2016). Forecasting student achievement in MOOCs with natural language processing. In Proceedings of the sixth international conference on learning analytics & knowledge. ACM.

Ruipérez-Valiente, J.A. et al. (2017). Early prediction and variable importance of certificate accomplishment in a MOOC. In European Conference on Massive Open Online Courses. Springer.

Shah, D. (2018). By The Numbers: MOOCs in 2018.

Shi, L. and Cristea, A.I. (2018). Demographic indicators influencing learning activities in MOOCs: Learning analytics of FutureLearn courses.

Tubman, P., Oztok, M., and Benachour, P. (2016). Being social or social learning: A sociocultural analysis of the FutureLearn MOOC platform. In Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on. IEEE.

Vail, A.K., et al. (2015). The mars and venus effect: the influence of user gender on the effectiveness of adaptive task support. In International Conference on User Modeling, Adaptation, and Personalization. Springer.

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research, 166*(2), 557–575.

Wen, M., Yang, D., and Rose, C.. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? in Educational data mining 2014. Citeseer.

Xu, B., & Yang, D. (2016). Motivation classification and grade prediction for MOOCs learners. *Computational Intelligence and Neuroscience, 2016*, **1**–**7**.

Zhang, Q., Peck, K. L., Hristova, A., Jablokow, K. W., Hoffman, V., Park, E., & Bayeck, R. Y. (2016). Exploring the communication preferences of MOOC learners and the value of preference-based groups: Is grouping enough? *Educational Technology Research and Development, 64*(4), 809–837.

Zhang, K. Z., et al. (2018). Online reviews and impulse buying behavior: The role of browsing and impulsiveness. *Internet Research, 28*, 522–543.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.