Forum-based Prediction of Certification in Massive Open Online Courses

Mohammad Alshehri
Department of Computer Science, Durham University, Durham, DH1 3LE, UK. *mohammad.a.alshehri@durham.ac.uk*

Ahmed Alamri
Department of Computer Science, Durham University, Durham, DH1 3LE, UK. *ahmed.s.alamri@durham.ac.uk*

Alexandra I. Cristea
Department of Computer Science, Durham University, Durham, DH1 3LE, UK. *alexandra.i.cristea@durham.ac.uk*

Craig Stewart
Department of Computer Science, Durham University, Durham, DH1 3LE, UK. *craig.d.stewart@durham.ac.uk*

Abstract

Massive Open Online Courses (MOOCs) have been suffering a very level of low course certification (less than 1% of the total number of enrolled students on a given online course opt to purchase its certificate), although MOOC platforms have been offering low-cost knowledge for both learners and content providers. While MOOCs forums generated textual data (forums) have been utilized for the purpose of addressing many MOOCs key challenges like the high rate of dropout and tutor timely intervention, analysing learners' textual interaction for the purpose of predicting certification, remains limited. Thus, this paper investigates *if MOOC learner's comments can predict their purchasing decision (certification)* using a relatively large dataset of 5 MOOCs of 23 runs. Our model achieved promising accuracies, ranging between 0.71 and 0.96 across the five courses. The outcomes of this study are expected to help design future courses and predict the profitability of future runs.

Keywords: MOOCs, Certification Prediction, Learner Analytics.

1. Introduction

For decades, digital learning has been revolutionising and changing the means of modem education. More recently, massive open online courses (MOOCs) platforms were introduced with the goal of reaching a massively infinitive number of prospective students worldwide, specifically to reach a massively unlimited number of potential learners from around the world. Taking into consideration the commercially successful development of Stanford's Coursera in 2011 [1], the contemporary era of e-learning began. Consequently, Several MOOC platforms have been introduced over the next years, especially in 2012, coining 2012 as "the year of the MOOCs" [2]. Since then, Several platforms, such as FutureLearn[1], edX[2], Udemy[3] and Coursera[4] have been introducing free and paid online educational content to the public targeting learners worldwide [3, 4]. As of 2020, there has

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

been a dramatic growth in MOOC platforms which resulted in 16.3 thousand courses introduced by over 950 university partners to more than 180 million learners [5].

The real challenge of MOOCs, *which is the low completion and certification rate*, has been always a concerning issue, nevertheless these learning platforms have been successful in attracting millions of online learners. Dropout is considered a challenging issue with regard to MOOCs, where learners "leak out" at several points during their learning journey. [6, 7]. Literature show that the high dropout rate, exam grades and timely intervention have been the focus of several previous studies [8-10]. However, the efforts for developing accurate prediction models for predicting completion as well as course purchasing continue. Importantly, few studies have investigated the characteristics and temporal activities for the purpose of modelling learners' certification decision behaviours, nevertheless MOOCs have recently gained a considerable attention by the research society.

The literature shows that user purchasing behaviour has been widely studied on pure e-commerce platforms [11, 12]. However, educational domain lacks this kind of analysis as the majority of studies focus on the educational aspect of analysis e.g. analysing learning-related issues, even though MOOC providers have been struggling to build their own sustainable revenues [13].

This paper proposes a forum-based predictor of learners' financial decision (course certificate purchase) taking into account the recent MOOCs transition towards paid courses with affiliate university partners. Specifically, this paper attempts to answer the following research questions:

- **RQ1:** *Do MOOC non-paying learners behave differently to course purchasers as to their forum activities (comments, replies, likes)?*
- **RQ2:** *Can MOOC forum data predict course purchase decisions (certification)?*

While the research questions are relatively attempting to address on issue of certification prediction in the environment of MOOCs, it is essential to distinguish the purpose of each question separately. the first research question utilises learner's forum and interaction to compare the activities of non-paying learners (NL) versus certificate purchasers (CP) using a systematic statistical methodology as shown in section 3.5. Next, the second research question examines whether learners' forum activities can be used to predict later certificate purchasing decisions. This goes beyond comparing samples to employing some state-of-art ML algorithms to predict students' decisions of purchasing a certificate after finishing the course. This kind of prediction is considered essential as the learners' purchasing behaviour is usually taken after the end of the course i.e., after attending the whole course' weekly content.

2. Related Work

Looking through the few studies that investigated MOOC certification, [14] studied the relationship between intention of completion, actual completion, and certificate earning. The study applied on 9 HarvardX MOOCs showed that the correlation between the first two variables was a stronger predictor of certification than any demographic traits. [15] studied MOOC learners' subsequent educational goals after taking the course, by using consumer goal theory. They showed that MOOC completers satisfied with the course delivery were more likely to progress to the course-host institution, than the non-completers. It also 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. It thus became a roadmap for tertiary institutes on how to design an effective MOOC to target potential future students.

Using the only the first week behaviour, [16] predicted MOOC certification via an asset of features. This includes 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. Their Logistic Regression classifier model was trained and tested on one MOOC run only under certain conditions and incentives, by the provider; therefore, it

might need to be replicated, for the results to be generalisable. Qiu et al. [17] 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 features used, i.e. demographics (gender, age, education), 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. [18] 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 the dependent variable – the acquisition of a certificate (true or false).

More recently, [19] 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. Table 1 below summarises the surveyed certification prediction models. Data used included Click Stream (CS), Forum Posts (FP), Assignments (ASSGN), Student Information Systems (SIS), Demographics (DEM) and Surveys (SURV).

Table 1. Certification Prediction Models versus our Model.

Ref.	Data Source	#Courses	#Students	Data Description
[20]	Coursera	1	826	CS; FP
[16]	Coursera	1	37,933	ASSGN; FP; SIS
[14]	HarvardX	9	79,525	DEM; SURV
[21]	edX	1	43,758	CS
[22]	Coursera	1	84,786	FP
[17]	XuetangX	11	88,112	CS
[23]	HarvardX- MITx	10	n/a	CS; FP
[19]	Coursera; edX	1	65,203	CS; FP
Our Model	**FutureLearn**	**5**	**245,255**	**FP**

Unlike previous studies on certification, our proposed model aims to predict the financial decisions of learners on whether to purchase the course certificate. Also, our work is applied to a less frequently studied platform, FutureLearn (Table 1). Another concern we address is study size, with 6 out of the total 9 studies conducted on one course only. As students may behave differently based on the course attended, previous models generalisability is unclear. Instead, we used a variety of courses from different disciplines: Literature, Psychology, Computer Science and Business.

Another key novelty of our study is predicting the learner's real financial decision on buying the course and gaining a certificate. Most course purchase prediction models identify certification as an automatic consecutive step to the completion, making them not different from completion predictors. Our study additionally identifies the most representative factors for certification purchase prediction. It also proposes tree-based and regression classifiers to predict MOOC purchasability using relatively few input features.

3. Methodology

3.1. Data Collection

When a learner joins FutureLearn for a given course, the system generates logs to correlate unique IDs and time stamps to learners, recording learner activities in different datasets:

- Enrolment: contains the learners' demographics along with a mandatory timestamp of enrolment and voluntary timestamps of withdrawal, completion and certification.
- Question Responses: contains the learners' correct and wrong answers.
- Step Activities: contains the learners' access details to each step of the course.
- Comments: contains the learners' comments, replies to other comments along with the number of likes each comment has received [24].

While the first three types of generated data have been analysed in our previous studies "Anonymised for purposes of review", learners' comments seem to be a rich source of data for predicting certification in MOOCs, which is what this paper addresses.

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 "Anonymised for purposes of review". 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]: 6 weeks, and Babies in Mind [BIM]: 4 weeks); Computer Science (Big Data [BD]: 9 weeks) and Business (Supply Chains [SC]: 6 weeks). These courses were delivered repeatedly in consecutive years (2013-2017); thus, we have data on several '*runs'* for each course. Table 2 below shows the number of enrolled, non-paying learners (NL), as well as those having purchased a certificate (CP). Our data shows that students *accessed 3,007,789 materials* in total and declared *2,794,578 steps completed*. Regarding these massive numbers, Table 2 clearly illustrates the low certification rate (less than 1% of the enrolled students).

Table 2. The number of non-paying learners and certificate purchasers on 5 FutureLearn courses.

Course	#Runs	#Weeks	#Non-paying Learners	#Certificate Purchasers
BIM	6	4	48777	676
BD	3	9	33430	268
SP	5	10	63630	750
SC	2	6	5810	71
TMF	7	6	93608	321
Total	**23**	**35**	**245255**	**2086**

3.2. Data Preprocessing

The obtained dataset went through several processing steps, in order to be prepared and fed into the learning model. Since 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, to avoid any mismatch during joining student activities over "several runs" with their current activities.

The pre-processing further contained some standard data manipulations, such as processing (replacing) missing values with zeros, applying *lambda* and *factorize* functions along with Pandas [25] and NumPy [26] 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).

Sentiment Analysis

As sentiment analysis has been an active task while dealing with textual data, classifying learners' sentiment based on their commenting behaviour has played a significant role in predicting certification as shown in the results section. We use the outcomes of sentiment analysis as potential indicators (input features) measuring the learners' numbers of positive/negative/neutral comments or replies on a weekly basis. To achieve this, a well-known Natural Language Processing (NLP) tool called Textblob[5] has been employed, in order to classify students' comments into three categories: *positive*, *neutral* and *negative*.

TextBlob is an NLP-oriented Python library, which measures polarity and subjectivity of a textual dataset for certain tasks, such as sentiment analysis, classification, part-of-speech tagging, extraction and more complex text processing tasks [27]. The tool has been widely used on similar datasets extracted from several sources such as social media platforms. This would help understand learner's expectation and overall satisfaction with the course contents and outcomes. A first analysis of the courses rendered 240,352 positive, 38,743 neutral and 44,242 negative comments, out of which there were 82,355 positive, 20,690 neutral and 18,658 negative replies. Table 3 shows the raw and computed features

[5] https://textblob.readthedocs.io/en/dev/

analysed in this study.

Table 3. The main features utilised for comparing learner activities and predicting course purchasability

Source	Type	Activities (per week)
Comments	Raw	# comments
	Computed	% Positive comments
	Computed	% neutral comments
	Computed	% negative comments
Replies	Raw	# replies
	Computed	% Positive replies
	Computed	% neutral replies
	Computed	% negative replies
Comments & replies	Raw	# likes
	Computed	# character count

As MOOCs are usually delivered on a weekly basis, it was essential to compute the various weekly activities of each learner generating a temporal matrix of their weekly activities. The newly processed Students Activities matrix of each course is as below, keeping in mind the summarised shape due to page width limit:

$$sa = \begin{bmatrix} s_1 & c_{w(1-n)} & pc_{w(1-n)} & \dots & l_{w(1-n)} & wc_{w(1-n)} & cc_{w(1-n)} \\ s_2 & c_{w(1-n)} & pc_{w(1-n)} & \dots & l_{w(1-n)} & wc_{w(1-n)} & cc_{w(1-n)} \\ s_n & c_{w(1-n)} & pc_{w(1-n)} & \dots & l_{w(1-n)} & wc_{w(1-n)} & cc_{w(1-n)} \end{bmatrix}$$

where *s*=student, *c*=comments, *pc*=positive comments, *l*=likes, *wc*=word count, *cc*=character count, *w*= week, *n*=the number of the weeks in a given course.

3.3. Features Selection

Our pre-processed number of features as can be seen in the *sa* matrix above is considerably high due to multiplying the total number of the main extracted features (10) by the total number of weeks *w* in a given course *c*. This resulted in a large array of features, especially for long courses like SP, where the number of weeks was 10, hence generating 100 features. This would on one hand allow for: (1) a temporal fine-grain analysis of the course's content, (2) a timely and early prediction of student's behaviours. However, in order to highlight the most representative features, feature selection techniques were applied, as below. As algorithms employed include tree-based and regression, the features for the tree-based algorithms were selected using Mean Decrease in Impurity (MDI), whereas Variance Inflation Factor (VIF) was used to detect and reduce the multilinearity for the regression algorithms as further explained below [28].

Mean Decrease in Impurity (MDI)

MDI counts the times a feature is used to split a node, weighted by the number of samples it splits. It calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits. MDI is defined as the total decrease in node impurity (weighted by the probability of reaching that node - which is approximated by the proportion of samples reaching that node) averaged over all trees of the ensemble [29].

Variance Inflation Factor (VIF)

Prior to doing regression, multicollinearity among our input features should be taken into consideration. We use VIF (Variable Inflation Factor) to analyse multicollinearity.

$$\boldsymbol{vif_i = \frac{1}{1-R_i^2}} \quad (1)$$

where R_i^2 is the R^2 value obtained by regressing the i^{th} predictor on the remaining predictors. Dropping variables after calculating VIF was an iterative process, starting with the variable having the largest VIF value, as its trend is highly captured by other variables. It was noticed that dropping the highest VIF feature has sequentially reduced the VIF values for the remaining features.

3.4. Statistical Analysis

Normality Test

Our first step of exploring our dataset was examining whether it comes from a specific distribution. The three common procedures of normality verification procedures of: graphical method (Quantile-Quantile plot), numeric method (skewness and kurtosis) and formal normality tests (Shapiro-Wilk) were applied [30]. This has revealed that our data comes from non-Gaussian (normal) distribution and therefore nonparametric tests were conducted as below.

Mann-Whitney U Test

As our data is not normally distributed as well as the variables we are analysing are independent, we used Mann-Whitney U test (Mann–Whitney–Wilcoxon (MWW) [31]), a nonparametric test for testing the statistical significance of the difference of distributions. We use it here to compare the activities of non-paying learners with certificate purchasers.

3.5. Classification Algorithms

Further to the statistical inference, the current study applied four different classification and regression algorithms to predict MOOC learners' purchasing behaviour: Random Forest (RF), ExtraTree (ET), Logistic Regression (LR) and Support Vector Classifier (SVC). These algorithms were chosen due to the fact that they were able to predict course purchasability well, by dealing with massively imbalanced datasets and using at the same time only very few features, as shown in Table 3. These input features exist in any standard MOOC system, which further promotes our model as generalisable. There are some further features that can be utilised for learner behaviour prediction, e.g., demographics or leaving surveys; these features are either not generated by every MOOC platform, or logged later after the end of the course, making early prediction of purchasing behaviour challenging.

To simulate the real-world issue of the low certification rate in MOOCs, we fed the imbalanced data to the classification models as-is. We have initially used many other classification algorithms for this prediction tasks. However, the algorithms that do not deal well with imbalanced data, i.e., have a parameter to define the class weight during learning were excluded.

To deal with our imbalanced dataset, we used the Balanced Accuracy (BA) to report our results, beside the commonly used metric of accuracy (Acc), which is defined as the average of recall obtained on each class [32]. BA equals the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate) as follows:

$$\boldsymbol{ba = \frac{1}{2}\left(\frac{tp}{tp+fn} + \frac{tn}{tn+fn}\right)} \quad (2)$$

Having applied the above preprocessing steps, the shape of X and Y passed to the prediction model was as depicted in Table 4.

Table 1. Number of observations in each class of 0 and 1 by number of selected features

Course	1st Week		1st - Mid Week		All Course	
	Class_0	Class_1	Class_0	Class_1	Class_0	Class_1
BIM	(25508, 5)	(625, 5)	(25508, 13)	(625, 13)	(25508, 25)	(625, 25)
BD	(16010, 7)	(232, 7)	(16010, 25)	(232, 25)	(16010, 36)	(232, 36)
SC	(2840, 6)	(59, 6)	(2840, 19)	(59, 19)	(2840, 32)	(59, 32)
SH	(28920, 10)	(497, 10)	(28920, 32)	(497, 32)	(28920, 47)	(497, 47)
TMF	(39533, 6)	(308, 6)	(39533, 19)	(308, 19)	(39533, 32)	(308, 32)

4. Results and Discussion

The results explore how our processed features can temporally identify course buyers based on their forum activity data. Our temporal analysis showed some statistical significance at various levels when comparing Non-paying Learners and Certificate Purchasers' behaviours across the five courses analysed. Due to the paper limit, we are reporting the most important results here only ordered by the activity categories as shown in Tables 5 – 11, where bold values mean the most significant value in a given course. As the courses analysed spanned over different numbers of weeks, we have selected the first, middle and last weeks to report the results, for fairness of comparison and easy visualisation. For courses with an even number of weeks, we have selected the middle week closer to the end of the course. Our results show that paying learners (CP) were generally more engaged with the course content, in terms of commenting, replying to other's comments, having more sentiment in their textual interaction and receiving more likes on their comments, i.e., being more engaged over their learning journey.

4.1. Number of Comments

Course purchasers have a higher number of comments compared to none-paying learners over the whole courses. This behavioural pattern (the difference between the number of comments of both groups) increases towards the end of the course. the purchasers' weekly number of comments as shown in Table 5 is increasing at different level of significance, but with the last week being the most significant for the majority of the courses.

Table 5. Comparison of the number of Comments for non-paying learners and purchasers at three different time points of the course

Course	M	(NL)			(CP)			*p-value 1st week*	*p-value Mid week*	*p-value Last week*
		1st Week	Mid Week	Last Week	1st Week	Mid Week	Last week			
BIM	μ	1.71	0.75	0.63	3.16	1.94	2.01	*1.0E-35*	*1.2E-47*	***8.5E-68***
	σ	2.81	1.87	1.96	3.73	2.98	3.35			
BD	μ	0.61	0.31	0.36	1.03	0.59	0.84	*2.5E-07*	*8.3E-06*	***9.2E-09***
	σ	1.54	1.01	1.18	2.01	1.29	1.84			
SH	μ	1.56	1.16	1.14	2.70	2.05	2.22	*1.3e-22*	***2.6e-22***	*1.6e-22*
	σ	2.89	2.26	2.22	3.75	2.91	3.22			
SC	μ	1.50	0.91	1.33	2.57	2.01	3.30	*0.008*	*0.001*	***0.000***
	σ	2.83	2.43	3.75	3.89	3.44	5.73			
TMF	μ	1.46	0.88	1.01	1.72	1.16	1.30	*0.049*	***0.027***	*0.134*
	σ	2.64	2.01	2.47	2.86	2.39	3.05			

4.2. Natural Comments

Learners' sentiment over their learning journey is another key factor to distinguish course purchasers from none-paying learners. The students who purchased a certificate at the end of course (CP) have more natural, negative and positive sentiment in their textual inputs (comments or replies) as shown in Table 6 - 8. Contrary to the trend, none-paying learners and course purchasers in BD course have more significant difference (*p-value*) in natural comments in the first week as shown in Table 6.

Table 6. Comparison of the number of Natural Comments for non-paying learners and purchasers at three different time points of the course

Course	M	(NL)			(CP)			*p-value 1st week*	*p-value Mid week*	*p-value Last week*
		1st Week	**Mid Week**	**Last Week**	**1st Week**	**Mid Week**	**Last week**			
BIM	μ	0.10	0.04	0.04	0.19	0.10	0.10	*2.9E-08*	*9.7E-10*	***1.9E-11***
	σ	0.38	0.26	0.28	0.57	0.38	0.41			
BD	μ	0.07	0.04	0.03	0.13	0.05	0.04	***0.055***	*0.151*	*0.083*
	σ	0.37	0.27	0.28	0.56	0.24	0.20			
SH	μ	0.17	0.10	0.13	0.21	0.18	0.22	*0.023*	***1.6e-06***	*2.1e-06*
	σ	0.55	0.37	0.43	0.58	0.47	0.53			
SC	μ	0.28	0.07	0.12	0.45	0.23	0.44	*0.114*	*0.011*	***0.000***
	σ	0.65	0.32	0.44	0.98	0.62	1.00			
TMF	μ	0.10	0.06	0.07	0.09	0.08	0.07	*0.490*	***0.127***	*0.313*
	σ	0.40	0.29	0.34	0.34	0.34	0.34			

4.3. Negative Comments

The significance of difference between none-paying learners and course purchasers varies across the five courses and the three-examining points of time. Additionally, Negative commenting behaviour in TMF course was not different between the two group.

Table 7. Comparison of the number of Negative Comments for non-paying learners and purchasers at three different time points of the course

Course	M	(NL)			(CP)			*p-value 1st week*	*p-value Mid week*	*p-value Last week*
		1st Week	**Mid Week**	**Last Week**	**1st Week**	**Mid Week**	**Last week**			
BIM	μ	0.29	0.09	0.02	0.56	0.22	0.09	***2.0E-26***	*8.7E-17*	*5.7E-14*
	σ	0.69	0.38	0.19	0.85	0.56	0.36			
BD	μ	0.07	0.05	0.03	0.15	0.06	0.06	***0.000***	*0.048*	*0.006*
	σ	0.35	0.26	0.19	0.47	0.27	0.25			
SH	μ	0.10	0.21	0.06	0.16	0.37	0.10	*0.000*	***9.6e-08***	*3.8e-05*
	σ	0.38	0.63	0.29	0.47	0.89	0.34			
SC	μ	0.23	0.09	0.17	0.35	0.18	0.42	*0.053*	*0.025*	***0.003***
	σ	0.66	0.44	0.68	0.76	0.47	1.02			
TMF	μ	0.24	0.16	0.13	0.27	0.17	0.15	*0.218*	*0.382*	***0.187***
	σ	0.65	0.53	0.52	0.73	0.58	0.49			

4.4. Positive Comments

Table 8, in relation to analysing learners' sentiment, compares the positive commenting behaviour between none-paying learner and course purchasers. The difference is more significant towards the end of the course for BIM and BD courses, whereas SH and TMF courses comparisons are more significant in the first week of the course.

Table 8. Comparison of the number of Positive Comments for non-paying learners and purchasers at three different time points of the course

Course	M	(NL)			(CP)			*p-value 1st week*	*p-value Mid week*	*p-value Last week*
		1st Week	**Mid Week**	**Last Week**	**1st Week**	**Mid Week**	**Last week**			
BIM	μ	1.30	0.61	0.56	2.40	1.61	1.81	*3.1E-32*	*4.3E-46*	***2.1E-66***
	σ	2.24	1.54	1.75	3.02	2.51	3.02			
BD	μ	0.46	0.22	0.29	0.75	0.46	0.74	*5.2e-06*	*1.1e-06*	***1.3-e08***
	σ	1.21	0.76	0.98	1.42	1.10	1.67			
SH	μ	1.27	0.84	0.94	2.31	1.49	1.89	***1.4e-23***	*1.9e-20*	*1.0e-19*
	σ	2.49	1.71	1.90	3.35	2.17	2.88			
SC	μ	0.98	0.73	1.03	1.76	1.59	2.44	*0.007*	***0.000***	*0.001*
	σ	2.05	2.02	2.90	2.98	2.85	4.24			
TMF	μ	1.11	0.66	0.80	1.35	0.90	1.07	***0.025***	*0.021*	*0.115*

	σ	2.06	1.55	1.97	2.27	1.94	2.49			

4.5. Number of Likes Received

While the dataset we have does not show who has given a certain like to a given comment or reply, the number of likes received on each comment or reply can be computed. Table 9 shows a comparison based on the number of likes received at three different time points of the course. Contrary to the trend, none-paying learners have a greater number of likes received in the mid-week of BD course.

Table 9. Comparison of the number of Likes Received for non-paying learners and purchasers at three different time points of the course

Course	**M**	**(NL)**			**(CP)**			***p-value 1st week***	***p-value Mid week***	***p-value Last week***
		1st Week	**Mid Week**	**Last Week**	**1st Week**	**Mid Week**	**Last week**			
BIM	μ	2.15	1.26	0.93	4.24	3.52	3.10	*2.1E-33*	*8.3E-42*	***2.8E-56***
	σ	5.21	3.98	3.76	7.34	7.64	7.31			
BD	μ	1.31	0.61	0.72	1.96	1.35	1.28	*6.7e-06*	*7.2e-05*	***5.7-07***
	σ	4.52	2.68	3.02	4.45	4.07	3.72			
SH	μ	2.53	2.17	1.95	3.70	3.59	3.49	***3.6e-15***	*2.1e-12*	*1.3e-13*
	σ	6.73	6.13	5.90	6.75	8.89	8.30			
SC	μ	1.7	0.73	1.06	2.71	1.81	1.74	*0.019*	***0.000***	*0.001*
	σ	4.24	3.02	3.93	5.02	4.28	3.78			
TMF	μ	1.72	1.46	1.65	1.82	1.66	1.92	***0.025***	*0.101*	*0.076*
	σ	5.02	4.29	5.28	4.76	4.69	5.65			

4.6. Character count

The length of comment/reply typed by learner is another indicating factor of course certification at the beginning, middle and end of the course as shown in Table 10. Course purchaser textual inputs tend to be significantly longer across the five courses and the testing points of time.

Table 10. Comparison of the Character Count for non-paying learners and purchasers at three different time points of the course

Course	**M**	**(NL)**			**(CP)**			***p-value 1st week***	***p-value Mid week***	***p-value Last week***
		1st Week	**Mid Week**	**Last Week**	**1st Week**	**Mid Week**	**Last week**			
BIM	μ	529.8	270.2	186.2	1033.9	741.9	624.9	*1.1E-35*	*5.9E-48*	***2.7E-66***
	σ	983.7	748.4	727.8	1407.3	1272.3	1338.7			
BD	μ	216.78	114.76	122.59	405.53	252.34	311.01	*8.9e-08*	*9.3e-06*	***8.6e-09***
	σ	654.13	414.62	486.36	885.68	724.66	851.88			
SH	μ	420.60	392.27	341.44	727.40	732.53	671.93	***5.5e-23***	*9.8e-22*	*6.3e-21*
	σ	949.25	964.51	811.71	1136.18	1348.09	1169.1			
SC	*M*	487.14	339.31	476.88	710.61	520.05	909.72	*0.012*	***0.001***	***0.001***
	σ	1142.1	1191.2	1601.5	1214.2	948.10	1894.8			
TMF	μ	595.14	349.39	430.84	720.55	470.69	573.01	*0.044*	***0.023***	*0.103*
	σ	1223.1	910.2	1256.4	1363.2	1087.3	1584.1			

4.7. Number of Replies Posted

Other exceptions to the overall trend of course purchasers having more engagement with the course are SC and TMF courses in the mid-week and first week respectively. Table 11 shows that in these two scenarios, the number of replies posted by none-paying learners was greater compared to course purchasers.

Table 11. Comparison of the number of Replies Posted for non-paying learners and purchasers at three different time points of the course

Course	M	(NL)			(CP)			*p-value 1st week*	*p-value Mid week*	*p-value Last week*
		1st Week	**Mid Week**	**Last Week**	**1st Week**	**Mid Week**	**Last week**			
BIM	μ	0.40	0.25	0.13	0.68	0.67	0.42	*5.0E-09*	*1.3E-16*	***1.6E-23***
	σ	1.64	1.39	0.82	2.26	2.58	1.56			
BD	μ	0.54	0.28	0.15	0.92	0.40	0.27	*3.4e-05*	***0.000***	*0.024*
	σ	2.67	1.70	1.07	2.80	1.21	1.32			
SH	μ	1.27	0.95	0.81	0.97	1.05	0.85	***0.000***	*0.055*	*0.020*
	σ	10.46	4.89	5.53	3.55	4.81	3.52			
SC	μ	0.31	0.13	0.27	0.66	0.11	0.40	*0.190*	*0.112*	***0.033***
	σ	1.39	0.82	1.22	2.20	0.37	1.05			
TMF	μ	0.85	0.72	0.82	0.78	0.82	0.94	*0.454*	***0.252***	*0.432*
	σ	3.51	4.32	5.10	3.67	3.25	4.57			

4.8. Prediction Performance

The results as shown in Table 12 achieved promising balanced accuracies (BA) across the five domain-varying courses. Keeping numbers of students from Table 2 in mind, it can be seen that there is a fairly inverse relationship between the number of times a course is delivered *#Runs* and the model performance. This suggests that learner activities may be different between runs of the same course, even though the content of each different run of a given course is almost the same - hence generating nosier data for the model to learn. The prediction results indicate that SVC was the best predictor among the tree-based and regression classifiers. Looking at the results from a temporality perspective, the model performance in general improved towards the end of the course. However, even the first-week-only results seems promising with accuracies ranging from 0.70 to 0.93. This qualifies this model to be confidently used for early predicting course purchasing (certification) on these five courses.

Table 12. Learner classification results distributed by course at three different time points of the course, class 0 = non-paying learners, class 1 = paid learners.

course	Classifier	1st Week		1st – Mid-week		All Course	
		BA	**Acc**	**BA**	**Acc**	**BA**	**Acc**
BIM	RF	0.71	**0.78**	0.74	0.86	0.77	0.89
	ET	0.70	0.70	0.72	0.87	0.75	0.90
	LR	0.70	0.73	0.74	**0.88**	0.74	0.90
	SVC	0.69	0.73	0.69	**0.88**	0.68	**0.91**
BD	RF	0.67	0.92	0.68	0.91	0.69	0.92
	ET	0.67	0.92	0.68	0.92	0.68	0.93
	LR	0.68	0.92	0.66	0.92	0.67	0.94
	SVC	0.62	**0.93**	0.53	**0.95**	0.54	**0.94**
CS	RF	0.62	0.83	0.58	**0.93**	0.76	**0.93**
	ET	0.63	0.84	0.58	0.90	0.76	**0.93**
	LR	0.63	0.85	0.58	0.91	0.54	0.91
	SVC	0.60	**0.87**	0.51	**0.93**	0.47	0.92
SP	RF	0.77	0.87	0.75	0.87	0.78	0.89
	ET	0.71	0.79	0.74	0.87	0.76	0.89
	LR	0.72	**0.90**	0.76	0.88	0.76	0.89
	SVC	0.74	0.84	0.64	**0.91**	0.66	**0.93**
TMF	RF	0.65	0.81	0.69	0.90	0.70	0.89
	ET	0.66	0.82	0.68	0.88	0.70	0.90
	LR	0.65	**0.84**	0.67	0.90	0.68	0.92
	SVC	0.61	**0.84**	0.60	**0.92**	0.56	**0.95**

5. Conclusion and Future Work

In this study, we found that students who paid for the course certificate were in general more engaged with the course content and interactive with their peers in terms of commenting, replying to other comments, having more sentiment classes and liking others' comments. We further compared four tree-based and regression classifiers to predict course purchasability based on learners' logged activities. Our proposed model achieved various accuracies, ranging between 0.71 and 0.96. Taking into consideration the real-life challenge of the massively imbalanced classes in MOOCs, our method aimed to solving this issue using the data as-is, without further balancing.

There are few experiments we are planning to conduct in the future. We will investigate the proper early personalisation needed for those who were classified as none-paying learners, and what course-design elements can be reengineered to enrich their learning experience and convince them to purchase the course certificate. This would be a promising research topic, taking into consideration the very low certification rate in MOOCs.

Content analysis are also planned to conduct just for further interpreting some certain results. This specifically will focus on the *"contrary to the trend"* results in order to in-depth understand this variation. One instance is deeply analysing comments where none-paying learners have greater average number of *natural comments* in the mid-week of BD course as shown in Table 6. Such a deeper analysis would infer more understanding of learners' interaction during their learning time.

Acknowledgment

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

References (*style: ReferenceList*)

1. Agarwal, R. The 5 Feature Selection Algorithms every Data Scientist should know. 27/07/2019 [cited 2021 30/03/2021]; Available from: https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2.
2. Alamri, A., et al. Predicting MOOCs dropout using only two easily obtainable features from the first week's activities. in International Conference on Intelligent Tutoring Systems. 2019. Springer.
3. Alshehri, M., et al. 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. 2018. ACM.
4. Alshehri, M., et al., Towards Designing Profitable Courses: Predicting Student Purchasing Behaviour in MOOCs. International Journal of Artificial Intelligence in Education, 2021: p. 1-19.
5. Breslow, L., et al., Studying learning in the worldwide classroom research into edX's first MOOC. Research & Practice in Assessment, 2013. 8: p. 13-25.
6. Castaño-Muñoz, J., et al., Does digital competence and occupational setting influence MOOC participation? Evidence from a cross-course survey. Journal of Computing in Higher Education, 2017. 29(1): p. 28-46.
7. Clow, D. MOOCs and the funnel of participation. in Proceedings of the third international conference on learning analytics and knowledge. 2013. ACM.
8. Coleman, C.A., D.T. Seaton, and I. Chuang. Probabilistic use cases: Discovering behavioral patterns for predicting certification. in Proceedings of the second (2015) acm conference on learning@ scale. 2015.
9. Cristea, A.I., et al. Earliest predictor of dropout in MOOCs: a longitudinal study of FutureLearn courses. 2018. Association for Information Systems.
10. Dellarocas, C. and M.W. Van Alstyne, Money models for MOOCs. Communications of the ACM, August, 2013. 56(8): p. 25-28.
11. developers, s.-l. Metrics and scoring: quantifying the quality of predictions. 2007-2020

[cited 2021 30/03/2021]; Available from: https://scikit-learn.org/stable/modules/model_evaluation.html#balanced-accuracy-score.
12. Gardner, J. and C. Brooks, Student success prediction in MOOCs. User Modeling and User-Adapted Interaction, 2018. 28(2): p. 127-203.
13. Gitinabard, N., et al., Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features. arXiv preprint arXiv:1809.00052, 2018.
14. Hansen, J.D. and J. Reich. Socioeconomic status and MOOC enrollment: enriching demographic information with external datasets. in Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. 2015. ACM.
15. Howarth, J., et al., MOOCs to university: a consumer goal and marketing perspective. Journal of Marketing for Higher Education, 2017. 27(1): p. 144-158.
16. Jiang, S., et al. Predicting MOOC performance with week 1 behavior. in Educational data mining 2014. 2014.
17. Joksimović, S., et al. Translating network position into performance: importance of centrality in different network configurations. in Proceedings of the sixth international conference on learning analytics & knowledge. 2016.
18. Loria, S., textblob Documentation. Release 0.15, 2018. 2.
19. McKinney, W. Data structures for statistical computing in python. in Proceedings of the 9th Python in Science Conference. 2010. Austin, TX.
20. McKnight, P.E. and J. Najab, Mann-Whitney U Test. The Corsini encyclopedia of psychology, 2010: p. 1-1.
21. Ng, A. and J. Widom, Origins of the Modern MOOC (xMOOC). Hrsg. Fiona M. Hollands, Devayani Tirthali: MOOCs: Expectations and Reality: Full Report, 2014: p. 34-47.
22. Oliphant, T.E., A guide to NumPy. Vol. 1. 2006: Trelgol Publishing USA.
23. Perrier, A. Feature Importance in Random Forests. 2015 [cited 2021 30/03/2020]; Available from: https://alexisperrier.com/datascience/2015/08/27/feature-importance-random-forests-gini-accuracy.html.
24. Pursel, B.K., et al., Understanding MOOC students: motivations and behaviours indicative of MOOC completion. Journal of Computer Assisted Learning, 2016. 32(3): p. 202-217.
25. Qiu, J., et al. Modeling and predicting learning behavior in MOOCs. in Proceedings of the ninth ACM international conference on web search and data mining. 2016. ACM.
26. Ramesh, A., et al. Modeling learner engagement in MOOCs using probabilistic soft logic. in NIPS workshop on data driven education. 2013.
27. Razali, N.M. and Y.B. Wah, Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of statistical modeling and analytics, 2011. 2(1): p. 21-33.
28. Reich, J., MOOC completion and retention in the context of student intent. EDUCAUSE Review Online, 2014. 8.
29. Ruipérez-Valiente, J.A., et al. Early prediction and variable importance of certificate accomplishment in a MOOC. in European Conference on Massive Open Online Courses. 2017. Springer.
30. Shah, D. By The Numbers: MOOCs in 2018. 2018.
31. Xu, B. and D. Yang, Motivation classification and grade prediction for MOOCs learners. Computational intelligence and neuroscience, 2016. 2016.
32. Zhang, K.Z., et al., Online reviews and impulse buying behavior: the role of browsing and impulsiveness. Internet Research, 2018.