# Predicting Certification in MOOCs based on Students' Weekly Activities

Mohammad Alshehri[1,2], Ahmed Alamri[1,3] and Alexandra I. Cristea[1]

[1] Department of Computer Science, Durham University, South Road, Durham, DH1 3LE, UK
[2] College of Business, University of Jeddah. Saudi Arabia
[3] College of Computer Science and Engineering. University of Jeddah. Saudi Arabia

[mohammad.a.alshehri,a.s.alamri,alexandra.i.cristea]@durham.ac.uk

**Abstract.** Massive Open Online Courses (MOOCs) have been growing rapidly, offering low-cost knowledge for both learners 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). This can impact seriously the business model of MOOCs. Nevertheless, MOOC research on learners' purchasing behaviour on MOOCs remains limited. Thus, the umbrella question that this work tackles is *if learner's data can predict their purchasing decision (certification)*. Our fine-grained analysis attempts to uncover the latent correlation between learner activities and their decision to purchase. We used a relatively large dataset of 5 courses of 23 runs obtained from the less studied MOOC platform of FutureLearn to: (1) statistically compare the activities of non-paying learners with course purchasers, (2) predict course certification using different classifiers, optimising for this naturally strongly imbalanced dataset. Our results show that learner activities are good predictors of course purchasibility; still, the *main challenge was that of early prediction*. Using only student *number of step accesses, attempts, correct and wrong answers*, our model achieve promising accuracies, ranging between 0.81 and 0.95 across the five courses. 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:** Learner Analytics, MOOCs, Certification Prediction.

## 1 Introduction

Online courses have been revolutionising and reforming education for decades. More recently, massive open online courses (MOOCs) were developed, specifically to reach a massively unlimited number of potential learners from around the world. This modern age of e-learning commenced with the commercially successful introduction of Stanford's Coursera in 2011 [1]. The following year witnessed the launch of many of today's MOOC platforms, coining 2012 as "the year of the MOOCs" [2]. Many

providers, such as *FutureLearn*, *edX*[1], *Udemy*[2] and *Coursera*[3] have started offering scalable online courses to the public, with a diverse set of learning content for learners from all over the world [3, 4]. This has resulted in 16.3 thousand MOOCs delivered via more than 950 university partners to more than 180 million learners by the end of 2020 [5].

Although MOOCs have been successful, attracting many online learners, the staggeringly *low completion and certification rates* is still one of the more concerning aspects to date, a funnel with students "leaking out" at various points along the learning pathway [6, 7]. While the high dropout rate has been the focus of many studies [8-10], 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 investigated the characteristics and temporal activities for the purpose of modelling learners' certification decision behaviours. Concomitantly, the literature shows that user purchasing behaviour has been widely studied on pure e-commerce platforms [11]. 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 [12]. Considering the recent MOOCs' transition towards paid macro-programmes and online degrees with affiliate university partners, this paper presents a fine-grain exploration of student behaviours from a different point of view, non-paying learners versus certificate purchasers. Specifically, this paper attempts to answer the following research questions:

- **RQ1:** *Do MOOC non-paying learners behave differently to course purchasers as to their activities of access and question answering (attempts, correct/wrong answers)?*
- **RQ2:** *Can MOOC learner's logged data predict course purchase decisions (certification)?*

It is worth mentioning that the first research question attempts to compare the activities of non-paying learners (NL) versus certificate purchasers (CP) using a systematic statistical methodology as shown in section 3.5. Subsequently, the second research question examines whether learner activities can be used to predict later certification behaviour. 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 type of prediction seems essential keeping in mind that the certificate purchasing decision is usually taken after the end of the course i.e. after attending the whole course' weekly content.

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

## 2 Related Work

Looking through the few studies that investigated MOOC certification, [13] 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. [14] 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, [15] 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. [16] 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. [17] 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, [18] 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 |
|---|---|---|---|---|
| [19] | Coursera | 1 | 826 | CS; FP |
| [15] | Coursera | 1 | 37,933 | ASSGN; FP; SIS |
| [13] | HarvardX | 9 | 79,525 | DEM; SURV |
| [20] | edX | 1 | 43,758 | CS |
| [21] | Coursera | 1 | 84,786 | FP |
| [16] | XuetangX | 11 | 88,112 | CS |
| [22] | HarvardX- MITx | 10 | n/a | CS; FP |

| | | | | |
|---|---|---|---|---|
| [18] | Coursera; edX | 1 | 65,203 | CS; FP |
| **Our Model** | **FutureLearn** | **9** | **245,255** | **CS; ASSGN; 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 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, such as weekly-based steps visited, completed, comments added, or question attempted [23]. 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 [university name removed]. 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 [24] and NumPy [25] 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). Table 3 shows the main four features analysed in this study.

**Table 3.** The features utilised for comparing student activities and predicting course purchasability

| Activity Source | Activities (per week) |
|---|---|
| Step Access *(a)* | # Accessed steps |
| Attempts *(t)* | # Attempts |
| Correct Answers *(r)* | # Correct Answers |
| Wrong Answers *(f)* | # Wrong Answers |

### 3.3 Feature Extraction

The preliminary data shape is a timestamp log spread on different data frames based on the data log source (access log, question answering log, comments and responses log). 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 follows:

$$sa = \begin{bmatrix} s_1 & a_{w(1-n)} & t_{w(1-n)} & r_{1-n} & f_{1-n} \\ s_2 & a_{w(1-n)} & t_{w(1-n)} & r_{1-n} & f_{1-n} \\ s_n & a_{w(1-n)} & t_{w(1-n)} & r_{1-n} & f_{1-n} \end{bmatrix}$$

where $s$=student, $a$=access, $t$=attempt, $r$=correct answers, $f$=wrong answers, $w$= week, $n$=the number of the weeks in a given course.

### 3.4 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 (4) 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 40 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 [26].

**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 [27].

**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.

$$vif_i = \frac{1}{1 - R_i^2}$$

where $R_i^2$ is the $R^2$ value obtained by regressing the $i^{\text{th}}$ 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.

### 3.5 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 [28]. 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) [29]), 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.

### 3.6 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), also known as the Area Under the ROC Curve, which is defined as the average of recall obtained on each class [30]. BA equals the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate) as follows:

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

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

**Table 4.** Number of observations in each class of 0 and 1 by number of selected features

| Course | Class_0 | Class_1 |
|---|---|---|
| BIM | (25508, 18) | (625, 18) |
| BD | (16010, 30) | (232, 30) |
| SC | (2840, 26) | (59, 26) |
| SH | (28920, 42) | (497, 42) |
| TMF | (39533, 26) | (308, 26) |

## 4 Results and Discussion

The results explore how our processed features can temporally identify course buyers based on their 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 Table 5- Table 8, 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 were generally more engaged with the course content, in terms of accessing the content more frequently, answering more questions correctly and being more socially interactive, i.e. having more comments and responses over their learning journey.

### 4.1 Access

Purchasers seem to have a higher number of accessed steps towards the end of the course. With the SC course as an exception, the purchasers' weekly number of access 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 Access for non-paying learners and purchasers at three different time points of the course.

| | | (NL) | | | (CP) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **C** | **M** | **1st Week** | **MidWeek** | **Last Week** | **1st Week** | **Mid Week** | **Last week** | ***p-value 1st week*** | ***p-value Mid week*** | ***p-value Last week*** |
| BIM | μ | 18.25 | 11.09 | 10.58 | 18.30 | 12.64 | 14.20 | *3.1E-06* | *2.2E-12* | ***8.4E-26*** |
| | σ | 2.09 | 6.16 | 8.35 | 2.30 | 5.27 | 7.38 | | | |
| SH | μ | 15.60 | 11.55 | 13.23 | 15.63 | 11.59 | 14.26 | *3.7E-01* | *2.3E-01* | ***3.3E-08*** |
| | σ | 1.30 | 2.18 | 5.75 | 1.04 | 2.11 | 4.94 | | | |
| TMF | μ | 14.71 | 12.06 | 15.53 | 14.70 | 11.88 | 15.61 | *4.7E-01* | *7.1E-03* | ***6.5E-03*** |
| | σ | 1.62 | 3.11 | 6.65 | 1.62 | 3.16 | 6.92 | | | |
| SC | μ | 18.66 | 16.48 | 21.17 | 18.24 | 17.10 | 21.49 | *2.3E-01* | *3.4E-01* | ***4.8E-01*** |
| | σ | 2.23 | 4.78 | 8.81 | 3.36 | 3.59 | 8.29 | | | |
| BD | μ | 11.72 | 9.32 | 7.83 | 11.73 | 9.61 | 10.03 | *1.8E-01* | *1.2E-02* | ***8.4E-08*** |
| | σ | 1.58 | 3.56 | 6.57 | 1.53 | 3.41 | 6.05 | | | |

### 4.2 Correct Answers

The students who purchased a certificate at the end of course have generally answered more correct answers compared to non-paying learners. Contrary to the trend, TMF has shown different results for both statistical analysis and the number of correct answers.

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

| | | (NL) | | | (CP) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **C** | **M** | **1st Week** | **MidWeek** | **Last Week** | **1st Week** | **Mid Week** | **Last week** | ***p-value 1st week*** | ***p-value Mid week*** | ***p-value Last week*** |
| BIM | μ | 4.70 | 4.13 | 3.34 | 4.76 | 4.12 | 3.54 | *1.9E-02* | *3.7E-15* | ***7.4E-31*** |
| | σ | 1.18 | 1.92 | 2.37 | 1.14 | 1.99 | 2.30 | | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| SH | μ | 11.41 | 11.11 | 9.24 | 11.45 | 11.30 | 9.80 | *1.7E-01* | *1.3E-02* | ***1.2E-03*** |
| | σ | 1.26 | 2.69 | 4.53 | 1.16 | 2.38 | 4.14 | | | |
| TMF | μ | 9.54 | 8.86 | 7.76 | 9.38 | 8.77 | 7.53 | ***3.3E-04*** | *3.5E-01* | *1.1E-0* |
| | σ | 1.40 | 2.54 | 3.82 | 1.51 | 2.65 | 3.94 | | | |
| SC | μ | 4.85 | 4.49 | 4.09 | 4.83 | 4.58 | 4.15 | *4.8E-01* | *3.1E-01* | ***4.6E-01*** |
| | σ | 0.83 | 1.50 | 1.87 | 0.91 | 1.40 | 1.81 | | | |
| BD | μ | 4.58 | 3.39 | 2.02 | 4.67 | 4.01 | 2.93 | *1.8E-01* | *1.3E-05* | ***1.5E-08*** |
| | σ | 1.38 | 2.32 | 2.26 | 1.25 | 2.03 | 2.18 | | | |

### 4.3 Number of Comments

The number of comments posted by learners seem to be the most effective predictor of course purchasability. We can see from Table 7 below that purchasers have commented more that non-paying learners across all weeks and all courses.

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

| | | (NL) | | | (CP) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **C** | **M** | **1st Week** | **Mid Week** | **Last Week** | **1st Week** | **Mid Week** | **Last week** | ***p-value 1st week*** | ***p-value Mid week*** | ***p-value Last week*** |
| BIM | μ | 1.71 | 0.76 | 0.63 | 3.17 | 1.95 | 2.01 | *1.0E-35* | *1.3E-47* | ***8.6E-68*** |
| | σ | 2.81 | 1.87 | 1.97 | 3.74 | 2.98 | 3.35 | | | |
| SH | μ | 1.56 | 1.17 | 1.14 | 2.70 | 2.06 | 2.23 | ***1.3E-22*** | *2.6E-22* | *1.7E-21* |
| | σ | 2.89 | 2.26 | 2.23 | 3.76 | 2.91 | 3.23 | | | |
| TMF | μ | 1.46 | 0.89 | 1.02 | 1.73 | 1.17 | 1.30 | *4.9E-02* | ***2.8E-02*** | *1.3E-01* |
| | σ | 2.65 | 2.01 | 2.47 | 2.86 | 2.39 | 3.06 | | | |
| SC | μ | 1.51 | 0.92 | 1.33 | 2.58 | 2.02 | 3.31 | *8.5E-03* | *1.1E-03* | ***8.9E-04*** |
| | σ | 2.84 | 2.44 | 3.76 | 3.90 | 3.44 | 5.74 | | | |
| BD | μ | 0.62 | 0.32 | 0.36 | 1.03 | 0.59 | 0.84 | *2.6E-07* | *8.4E-06* | ***9.2E-09*** |
| | σ | 1.55 | 1.02 | 1.19 | 2.02 | 1.30 | 1.84 | | | |

### 4.4 Number of Replies

The number of replies posted by both non- and paying learners have similar pattern to the number of comments discussed above. However, non-paying learners in SH and TMF courses have responded more during the first weeks only.

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

| | | (NL) | | | (CP) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **C** | **M** | **1st Week** | **Mid Week** | **Last Week** | **1st Week** | **Mid Week** | **Last week** | ***p-value 1st week*** | ***p-value Mid week*** | ***p-value Last week*** |
| BIM | μ | 0.41 | 0.26 | 0.14 | 0.68 | 0.67 | 0.43 | *5.0E-09* | *1.3E-16* | ***1.6E-23*** |
| | σ | 1.65 | 1.39 | 0.82 | 2.27 | 2.59 | 1.56 | | | |
| SH | μ | 1.28 | 0.95 | 0.82 | 0.98 | 1.05 | 0.85 | ***1.9E-04*** | *5.5E-02* | *2.0E-02* |
| | σ | 10.47 | 4.89 | 5.54 | 3.56 | 4.81 | 3.52 | | | |
| TMF | μ | 0.85 | 0.72 | 0.83 | 0.79 | 0.82 | 0.94 | *4.5E-01* | ***2.5E-01*** | *4.3E-01* |
| | σ | 3.52 | 4.33 | 5.10 | 3.67 | 3.25 | 4.57 | | | |
| SC | μ | 0.32 | 0.14 | 0.27 | 0.66 | 0.12 | 0.41 | *1.9E-01* | *1.1E-01* | ***3.3E-02*** |
| | σ | 1.39 | 0.83 | 1.23 | 2.21 | 0.38 | 1.05 | | | |
| BD | μ | 0.55 | 0.28 | 0.16 | 0.92 | 0.41 | 0.27 | ***3.4E-05*** | *1.4E-04* | *2.5E-02* |
| | σ | 2.68 | 1.70 | 1.07 | 2.80 | 1.21 | 1.33 | | | |

### 4.5 Prediction Performance

The results as shown in Table 9 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 an 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. This may also explain why the CS course, with the lowest number of purchasers, has achieved the highest results on both classes' recalls, compared to the other courses. Class-wise, it is worth mentioning that Recall_1 prediction *our main target, paying students* was greater than Recall_0 over all the five courses.

**Table 9.** Learner classification results distributed by course, class 0 = non-paying learners, class 1 = paid learners.

| Course | Classifier | 1st Week | | | Mid Week | | | Last Week | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Rec_0 | Rec_1 | BA | Rec_0 | Rec_1 | BA | Rec_0 | Rec_1 | BA |
| BIM | RF | 0.61 | 0.95 | 0.78 | 0.79 | 0.85 | 0.82 | 0.80 | 0.85 | 0.83 |
| | ET | 0.60 | 0.95 | 0.77 | 0.80 | 0.82 | 0.81 | 0.81 | 0.82 | 0.81 |
| | LR | 0.60 | 0.95 | 0.78 | 0.78 | 0.86 | 0.82 | 0.80 | 0.86 | 0.83 |
| | SVC | 0.59 | 0.96 | 0.78 | 0.79 | 0.87 | 0.83 | 0.80 | 0.87 | **0.84** |
| BD | RF | 0.78 | 0.96 | 0.87 | 0.87 | 0.86 | 0.86 | 0.87 | 0.95 | **0.91** |
| | ET | 0.76 | 0.98 | 0.87 | 0.85 | 0.90 | 0.88 | 0.86 | 0.95 | 0.91 |
| | LR | 0.76 | 0.98 | 0.87 | 0.86 | 0.88 | 0.87 | 0.86 | 0.95 | 0.91 |
| | SVC | 0.76 | 0.98 | 0.87 | 0.85 | 0.90 | 0.87 | 0.85 | 0.95 | 0.90 |
| CS | RF | 0.78 | 1.00 | 0.89 | 0.90 | 0.90 | 0.90 | 0.90 | 1.00 | **0.95** |
| | ET | 0.78 | 1.00 | 0.89 | 0.89 | 0.90 | 0.89 | 0.89 | 1.00 | 0.95 |
| | LR | 0.78 | 1.00 | 0.89 | 0.90 | 0.90 | 0.90 | 0.90 | 1.00 | **0.95** |
| | SVC | 0.78 | 1.00 | 0.89 | 0.90 | 0.85 | 0.87 | 0.89 | 1.00 | 0.95 |
| SP | RF | 0.55 | 0.98 | 0.77 | 0.79 | 0.96 | 0.87 | 0.84 | 0.91 | 0.87 |
| | ET | 0.55 | 0.98 | 0.77 | 0.79 | 0.96 | 0.88 | 0.84 | 0.92 | **0.88** |
| | LR | 0.58 | 0.95 | 0.76 | 0.84 | 0.90 | 0.87 | 0.84 | 0.90 | 0.87 |
| | SVC | 0.55 | 0.98 | 0.77 | 0.79 | 0.96 | 0.87 | 0.84 | 0.91 | 0.87 |
| TMF | RF | 0.66 | 0.96 | 0.81 | 0.80 | 0.93 | 0.86 | 0.85 | 0.86 | **0.86** |
| | ET | 0.66 | 0.98 | 0.82 | 0.81 | 0.89 | 0.85 | 0.84 | 0.86 | 0.85 |
| | LR | 0.66 | 0.98 | 0.82 | 0.80 | 0.93 | 0.86 | 0.84 | 0.86 | 0.85 |
| | SVC | 0.66 | 0.98 | 0.82 | 0.81 | 0.89 | 0.85 | 0.84 | 0.86 | 0.85 |

## 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. We further compared four tree-based and regression classifiers to predict course purchasability based on learners' logged activities. Our proposed model achieved various balanced accuracies, ranging between 0.81 and 0.95. 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. This is particularly competitive when considering that there could be many other factors influencing the financial decision, such as financial resources, need to document certification, which may have little to do with how students do during the course.

There are few experiments we are planning to conduct in the future. We will investigate the students' sentiments during the course, in order to infer if they correlate with the decision to purchase the certificate. This would be a promising research topic, taking advantage of recent developments in textual data analysis. This could further help in classifying students as early as possible, to provide them with timely intervention and guidance. Another avenue for further research is what to do when students have been categorised, if (and how) to lead them to certification.

**References**


1. 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.
2. Gardner, J. and C. Brooks, *Student success prediction in MOOCs.* User Modeling and User-Adapted Interaction, 2018. **28**(2): p. 127-203.
3. 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.
4. Cristea, A.I., et al. *Earliest predictor of dropout in MOOCs: a longitudinal study of FutureLearn courses*. 2018. Association for Information Systems.
5. Shah, D. *By The Numbers: MOOCs in 2018*. 2018.
6. Clow, D. *MOOCs and the funnel of participation*. in *Proceedings of the third international conference on learning analytics and knowledge*. 2013. ACM.
7. 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.
8. 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.
9. 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.
10. 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.
11. Zhang, K.Z., et al., *Online reviews and impulse buying behavior: the role of browsing and impulsiveness.* Internet Research, 2018.
12. Dellarocas, C. and M.W. Van Alstyne, *Money models for MOOCs.* Communications of the ACM, August, 2013. **56**(8): p. 25-28.
13. Reich, J., *MOOC completion and retention in the context of student intent.* EDUCAUSE Review Online, 2014. **8**.
14. 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.
15. Jiang, S., et al. *Predicting MOOC performance with week 1 behavior*. in *Educational data mining 2014*. 2014.
16. 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.
17. 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.

18. 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.
19. Ramesh, A., et al. *Modeling learner engagement in MOOCs using probabilistic soft logic*. in *NIPS workshop on data driven education*. 2013.
20. 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.
21. 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.
22. Xu, B. and D. Yang, *Motivation classification and grade prediction for MOOCs learners.* Computational intelligence and neuroscience, 2016. **2016**.
23. 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.
24. McKinney, W. *Data structures for statistical computing in python*. in *Proceedings of the 9th Python in Science Conference*. 2010. Austin, TX.
25. Oliphant, T.E., *A guide to NumPy*. Vol. 1. 2006: Trelgol Publishing USA.
26. 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.
27. 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.
28. 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.
29. McKnight, P.E. and J. Najab, *Mann-Whitney U Test.* The Corsini encyclopedia of psychology, 2010: p. 1-1.
30. 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.