Predict and Intervene: Addressing the Dropout Problem in a MOOC-based Program

Inma Borrella
inma@mit.edu
Massachusetts Institute of Technology
Cambridge, Massachusetts

Sergio Caballero-Caballero
sergioac@mit.edu
Massachusetts Institute of Technology
Cambridge, Massachusetts

Eva Ponce-Cueto
eponce@mit.edu
Massachusetts Institute of Technology
Cambridge, Massachusetts

ABSTRACT

Massive Open Online Courses (MOOCs) are an efficient way of delivering knowledge to thousands of learners. However, even among learners who show a clear intention to complete a MOOC, the dropout rate is substantial. This is particularly relevant in the context of MOOC-based educational programs where a funnel of participation can be observed and high dropout rates at early stages of the program significantly reduce the number of learners successfully completing it. In this paper, we propose an approach to identify learners at risk of dropping out from a course, and we design and test an intervention intended to mitigate that risk. We collect course clickstream data from MOOCs of the MITx MicroMasters® in Supply Chain Management program and apply machine learning algorithms to predict potential dropouts. Our final model is able to predict 80% of actual dropouts. Based on these results, we design an intervention aimed to increase learners' motivation and engagement with a MOOC. The intervention consists on sending tailored encouragement emails to at-risk learners, but despite the high email opening rate, it shows no effect in dropout reduction.

KEYWORDS

MOOC, Dropout, Retention, Machine learning, Predictive model, Intervention, Online education, Higher education.

ACM Reference format:
Inma Borrella, Sergio Caballero-Caballero, and Eva Ponce-Cueto. 2019. Predict and Intervene: Addressing the Dropout Problem in a MOOC-based Program. In *Proceedings of Sixth (2019) ACM Conference on Learning @ Scale, Chicago, IL, USA, June 24–25, 2019 (L@S '19),* 9 pages.

1 INTRODUCTION

Open online learning has gone a long way since the first Massive Open Online Course (MOOC) appeared in 2006 [12]. By 2012, MOOCs had become one of the most accessible ways to achieve lifelong learning goals. MOOCs are more affordable and provide more flexibility than in-person courses, and usually grant a certificate upon successful completion. Innovative educational programs consisting of concatenation of MOOCs have emerged, such Coursera's Specializations and edX's MicroMasters®, and are gaining popularity.

MOOC-based educational programs provide deep learning in a specific career field. They include a series of MOOCs which often equate to graduate level courses at top institutions. The value of these programs is increasingly recognized by employers and, in certain cases, they are a path for credit at higher education institutions [27].

MOOC-based programs are changing the way higher education is delivered, and expanding access to credentials from top institutions to underserved populations. However, the high level of dropout associated to MOOCs represents a challenge to their success [15]. A funnel of participation has been identified and characterized for individual MOOCs [9], and this effect is accentuated in MOOC-based programs. High dropout rates in initial courses of the program drain away the number of learners in later courses.

Our main goal is to understand, identify and reduce the dropout rate in MOOC-based educational program. In this paper, we analyze the dropout problem in an edX MicroMasters® program, we develop a predictive model to identify learners at risk of dropping out, and we design and test one intervention intended to increase learners' motivation. The accuracy of the predictive model and the effectiveness of the intervention are also discussed.

The remainder of the paper is organized as follows. In Section 2, we introduce the problem that motivates this research. In Section 3, we review the literature and identify related work. In Section 4, we explain the methodology. Next, in Section 5, we describe the predictive model and the intervention. Results are discussed in Section 6. Finally, we present our conclusions and further research avenues in Section 7.

2 PROBLEM DESCRIPTION

The MITx MicroMasters® credential in Supply Chain Management was announced in Fall 2015 and consists of five MOOCs and a proctored comprehensive final exam. The five graduate level courses include Supply Chain Analytics (SC0x), Supply Chain Fundamentals (SC1x), Supply Chain Design (SC2x), Supply Chain Dynamics (SC3x), and Supply Chain Technology and Systems (SC4x). From now on, we will refer to these five MOOCs as SCx courses.

There are two types of learners in SCx courses: audit and verified. Audit learners enroll in the course and get access to the contents for free. Verified learners pay a fee (US$ 200) that grants them a certificate if they pass (final grade equal or above 60%). Achieving a certificate in all five courses of the MicroMasters® program and passing the Comprehensive Final Exam is the only way of getting the MicroMasters credential.

L@S '19, June 24–25, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6804-9/19/06…$15.00

DOI: 10.1145/3330430.3333634

Since anyone can take SCx courses for free, our assumption is that learners who pay to become verified are demonstrating an intention to complete the course. Surprisingly, we have observed that, on average, 31% of verified learners drop out before the end of the course.

We define dropout occurrence as the moment at which a verified learner stops submitting graded problems in a course. According to this definition of dropout occurrence, any learner who does not complete the final exam has dropped out at some point during the course. Historical data (data from the nine SCx MOOCs run in 2017) reveals that the dropout rate is considerably higher in the first courses of the MicroMasters® program (Figure 1). The dropout rates in SC0x, SC1x, and SC2x range between 32 and 48%, while in SC3x and SC4x are below 22% (Table 1).

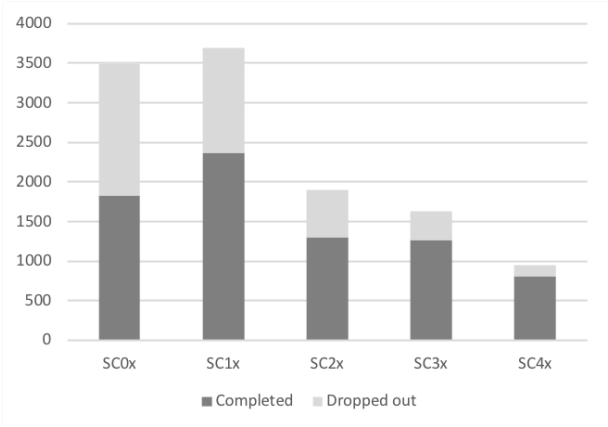

Figure 1: Number of verified learners per course in 2017

This difference could be attributed to many different factors. Here we suggest two potential factors. First, the SC0x, SC1x and SC2x are heavily focused on mathematical models and techniques, which makes them more challenging for many students than SC3x and SC4x, which are more qualitative. Second, the SC0x, SC1x and SC2x usually are the entry point to the MicroMasters®, and while taking them, some people may realize that this is not the educational program they want to pursue. On the other hand, learners taking SC3x and SC4x are already invested in the program and therefore might be less likely to drop out.

High dropout rates at early stages of the program translate into fewer people going down the pipeline, completing the program, and earning a credential.

Table 1: % of verified learners who dropped-out in 2017

Course	Dropout rate	# verified learners
SC0x	48%	3,495
SC1x	37%	3,695
SC2x	32%	1,901
SC3x	22%	1,629
SC4x	16%	995

3 RELATED WORK

Since their origin, MOOCs have been gaining popularity and the offering is still expanding. The delivery and dynamics of a MOOC are very different to those of traditional courses. Early analysis of MOOC enrollment and completion data highlight that it is inadequate to compare completion rates in MOOCs to traditional in-person courses [21]. Understanding the reasons behind MOOCs' low completion rates has attracted a growing interest of academics in the learning analytics area.

Lee and Choi [25] reviewed the literature about online course dropout research and identified relevant factors that influence students decisions to dropout. They classified these factors into three main categories: (1) student factors, (2) course/program factors and (3) environmental factors. Table 2 includes the most relevant factors in each category.

Table 2: Dropout factors in online courses [25]

Categories	Factors
Student	Academic Background
	Relevant Experiences
	Skills
	Psychological attributes
Course/Program	Course Design
	Institutional Support
	Interactions
Environmental	Work commitments
	Supportive Environments

According to Lee and Choi [25] student factors are the dropout factors most frequently studied in the papers they reviewed (55%), followed by environmental factors (25%) and course/program factors (20%). These factors are not independent but influence each other. Some researchers [19, 31] pointed that it is the interaction of numerous factors what leads a student to complete a course or not.

In this paper, we identify students at risk of dropping out by looking at course factors. More specifically, we analyze learners' interaction with the online platform. We monitor their engagement with different learning activities (videos, quick questions, problems,...) and their progress in the course (grades), and use this data to build a predictive model.

We also intend to reduce the dropout rate by influencing student factors. We try to address psychological attributes (motivation) through an intervention that encourages learners to complete an important course activity. Several studies [7, 8, 20] indicate a significant correlation between successfully completing an online course and student's internal motivation. In the following sub-sections we present the most relevant work related to predictive models and interventions in MOOCs to reduce dropout rates.

3.1 Predictive models

Much research has been dedicated to develop models to predict dropout. Existing approaches vary across several dimensions including the data collected, the algorithm used for training and testing, the training set used, and the performance metric.

Regarding the data collected, most of these models use various types of in-course click-stream data. Such data involves records of the students' interactions with course content, including video lectures, discussion forums, assignments, and additional course content. Some researchers favor specific types of course activity such as clicks on videos [4, 16, 18, 24, 26, 29, 36, 37, 41] or forum activity [2, 10, 13, 38, 44]. Some works also look outside the course click-stream data and into demographic information of learners, finding geographic and gender achievement gaps [14, 22].

A variety of machine learning algorithms have been employed to build models to predict learner's dropout. Some of these algorithms include logistic regression [16, 18, 38, 41], decision trees [26, 29, 37], random forests [4], support vector machines (SVM) [4, 24], neural networks [13, 41], and survival analysis [14, 44]. Based on the existing literature, most of these algorithms provide similar predictive power. There is no consensus about a single algorithm outperforming the others. That is why the decision about the algorithm to employ is based on additional criteria rather than merely on predictive performance.

The vast majority of the existing literature train and test the algorithm on the same course [2, 4, 10, 13, 14, 16, 24, 26, 29, 37, 38, 44]. However, some academics argue that this approach can lead to overly optimistic accuracy estimates [18, 41]. They suggest to train the algorithm on data that is collected only after a course has finished and test it in the next run of the same course. For instance, in [22], the authors trained the prediction model on prior course's click-stream to anticipate dropout in an ongoing course.

Most of the authors report the predictive or classification performance of their models based on two well known metrics: accuracy [2, 10, 24, 26, 29, 37] or AUC-ROC (Area Under The Curve - Receiver Operating Characteristics) curve [4, 13, 18, 38, 41]. Very few papers use metrics such as recall or specificity [17] that allow to focus only on the fraction of learners who actually dropped out that was predicted by the model.

Our models use click-stream data as a proxy for learners' activity in the course. As suggested by the literature, we followed a chronological approach, training our models in older course runs and testing them in more recent course runs. We used two very well-known machine learning algorithms: logistic regression and random forests. These algorithms are known for their robustness, ability to identify most relevant predictors, and high predictive power. We tested these algorithms based on their recall values.

3.2 Interventions

In order to reduce the dropout rate in MOOCs, it is increasingly important for the academic community to go beyond predictive models and into the design and implementation of effective interventions.

Identifying the causes of attrition among online learners is very difficult, because dropout is a complicated response to multiple factors [25, 34]. Among the myriad of factors that may influence a student's decision to drop out or persist, psychological attributes are the ones that have received more attention in the literature dedicated to MOOC interventions [25]. And there is an increasing consensus around certain factors that increase persistence: social belonging, motivation, satisfaction, and self-regulation [22, 25, 34].

The efforts to reduce dropout in MOOCs by influencing these internal factors have been manifold, but not all of them have had the desired effect. MOOC interventions are usually implemented as experiments: randomized controlled trials [5, 11, 33, 39, 42, 46], sequential randomized trials [30] or natural field experiments[3, 23, 40, 45]. Some interventions are implemented just once during a course [3, 23, 45, 46], others are repeated at specific times or periodically [5, 11, 30, 40], others are delivered automatically when certain conditions are met [1, 39, 42]. Most interventions are addressed directly to students, and aimed to increase their sense of social belonging, their self-regulation ability and/or their motivation and engagement with the course.

Research has shown that social belonging is a driver of persistence in both in-person and online courses [22, 25, 34]. But increasing the sense of social integration in an online course is challenging because the students work remotely and at their own pace. This difficulty increases even more in MOOCs, with classes consisting of thousands of students from different cultures and time zones. The discussion forum, a space for peer-to-peer and instructor-learners interaction, is the main tool being used by researchers to improve social integration. The discussion forum has been used, for example, to detect and address confusion around a certain topic [1], and to encourage collaboration with peers through small group work [46] and collaborative chats [40].

Self-regulated learning strategies have been studied for years in traditional classroom environments [32, 47], and there is increasing evidence of its importance for online learning settings [6, 28]. In order to improve the self-regulation ability of learners, different interventions have been tested before the course start and during the course. Before the course start, learners are encouraged to make a plan. Two interventions tried to do that with very different results. Yeomans and Reich [45] prompted learners with a plan-making survey at beginning of the course and observed a positive impact on completion and verification. Baker [3] emailed learners and encouraged them to commit a day and time to watch the first lecture of the course, but this intervention had null or even negative effect in some groups of learners. During the course, suggesting SRL strategies [23] does not seem to have any effect on reducing attrition, but introducing social comparison with successful peers in a personalized feedback seems to be a good mechanism to improve SRL ability and reduce dropout [11].

Finally, there are a myriad of interventions trying to increase students' motivation and engagement. Some interventions are focused on encouraging learners by generating greater interest in the course topics [30], or leveraging opportunities to increase their social status [5]. Other interventions focus on struggling learners, and try to reduce their demotivation by proposing mechanisms to cope with their challenges through email [39] or forum communications [1]. Some interventions just try to get absent learners back into the course and/or collect information about their reasons to leave [42].

Our intervention also intends to address learners' lack of motivation at a specific moment in the course. Similarly to many of the papers reviewed, we chose the email as the most appropriate channel to deliver a personalized message directly to each learner. We implemented the intervention as an A/B test experiment and

applied a rigorous statistical analysis to evaluate its effect on learners.

4 METHODOLOGY

The main goal of this research project is to identify MOOC learners at risk of dropping out of the course and take actions to reduce that risk. In this sense, our methodological approach could be divided in two main stages: the "predictive phase" and the "intervention phase" (see Figure 2). In the "predictive phase", we apply machine learning algorithms to determine which learners will drop out from the course. The "intervention phase" targets these potential dropouts deploying a timely intervention intended to persuade them to became more engaged in the course.

In the "predictive phase", we develop a predictive model using course clickstream data and applying two different machine learning algorithms (random forest and logistic regression). The algorithms learn from 42 different predictors. We use the SCx courses to train and test the predictive model. These courses are part of the same program and, despite delivering different contents, they have exactly the same structure, length and grading rules (see Course Description in Section 5.1). This, in addition to a rich amount of historical data, represents an ideal testing ground for building an effective MOOC predictive models.

We train and test our model in a manner that is consistent with how it will be used in practice. We use data from older course runs to train the model and data from more recent courses to test them.

In the "intervention phase", we design an intervention intended to increase motivation and we implement it as an experimental study (A/B test) to measure its impact. We apply the model developed in the "predictive phase" to a current course, and identify the group of learners who are at risk of dropping out. Half of these learners will receive the intervention (treatment group), while the other half will not (control group).

5 EXPERIMENTAL STUDY

For our experimental study, we focused on the first three courses of the MITx MicroMasters® in Supply Chain Management program (SC0x, SC1x, SC2x). In Section 2, we outlined the similar nature of their contents, their large dropout rate, and their importance on overall program retention.

5.1 Course description

Each of these courses is structured in 13 weeks: one intro-to-course week (week 0), 8 content-based weeks (weeks 1 to 4 and 7 to 10), two off-weeks (weeks 5 and 11), and two weeks for midterm (week 6) and final (week 12) exams. Every course week is released on Wednesday at 15:00 UTC.

The concepts of every content-based week are explained in two lessons, each of them provides a series of lecture videos interspersed with short questions called quick questions. Following the lessons, a set of practice problems offers learners the opportunity to reinforce what they just learned. These problem sets provide a space to practice and receive immediate feedback. Verified learners are benefited with additional practice problems and supplemental reading material (e.g., scientific articles, thesis). Finally, a graded assignment evaluates learners' understanding of the lessons taught during the content-based week. The graded assignment is due 14 days after releasing the content-based week material.

Table 3 summarizes the number of videos, quick questions (QQs), practice problems (PPs), and graded assignments (GAs) offered in each course.

Table 3: Course contents

Course	#Videos	#PPs	#QQs	#GAs
SC0x	187	66	109	17
SC1x	178	92	109	17
SC2x	173	64	116	23

5.2 Sample description

The temporal scope of our study was the year 2017. The program offers each course twice a year. The first run (1T) is launched in the first semester, while the second run (2T) starts in the second semester. The course contents do not change, but new Midterm and Final Exams are created for every run.

Figure 3 shows the funnel of completion in these six course runs during 2017. Total enrollment was 137,259. A total of 9,091 learners converted into verified learners, this is our sample. In our sample, 5,509 learners completed the course (took the final exam), and 4,683 passed the course and earned a certificate. The total dropout rate of this sample was 39.4%.

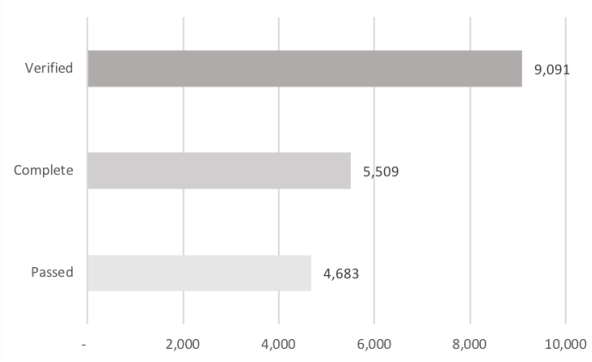

Figure 3: Funnel of completion in selected courses in 2017

Regarding the demographics of our sample, more than 141 nationalities are represented, with the top five countries being: USA (28%), India (9%), Brazil (6%), Egypt (4%), and Spain (3%). The median age is 33 years old, 22% of the learners are female, and the majority of the learners hold a college degree (49%), or a M.Sc. degree (39%).

5.3 Building the predictive model

For our experimental study, we used the first course runs of 2017 (SC0x_1T, SC1x_1T, and SC2x_1T) to train the machine learning algorithms. The second runs of 2017 (SC0x_2T, SC1x_2T, and SC2x_2T) were used to test the models.

The data collected reflects recent and past interaction activity of each individual learner with the platform. As suggested by prior

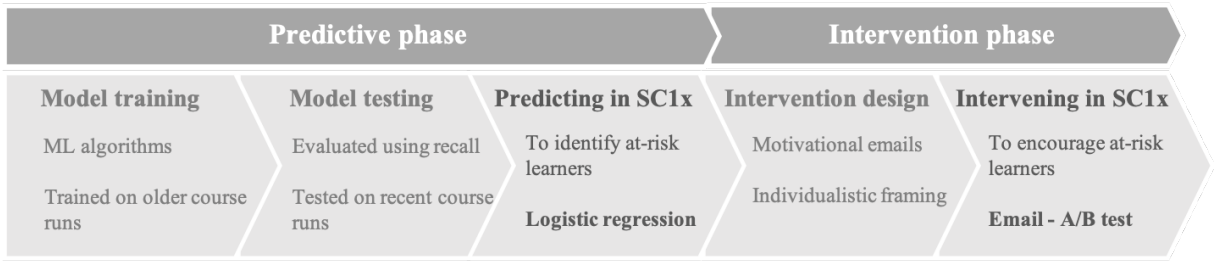

Figure 2: Methodological approach

work [16, 41], we focus on variables that suggest learner's lack of ability or lack of interest or time. These features demonstrated to correlate strongly with dropout. Among the variables that imply student's ability or lack of it are grade scores (grade to date and grade achieved in week). Variables that suggest lack of interest or time are represented by the time the learner spent in the platform, the number of clicks in different content material, and time elapsed since last log-in. The data collected could be grouped into five main categories, as shown in Table 4.

The first category, enrollment data, provides information about when the student enrolled and verified in the course. Enrollment variables are measured in days and are calculated based on the date in which the course was launched. The second category, grade data, gives information about graded activity such as number of missed assignments, grade achieved to date, and grade obtained in current week. The third category, time data, includes time elapsed since the learner's last click in the course, and the time spent in the course during the past seven days. The fourth category, click in time frames, provides the total number of clicks up to date, and the number of clicks in the course and discussion forums during the last seven days. The last category, clicks in content weeks, includes similar information but associated with lecture videos, quick questions, practice problems, and graded assignment of specific lessons (content-based weeks).

The database can be updated on a weekly basis. The most recent information is included (between $t-1$ and t) and information about last 3 weeks is maintained (M_1, M_2 and M_3). Figure 4 shows a schematic view of the first part of the course. In this example, the model is updated after closing content week 2's (lecture 2) graded assignment ($t = 2$). Grade, time, and clicks information about the most recent week (from $t = 1$ to $t = 2$) will be incorporated in the model and similar information about previous weeks kept in the database.

5.4 Designing the intervention

The main goal of our intervention was to increase learners' motivation before the Midterm Exam. The Midterm Exam is an important milestone in SCx courses. It happens halfway through the course and accounts for 35% of the final grade. Therefore, passing the course (grade $\geq 60\%$) becomes very difficult for someone who misses the Midterm. In the six courses of our sample, on average, 70% of the dropouts within a course abandoned during the first half of the course (meaning that they never completed the Midterm Exam). In some cases, this might be motivated by external factors,

Table 4: Predictive variables collected in each week

Feature description	Time interval
Enrollment	
Enrollment time	
Verification time	
Grade	
Number of missed assignments	
Grade to date	
Grade achieved in week	t, M_1, M_2, M_3
Time	
Time elapsed since last click	
Time spent in the course during 7 days	t, M_1, M_2, M_3
Clicks in time frames	
Clicks to date	
Clicks in the course during 7 days	t, M_1, M_2, M_3
Clicks in the forums during 7 days	t, M_1, M_2, M_3
Clicks in content weeks	
Clicks in graded assignment of a certain week	
Clicks in lecture videos of a certain week	t, M_1, M_2, M_3, P_1
Clicks in practice problems of a certain week	t, M_1, M_2, M_3, P_1
Clicks in quick questions of a certain week	t, M_1, M_2, M_3, P_1

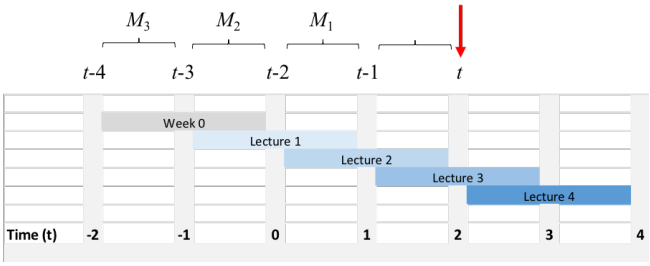

Figure 4: Schematic view of data scheme

but in other cases internal factors such as lack of motivation or self-confidence may play a role. Our intervention intended to encourage learners to take the Midterm Exam.

We identify learners at risk of dropping out at the Midterm Exam (not taking the exam) applying our predictive model. The at-risk learners are divided in two groups, the test group and the control group, and an encouragement email is sent to each learner in the control group. The email is personally addressed to the each learner

(using his/her name), it is signed by the course lead and the director of the program, and it contains a link to the Midterm Exam.

The content of the email is tailored to each learner depending on how many graded assignments he/she has completed at the date of the intervention. Therefore, our at-risk learners are divided in three subgroups. The learners who have not completed any graded assignments are assigned to Group 0; the ones who have completed one graded assignment are included in Group 1; and the rest (2 or 3 graded assignments completed) are part of Group 2.

In all three messages, we use an individualistic framing, because the study by Davis et al. [11] suggested that it is more successful in MOOC interventions than a collectivist framing. From an expectancy-value theory perspective [43], the messages for Group 0 and Group 1 focused on outcome expectancies. The content of these messages highlight that the learner can still catch up with the course and get the certificate, if he/she changes current behaviour and take the Midterm Exam. For Group 2, the message focuses on recognizing the extrinsic value of completing the course and achieving a certificate. The content of the three emails can be seen in Table 5.

Table 5: Overview of emails content

Group	Content
G0	We know you are interested in SC1x, but we haven't seen you around in the course much. Can you tell us what is holding you back and how we can help you? [Big blue button with link to an open response survey] You can still catch up with the course, you know? The Midterm will open on February 7 and is worth 35% of the final grade, so give it a shot. We know you can make it!
G1	We know you have missed some Graded Assignments, but don't worry, it is not a big deal! You can still catch up and get your SC1x certificate. Take a shot at the Midterm Exam! It is worth 35% of the final grade. We know you can make it!
G2	We can see that you are working hard on SC1x. Sometimes it may be challenging, but it will be worth it! You are learning a lot and the SC1x certificate will be useful in your career. Good luck in the Midterm Exam, we know you can make it!

6 RESULTS AND DISCUSSION

6.1 Training and testing the model

We applied two machine learning algorithms to build our predictive models: random forest and logistic regression. These algorithms aim to predict if a learner will drop out the course based on her or his activity and performance in the course (see Table 4).

The predictive models can be run on any given week. But in this study we have focused on two particular time periods. Due to their significant weight in the final grade, the Midterm and Final Exams are key moments to determine who would drop out the course. Therefore, we have evaluated our predictive models according to how well they can predict who will drop out the exam one week before the exam becomes available in the platform.

Our models were evaluated based on a well known metric: recall. The recall (or sensitivity) of a classifier measures its ability to detect the important class member correctly [35]. In our case, recall indicates the proportion of total learners who actually dropped out that was correctly predicted by the model. We focus on recall because we want the algorithm to minimize false negatives (actual dropouts who are not predicted as dropouts). Moreover, there is low cost (risk) of intervening (sending emails to) learners who are false positives (learners who actually did not drop out but are predicted as dropouts). The recall was computed for both training and test data sets.

Table 6 shows the recall and precision values obtained to predict learners who would skip the Midterm or Final Exam. Recall values are pretty consistent among algorithms. However, these values show some differences among courses, specially if we compare SC0x and SC1x with SC2x.

In general, it is more challenging to identify learners who skip the Final Exam. On average, 86% of the learners who actually skipped the Midterm Exam were identified by the model, compared with the 76% in the case of the Final Exam. This might be explained by the fact that to predict who would skip the Midterm Exam, the model considered all verified learners, including those with very little activity in the course. This group of learners are very likely to skip the Midterm Exam and consequently can be easily detected by the model. In the case of the Final Exam, to predict who will become a dropout, the model only considers learners who have taken the Midterm Exam.

The most relevant factors to predict if learners would skip the Midterm Exam are *Grade to date, Number of missed assignments, Clicks in graded assignment,* and *Time spent in the course during the last 7 days*. Similarly, the most relevant factors to predict Final Exam participation are *Grade to date, Grade in midterm, Number of missed assignments, Clicks in graded assignment, Time spent in the course during the last 7 days, Time since last click,* and *Clicks in lecture videos*.

In summary, both machine learning algorithms provided similar results in terms of predicting potential dropouts and identifying the most relevant drivers of dropping out, as already highlighted by prior research. The rest of the paper will show the results of applying logistic regression to predict dropouts in a current SCx course. We decided to move forward and use this algorithm because it is easier to understand how it works and interpret how the dropouts are being predicted.

6.2 Predicting in SC1x_1T 2018

We chose the first run of SC1x in 2018 to implement our experiment. The total number of verified learners in the course was 1,506. We used logistic regression to identify the learners at risk of not taking the Midterm Exam. The previous course run of SC1x (2T 2017) was used for training purposes.

The predictive model identified 365 learners who were likely to skip the Midterm Exam. These at-risk learners were the target of the intervention. Once the Midterm Exam was closed, we could evaluate the effectiveness of the predictive model. Out of the 405

Table 6: Midterm and final exam recall and precision values for 2017 SCx courses

Course	Exam	Algorithm	Recall	Precision
SC0x	Midterm	Logistic regression	0.88	0.88
		Random forest	0.90	0.86
	Final	Logistic regression	0.73	0.86
		Random forest	0.81	0.83
SC1x	Midterm	Logistic regression	0.78	0.85
		Random forest	0.79	0.87
	Final	Logistic regression	0.88	0.80
		Random forest	0.88	0.82
SC2x	Midterm	Logistic regression	0.91	0.93
		Random forest	0.90	0.95
	Final	Logistic regression	0.62	0.69
		Random forest	0.64	0.88

learners who skipped the test, 311 learners were predicted by the model (77%). More details can be seen in the confusion matrix shown in Table 7.

Table 7: Confusion matrix for prediction in SC1xT 2018

		Predicted		
		Take exam	Skip exam	Total
Actual	Take exam	1047	54	1101
	Skip exam	94	311	405
	Total	1141	365	1506

6.3 Intervening in SC1x_1T 2018

The 365 at-risk learners identified by the predictive model became the target of our intervention. 24 of these at-risk learners were unenrolled from the course at the date of the intervention, so our final target group was 341.

We divided our at-risk learners in subgroups, according to the number of graded assignments submitted. At the time of the intervention, a maximum of 3 graded assignments could have been completed. This classification resulted in the three subgroups shown in Table 8:

We conducted a A/B test experiment. From each group, 50% of the learners were randomly selected as receivers of the intervention, keeping the other 50% as the control group (see Table 8). Encouragement emails (see Table 5) were sent to the selected receivers exactly one week before the Midterm Exam. The mails were sent through our marketing platform, MailChimp. However, not all the mails sent were delivered, because some of the learners had previously unsubscribed from our mailing list. This resulted in a control group of 187 learners who did not receive our email communication, and a treatment group of 154 learners who received an email. These mails had an overall opening rate of 62%.

After the Midterm Exam closed, we could evaluate the impact of the intervention. Of the 154 learners who were the target of the intervention, 33 took the Midterm Exam (22%); while among the 187 who did not receive any email, 46 took the Midterm Exam (25%). This results indicate that the intervention did not have the intended effect of reducing the dropout at the Midterm Exam (see Table 9).

Table 8: Emails sent and delivered.

Group	# GA submitted	# learners	# mails sent	# mails delivered	Opening rate
G0	0	211	105	96	63%
G1	1	97	48	43	56%
G2	2 or 3	33	16	15	67%
All	any	341	169	154	62%

Considering these results, a Chi-Square test of independence was performed to examine the relation between receiving an intervention email and taking the Midterm Exam. The relation between these variables was not significant, $p-$value = 0.86 (>0.05). Therefore, we can conclude that the email intervention did not have any impact on the likelihood of learners taking or not the Midterm Exam.

Table 9: Results of intervention experiment

Test group	# learners	# took Midterm	% took Midterm
Intervention group	154	33	21%
Control group	187	46	25%

While these email interventions did not have the impact that we expected, we did not analyze other potential effects on learners' performance (grade) or engagement (activity in the course).

Twenty learners from Group 0 shared information with us through the open response survey that we included in the body of the email. They thanked us for the encouragement, apologized for their poor performance and/or explained their lack of engagement with the course. In most of their messages, the learners argued that some external impact have prevented them from dedicating enough time to the course. For example, two learners reported health issues, twelve learners reported increasing work responsibilities and travels, and two learners reported family events. All these learners suggested that more flexibility with deadlines and allowing some extra time would help them. Only three of the learners reported to be struggling with the contents and requested some more guidance.

This information might be an indicator that the reason behind learners dropout in SCx courses is not a lack of motivation but a lack of time. This is consistent with the findings of [22], who concluded that the major reason for attrition among the online learners in their study was having not enough time.

7 CONCLUSIONS AND FUTURE WORK

This research explores dropout on a MOOC-based program, and argues that predictive and intervention efforts should be combined to address the dropout problem.

A predictive model was proposed based on machine learning algorithms fed with course click-stream data. This model was trained in past courses data and tested in most recent courses data. This training-testing approach replicates the way the models will be used in a real course setting, and avoids the overoptimistic results that often appear when training and testing the model in data extracted from the same course. The measure selected to report the predictive power of the model was recall, a measure that indicates how many of the actual dropouts were rightly predicted by the model. Our model was able to predict four out of five actual dropouts in the courses that were part of our sample.

An intervention intended to reduce the dropout rate halfway through the course (before the Midterm Exam) was designed and tested in one of the courses. The target group of learners was identified by applying the predictive model described above. The intervention consisted of personalized emails that contained an encouragement message. The intervention was implemented as an A/B test experiment. Sixty-two per cent of the learners who received the intervention opened the email, and some of them provided information about why they were falling behind in the course. But the statistical analysis demonstrated that the mail intervention had no effect on reducing the dropout rate associated with the Midterm Exam.

The ineffectiveness of our email intervention and the insights from our learners' feedback suggest that supporting self-regulation and providing mechanisms to cope with unexpected life events (e.g. allowing more flexibility with graded assignments) might be more effective than motivational messages to reduce the dropout rate in MOOCs.

As future developments, we plan to introduce other variables in our predictive model such as demographic data and program-related data, which might include: number of previous SCx courses enrolled, average activity, and performance in those courses. Additionally, we will further analyze dropout learners' clickstream data to identify patterns, and collect survey data to understand the main drivers of dropout.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support of Dr. Chris Caplice, Director of the MITx MicroMasters Program in SCM, to this project. This work benefited from his feedback and also the feedback provided by three anonymous reviewers. Special thanks to Connor Makowski who helped with the data extraction. Finally, thanks to MIT Integrated Learning Initiative for funding this project.

REFERENCES

[1] Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke. 2015. YouEDU: Addressing Confusion in MOOC Discussion Forums by Recommending Instructional Video Clips. (2015), 8 pages.

[2] U Andersson, T Arvemo, and M Gellerstedt. 2016. How well can completion of online courses be predicted using binary logistic regression?. In *IRIS39-The 39th Information Systems Research Conference in Scandinavia*. Ljungskile, Sweden.

[3] Rachel Baker, Brent Evans, and Thomas Dee. 2016. A Randomized Experiment Testing the Efficacy of a Scheduling Nudge in a Massive Open Online Course (MOOC). *AERA Open* 2, 4 (oct 2016), 1–18.

[4] Miguel L Bote-Lorenzo and Eduardo Gómez-Sánchez. 2017. Predicting the decrease of engagement indicators in a MOOC. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK'17)*. Vancouver, Canada, 143–147.

[5] Katherine Brady, Douglas Fisher, and Gayathri Narasimham. 2016. Exploring the Effects of Lightweight Social Incentives on Learner Performance in MOOCs. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*. ACM Press, Edinburgh, UK, 297–300.

[6] J Broadbent and W L Poon. 2015. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *Internet and Higher Education* 27 (2015), 1–13.

[7] Jane Castles. 2004. Persistence and the adult learner: factors affecting persistence in Open University students. *Active learning in higher education* 5, 2 (2004), 166–179.

[8] Seung Youn Chyung. 2001. Systematic and systemic approaches to reducing attrition rates in online higher education. *American Journal of Distance Education* 15, 3 (2001), 36–49.

[9] Doug Clow. 2013. MOOCs and the funnel of participation. In *Proceedings of the 3rd Conference on Learning Analytics and Knowledge (LAK 2013)*. Leuven, Belgium, 185–189.

[10] Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S. McNamara, and Ryan S. Baker. 2016. Combining click-stream data with NLP tools to better understand MOOC completion. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* (2016), 6–14.

[11] Dan Davis, Ioana Jivet, René F Kizilcec, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. 2017. Follow the Successful Crowd: Raising MOOC Completion Rates through Social Comparison at Scale *. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK'17)*. Vancouver, Canada, 454–463.

[12] Stephen Downes. 2008. Places to Go: Connectivism & Connective Knowledge Recommended APA Citation. *Journal of Online Education* 5, 1 (2008), Article 6.

[13] Mi Fei and Dit-Yan Yeung. 2015. Temporal Models for Predicting Student Dropout in Massive Open Online Courses. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 256–263.

[14] J. A. Greene, C. A. Oswald, and J. Pomerantz. 2015. Predictors of Retention and Achievement in a Massive Open Online Course. *American Educational Research Journal* 52, 5 (2015), 925–955.

[15] Christian Gütl, Rocael Hernández Rizzardini, Vanessa Chang, and Miguel Morales. 2014. Attrition in MOOC: Lessons Learned from Drop-Out Students. *Communications in Computer and Information Science* 446 CCIS (2014), 37–48.

[16] Sherif Halawa, Daniel Greene, and John Mitchell. 2014. Dropout Prediction in MOOCs using Learner Activity Features. *eLearning Papers* 37, March (2014), 1–10.

[17] Sherif Halawa, Daniel Greene, and John Mitchell. 2014. Dropout Prediction in MOOCs using Learner Activity Features. *eLearning Papers* 37 (2014), 1–10.

[18] Jiazhen He, James Bailey, Benjamin I P Rubinstein, and Rui Zhang. 2015. Identifying At-Risk Students in Massive Open Online Courses. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence Identifying*.

[19] Bruce Holder. 2007. An investigation of hope, academics, environment, and motivation as predictors of persistence in higher education online programs. *The Internet and higher education* 10, 4 (2007), 245–260.

[20] Nataliya V Ivankova and Sheldon L Stick. 2007. StudentsâĂŹ persistence in a distributed doctoral program in educational leadership in higher education: A mixed methods study. *Research in Higher Education* 48, 1 (2007), 93.

[21] K. Jordan. 2014. Initial trends in enrolment and completion of massive open online courses Massive Open Online Courses. *International Review of Research in Open and Distance Learning* 15, 1 (2014), 133–160.

[22] René F Kizilcec and Sherif Halawa. 2015. Attrition and Achievement Gaps in Online Learning. In *Proceedings of the Second ACM Conference on Learning@ Scale*.

[23] René F. Kizilcec, Mar Pérez-Sanagustín, and Jorge J. Maldonado. 2016. Recommending Self-Regulated Learning Strategies Does Not Improve Performance in a MOOC. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*. ACM Press, New York, New York, USA, 101–104.

[24] Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. 2014. Predicting MOOC Dropout over Weeks Using Machine Learning Methods. In *Proceedings ofthe 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 60–65.

[25] Youngju Lee and Jaeho Choi. 2011. A review of online course dropout research: implications for practice and future research. *Educational Technology Research and Development* 59, 5 (oct 2011), 593–618.

[26] Jiajun Liang, Chao Li, and Li Zheng. 2016. Machine learning application in MOOCs: Dropout prediction. In *ICCSE 2016 - 11th International Conference on Computer Science and Education*. IEEE, 52–57.

[27] Joshua Littenberg-tobias and Justin Reich. 2018. âĂIJ Yeah , I know this âĂİ: Student Experiences in a Blended MicroMasters. (2018), 1–3.

[28] Allison Littlejohn, Nina Hood, Colin Milligan, and Paige Mustain. 2016. Learning in MOOCs: Motivations and self-regulated learning in MOOCs. *The Internet and Higher Education* 29 (apr 2016), 40–48.

[29] Saurabh Nagrecha, John Z. Dillon, and Nitesh V. Chawla. 2017. MOOC Dropout Prediction. In *Proceedings of the 26th International Conference on World Wide*

Web Companion - WWW '17 Companion. ACM Press, New York, New York, USA, 351–359.

[30] Timothy NeCamp, Josh Gardner, and Christopher Brooks. 2018. Beyond A/B Testing: Sequential Randomization for Developing Interventions in Scaled Digital Learning Environments. *arXiv preprint arXiv:1810.11185* (oct 2018), 14. arXiv:1810.11185

[31] Beth Perry, Jeanette Boman, W Dean Care, Margaret Edwards, and Caroline Park. 2008. Why Do Students Withdraw from Online Graduate Nursing and Health Studies Education?. *Journal of Educators Online* 5, 1 (2008), n1.

[32] Paul R Pintrich. 1999. The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research* 31 (1999), 459–470.

[33] Jan Renz, Daniel Hoffmann, Thomas Staubitz, and Christoph Meinel. 2016. Using A/B Testing in MOOC Environments. In *Proceedings of the 6th International Learning Analytics & Knowledge Conference*. Edinburgh, UK, 304–313.

[34] Alfred P Rovai. 2003. In search of higher persistence rates in distance education online programs. *The Internet and Higher Education* 6, 1 (2003), 1–16.

[35] Peter C Shmueli, Galit; Patel, Nitin R; Bruce. 2010. *Data Mining for Business Intelligence*. Wiley. arXiv:arXiv:1011.1669v3

[36] Tanmay Sinha, Patrick Jermann, Nan Li, and Pierre Dillenbourg. 2014. Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. (2014). arXiv:arXiv:1407.7131v2

[37] Jeff K. Tang, Haoran. Xie, and Tak-Lam Wong. 2015. A Big Data Framework for Early Identification of Dropout Students in MOOC. In *In International Conference on Technology in Education*. 127–132.

[38] Colin Taylor, Kalyan Veeramachaneni, and Una-May O 'reilly. 2014. Likely to stop? Predicting Stopout in Massive Open Online Courses. (2014).

[39] Ralf Teusner, Thomas Hille, and Thomas Staubitz. 2018. Effects of automated interventions in programming assignments. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale - L@S '18*. ACM Press, New York, New York, USA, 1–10.

[40] Gaurav Singh Tomar, Sreecharan Sankaranarayanan, Xu Wang, and Carolyn Penstein Rosé. 2017. Coordinating Collaborative Chat in Massive Open Online Courses. *arXiv preprint arXiv:1704.05543*. (apr 2017). arXiv:1704.05543

[41] Jacob Whitehill, Kiran Mohan, Daniel Seaton, Yigal Rosen, and Dustin Tingley. 2017. Delving Deeper into MOOC Student Dropout Prediction. (2017).

[42] Jacob Whitehill, Joseph Williams, Glenn Lopez, Cody Coleman, and Justin Reich. 2015. Beyond Prediction : First Steps Toward Automatic Intervention in MOOC Student Stopout. *Proceedings of the 8th International Conference on Educational Data Mining* (2015), 171–178.

[43] Allan Wigfield and Jacquelynne S Eccles. 2000. Expectancy-value theory of achievement motivation. *Contemporary educational psychology* 25, 1 (2000), 68–81.

[44] Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rose. 2013. Turn on, Tune in, Drop out: Anticipating student dropouts in Massive Open Online Courses. In *Proceedings of the 2013 NIPS Data-driven education Workshop*. Vol.11, p.14.

[45] Michael Yeomans and Justin Reich. 2017. Planning Prompts Increase and Forecast Course Completion in Massive Open Online Courses. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK'17)*. Vancouver, Canada, 464–473.

[46] Zhilin Zheng, Tim Vogelsang, and Niels Pinkwart. 2015. The Impact of Small Learning Group Composition on Student Engagement and Success in a MOOC. In *Proceedings of the 8th International Conference on Educational Data Mining*. 500–503.

[47] Barry J Zimmerman. 1990. Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist* 25, 1 (1990), 3–17.