Maldonado-Mahauad, Jorge; Pérez-Sanagustín, Mar; Moreno-Marcos, Pedro Manuel; Alario-Hoyos, Carlos ; Muñoz-Merino, Pedro J.; Delgado-Kloos, Carlos. Predicting Learners' Success in a Self-Paced MOOC through Sequence Patterns of Self-Regulated Learning. In: Pammer-Schindler V., et al. (eds.). (2018) *Lifelong Technology Enhanced Learning: 13th European Conference on Technology Enhanced Learning, EC-TEL 2018, Leeds, UK, September 3-5, 2018: proceedings* (pp. 355-369). (Lecture Notes in Computer Science, 11082).

DOI: https://doi.org/10.1007/978-3-319-98572-5_27

Predicting Learners' Success in a Self-Paced MOOC through Sequence Patterns of Self-Regulated Learning

Jorge Maldonado-Mahauad[1,3], Mar Pérez-Sanagustín[1], Pedro Manuel Moreno-Marcos[2], Carlos Alario-Hoyos[2], Pedro J. Muñoz-Merino[2], Carlos Delgado-Kloos[2]

[1] Department of Computer Science, Pontificia Universidad Católica de Chile, Santiago, Chile
[2] Department of Telematics Engineering, Universidad Carlos III de Madrid, Madrid, Spain
[3] Department of Computer Science, University of Cuenca, Cuenca, Ecuador

{jjmaldonado, mar.perez}@uc.cl;{pemoreno, calario, pedmume, cdk}@it.uc3m.es

Abstract. In the past years, predictive models in Massive Open Online Courses (MOOCs) have focused on forecasting learners' success through their grades. The prediction of these grades is useful to identify problems that might lead to dropouts. However, most models in prior work predict categorical and continuous variables using low-level data. This paper contributes to extend current predictive models in the literature by considering coarse-grained variables related to Self-Regulated Learning (SRL). That is, using learners' self-reported SRL strategies and MOOC activity sequence patterns as predictors. Lineal and logistic regression modelling were used as a first approach of prediction with data collected from N = 2,035 learners who took a self-paced MOOC in Coursera. We identified two groups of learners: (1) Comprehensive, who follow the course path designed by the teacher; and (2) Targeting, who seek for the information required to pass assessments. For both type of learners, we found a group of variables as the most predictive: (1) the self-reported SRL strategies 'goal setting', 'strategic planning', 'elaboration' and 'help seeking'; (2) the activity sequences patterns 'only assessment', 'complete a video-lecture and try an assessment', 'explore the content' and ' try an assessment followed by a video-lecture'; and (3) learners' prior experience, together with the self-reported interest in course assessments, and the number of active days and time spent in the platform. These results show how to predict with more accuracy when students reach a certain status taking in to consideration not only low-level data, but complex data such as their SRL strategies.

Keywords: Self-Regulated Learning, Prediction, Massive Open Online Courses, Sequence Patterns, Achievement, Success.

1 Introduction

The massive and open nature of Massive Open Online Courses (MOOCs) contribute to attract a great diversity of learners, who have seen in MOOCs an opportunity for their personal growth. Most of the learners who enroll in a MOOC decide which parts of the course content they choose to engage with, and eventually only a small proportion of these enrollees complete the course (typically less than the 10%) [8]. This has aroused the interest on studying the causes why learners complete or drop out a MOOC.

Prior research shows that self-regulation is one of the critical skills needed to achieve personal learning goals in a MOOC [20]. Self-regulated learners are characterized by their ability to initiate cognitive, metacognitive, affective and motivational processes [4]. Moreover, recent research in self-regulated Learning (SRL) suggests that successful learning and academic achievement are associated with the deployment of regulatory activities such as goal-setting, planning or monitoring [2].

MOOC enrollees present a diversity of behaviours depending on: learner's previous knowledge, prior experience, intentions and motivations [18, 25]. In a MOOC platform, this behaviour is recorded as the interactions of the learners with the course content, generating a great deal of information that offers an opportunity for identifying patterns and predict trends [11]. Actually, using all these data to run predictions about learner's success in a MOOC is of special relevance. Understanding enrollees' learning behaviour can help to detect learners who "probably" will not pass the course [29]. Moreover, this analysis could be used to better understand how learners work in the course and what kind of support he/she may need, anticipating problems which may lead to learners' dropouts.

Several studies have tried to predict attrition, retention and completion in MOOCs. Most of these studies have been carried out in cohort MOOC settings (e.g., instructor based), where time is typically structured, learners follow a fixed schedule, and course materials are released at specific times. However, in self-paced MOOCs, this prediction models may be more critical. On the one hand, the success in self-paced courses, without the support of an instructor, depends on the ability of enrollees to be able to self-regulate their behaviour [21]. On the other hand, learners' behaviour could be more variable, since students do not follow a strict schedule, all materials are released when the course starts, and dates for assessments are flexible [15].

As a consequence, to detect and predict trends in self-pace MOOCs is still a challenge that have been addressed in prior works with different approaches. For example, authors in [27] developed a grade predictive method that uses learner activity features to forecast whether or not a learner may get a certificate. Authors in [5] developed a predicting model to understand when learners will answer a question correctly. In [26], authors analysed the relationship between interactions and the number of days in which learners interact with the content.

Despite of the predictive power of the models proposed, these models raised some discussions in the community. On the one hand, some researchers argue that frequency and events count are not the best metrics to obtain practical indicators to explain individual differences in online learning [28]. On the other hand, existing models are based on the use of low-level indicators of learners' interaction with the course, but this makes it difficult to obtain meaningful patterns of more complex behaviours, such the use of SRL strategies[28]. Therefore, there is an opportunity to improve these predictive models by considering both, data informing about the heterogeneity of learners (e.g. self-reported data about learning strategies) and more complex behaviours represented by activity sequences instead of individual events.

As a first proposal in this line, we present an exploratory study that uses SRL behavioural patterns related with learners' success as coarse-grained data to predict their behaviour in a self-paced MOOC. Specifically, we investigate whether or not learners pass the course based on these patterns together with demographic variables, SRL self-reported strategies and learners' intentions. As a result, we identified new factors to improve predictive models of learners' success in self-paced MOOCs.

2 Prior Work

2.1 Prediction in MOOCs and Self-Regulated Learning

MOOCs have special features that differentiate them from other online courses. First, the big amount of global data that can be collected about learners' activity with the course content. Second, the variety of this data, in which we can identify heterogeneous profiles in terms of personality, learning preferences, education, etc. And third, the number of the interactions related to intensive use of video-lectures and assessments, less frequent in traditional online courses [23]. All these data have been used to discover predictive patterns of persistence or attrition through MOOC success and completion. Specifically, the data sources used in previous work is usually: (1) learners' demographic data, (2) learners' self-reports data (as intentions regarding the course), (3) clickstream data, (4) forums and social media data and (5) other clickstream traces [14]. In the past years, recent studies started considering not only learners' demographic data for predicting behaviour, but also self-reported data related with more complex students' learning strategies. For example, studies [6, 7] found positive relationship between learners' self-reported SRL strategies and academic achievement. According to these studies, the use of SRL strategies affects the learning outcomes achieved and is typically associated with better academic performance in both traditional and online learning situations. In study [10, 19] authors found 15 learning strategies were correlate with learners' academic performance (final grades) in online environments, and 5 were found to predict learners' grades. In another example with 50,000 learners[13], authors found significant differences in the scores obtained by learners who were already familiar or working in fields related with the MOOC content, with higher self-efficacy, than their counterparts. In another study with 4,831 learners [14], authors found that goal setting and strategic planning predicted attainment of personal course goals. Further, in [9] in a study with 2,439 learners, authors found that having a particular help seeking strategy predicts better performance in the course.

Regarding clickstream, data with video-lectures, assessments and forums have been used in predictive models. For example, studies [14, 20] use video-lectures actions related to pause, play, stop video, watch, complete or review as a method for measuring learners' engagement the course content. Results of these studies showed that the amount of video-lectures intended and completed are predictors of course completion and showed that it is not necessary for learners to watch video-lectures from the beginning to the end to demonstrate its predictive effect [26]. In relation to assessment, different types of clickstream such as trying or completing an assessment, have been found to be predictors of course completion [20]. Researchers in [3], for example, found that the number of assessments' attempts is predictor of course completion; even more, those who try the first assessment were 30% less likely to drop out the course. Regarding the activity recorded in forums, the study [26] found that the number of forum pages viewed, or activities within the forum, such as voting up or down, were found as predictors of MOOC completion and persistence. Finally, some others clickstream traces have been found as predictors to MOOC persistence and completion, such as the number of active days that learners spent in a MOOC and the learners' pace through the contents [23].

Despite of their demonstrated predictive power, these models have some limitations. On the one hand, the use of these data sources as indicators for predict success in a MOOC are not always the more adequate. Learners' self-reported data captures only the intentions of the learners regarding the course, but not their actual behaviour. Since SRL is a continuous process rather than a single picture in time, considering indicators that come from the learners' activity within the course could be a better potential indicator. On the other hand, frequency counts of events from clickstream data and other clickstream traces that are obtained directly from low-level data are limited for detecting learners more complex behaviour in a MOOC for suggesting learning guidance. Moreover, as other studies already demonstrated, clickstream data in isolation do not necessarily build better predictive models [29]. Therefore, predictive models could be improved by adding variables built on longer activity sequences resulting from learners' interaction with the course content. That is, to propose new indicators that represent how learners adhere to the designed paths of the course, such as activity sequences extracted from coarse-grained data. This idea is built upon previous studies, which investigated the relationships between interaction sequences and learning outcomes using methods such as transition graphs, process mining, sequential pattern analysis, and Markov models [14, 20, 24].

Therefore, and based on prior work, this paper tackles the following research question: *Which indicators of SRL obtained from self-reported questionnaires and activity sequence extracted from trace data can predict course success in self-paced MOOCs?*

3 Methodology

3.1 Context: Sample and MOOC

This study uses data from one MOOC on Electronics[1] offered by Pontifical University Catholica of Chile in Coursera. The course was taught in Spanish and the materials were organised in four modules. In total the course included 17 lessons, 83 video-lectures and 16 summative assessments. The course followed a self-paced delivery mode in which course materials were available all at once, and without specific predefined deadlines. Data collection occurred between April and December of 2015.

A total of 25,706 learners registered for the MOOC, but the study sample is N = 2,035 which corresponds with those learners who answered a self-reported SRL questionnaire

[1] Coursera MOOC: Electrones en acción

that was introduced at the beginning of the course to define SRL learners' profile. Learners' average age was 30.7 years (SD = 11.06); and the 11% were women.

3.2 Measures

The instrument used to define learners' SRL profile was already validated in previous studies. It contains 35 questions about learners' intentions with the MOOC content (e.g., hours expected to be dedicated to the MOOC, interest in the topic, etc.), demography (e.g., age, gender, employment status, etc.) and a measure of SRL [14][2]. The SRL measure consisted of 24 statements related to six SRL strategies: goal-setting strategies (4 statements), strategic planning (4), self-evaluation (3), task strategies (6), elaboration (3) and help seeking (4). Learners rated statements using a 5-point scale (coded from 0 to 4), where a total average of 4 means a high SRL profile. The SRL measure exhibited high reliability for all strategy subscales with Cronbach's alpha of at least 0.70.

For this study, we also defined success in a self-paced MOOC based on the grades that learners achieve in the course. Therefore, success learners include any enrollee who meets one of the following two conditions:

1. obtains at least the minimum score to pass the course (80%) independently if he/she tackle most of the course materials (most common form of success),
2. obtains at least the minimum score to pass the course attempting at least 50% of the videos in the course materials

This choice is based on the common patterns that learners follow in a MOOC that were found in a previous work [20].

3.3 Procedure

In order to extract sequence patterns from a self-paced MOOC, we used the Process Mining method that was reported in [20]. This process is structured into four stages (see Fig. 1):

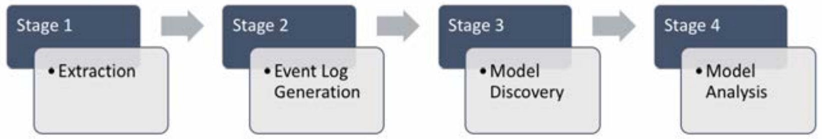


Fig. 1. Stages for extracting sequence patterns using Process Mining method.

(1) Extraction stage. In this stage, the data is extracted from the Information System databases (Coursera in our case). We obtained the trace data from Coursera database in order to study the interaction sequences of learners in the MOOC. This raw data is organised into three categories: (a) general data, (b) forums, and (c) personal data that contain relevant information about learners' behaviour.

[2] SRL measure questionnaire in Spanish and English are available at https://doi.org/10.6084/m9.figshare.1581491.

(2) Event log generation stage. In this stage gathered data is modeled in terms of event logs, defining the concepts of case (execution of a process), activities (steps of the process), and temporal order of the activities. We defined the main event log file including the learners' interactions in the MOOC within a session, their SRL scores, as well as information required to perform the analysis, such as the case id, time stamp and other resources. In this stage, we defined the concepts of (1) session and (2) interaction.

A session is defined as a period of time in which the Coursera trace data registers continuous activity of a learner within the course, with intervals of inactivity no greater than 45 minutes; this definition of study session has been already adopted in prior works [16].

An interaction is defined as an action recorded in the Coursera trace data that registers the interaction of a learner with a MOOC content. We defined six types of interactions depending on the content that learners interact with (video-lectures / assessments):

- **Video-lectures:** *(1) start a video-lecture* (begin to watch a video-lecture for the first time without completing it), *(2) complete a video-lecture* (watch a video-lecture entirely for the first time), *(3) review a video-lecture already completed* (go back to a video-lecture which was already completed)
- **Assessments:** *(1) try an assessment* (attempt to solve an assessment), *(2) pass an assessment* (successful attempt to solve an assessment for the first time), *(3) review an assessment already passed* (go back to an assessment that was previously completed successfully).

After defining these key concepts, we extracted the study sessions and coded as consecutive learning actions (interactive sequences) performed by learners when interacting with MOOC resources, such as video-lectures and assessments. Finally, we defined an event log that included a label to identify the first (begin session) and last interaction of the learner with the course (end session). Besides the interactions with the course, the event log also included learners' SRL scores obtained from the self-report questionnaire. The Table 1 shows an example of the event log generated.

Table 1. Example of the event log generated.

Case ID	Time stamp	Interaction	SRL Scores
1acc92cf40b27c8a36ea9d	1451023929	Begin session	3,162
1acc92cf40b27c8a36ea9d	1448567431	Video-Lecture.begin	3,162
1acc92cf40b27c8a36ea9d	1448567737	Video-Lecture.complete	3.162
1acc92cf40b27c8a36ea9d	1448568139	Assessment.try	3.162
1acc92cf40b27c8a36ea9d	1449105157	End session	3,162

(3) Model discovery stage. In this stage, Process Mining (PM) discovery algorithms are applied to the event log to obtain a process model (process map). This model represents the behaviour of the learners in the MOOC as a result of its interaction with the video-lectures and assessments. We selected the Disco algorithm and their implementation in the Disco commercial tool [12]. This algorithm is based on the Fuzzy algorithm concept

combined with some features from the Heuristic algorithm family [1]. We use this algorithm given that the exploratory context of this study in which is necessary to handle complex processes and the resulting models can be understood by experts in the domain without experience in PM [10].

(4) Model analysis stage. In this stage, the discovered process models are analysed in order to understand the observed behaviour (see Fig. 2). Once the process model was generated, we identified learners' most frequent interaction sequences that characterize each session for a learner (an interaction sequence is defined as a set of concatenated interactions, from one interaction to another one, of the same learner within a session). That is the learner's path followed in the MOOC within a session (see Fig. 3).

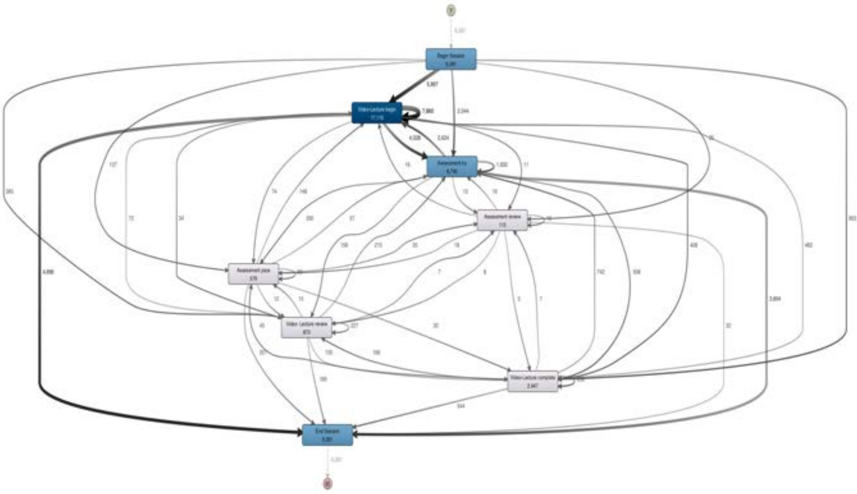
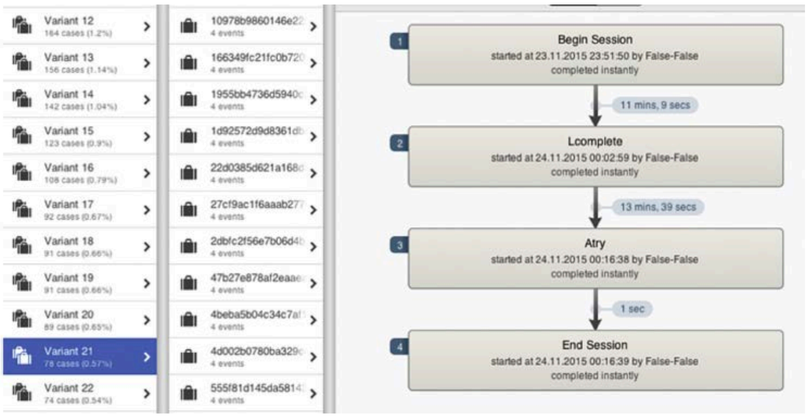


Fig. 2. Process Model obtained containing all interaction sequences by sessions.



Variant 12
164 cases (1.2%)

Variant 13
156 cases (1.14%)

Variant 14
142 cases (1.04%)

Variant 15
123 cases (0.9%)

Variant 16
108 cases (0.79%)

Variant 17
92 cases (0.67%)

Variant 18
91 cases (0.66%)

Variant 19
91 cases (0.66%)

Variant 20
89 cases (0.65%)

Variant 21
78 cases (0.57%)

Variant 22
74 cases (0.54%)

10978b9860146e22
4 events

166349fc21fc0b720
4 events

1955bb4736d5940c
4 events

1d92572d9d8361db
4 events

22d0385d621a168c
4 events

27cf9ac1f6aaab27
4 events

2dbfc2f56e7b06d4b
4 events

47b27e878af2eaae
4 events

4beba5b04c34c7af
4 events

4d002b0780ba329c
4 events

555f81d145da5814
4 events

1 Begin Session
started at 23.11.2015 23:51:50 by False-False
completed instantly

11 mins, 9 secs

2 Lcomplete
started at 24.11.2015 00:02:59 by False-False
completed instantly

13 mins, 39 secs

3 Atry
started at 24.11.2015 00:16:38 by False-False
completed instantly

1 sec

4 End Session
started at 24.11.2015 00:16:39 by False-False
completed instantly

Fig. 3. List of the 1366 sessions obtained using Disco software. Session 21 shows the begin and the end of the session and 2 interactions (events) with 3 interaction sequences and the time associated with the duration of the session (variant 21).

3.4 Proposed approach

Once the process model was generated and in order to answer the research question, we set up the proposed approach in two steps: (1) extracting meaningful SRL patterns, (2) applying predictive models.

(1) Extracting SRL patterns. We used Process Mining techniques following the PM^2 method used in [20] to identify the most frequent interaction sequences of learners. As a result, six interaction sequences patterns were identified: (1) *only video-lectures*, (2) *only assessment*, (3) *explore*, (4) *assessment-try to video-lecture*, (5) *video-lecture-complete to assessment-try*, and (6) *video-lecture to assessment-complete*. Then, the interaction sequences patterns extracted were used as input for grouping learners with similar behaviour. This was done through agglomerative hierarchical clustering based on Ward's method. This clustering technique is advisable for detecting learner groups in online contexts [16]. To select the optimal number of clusters, we inspected the resulting dendrogram and looked for different ways of cutting the tree structure, in order to obtain a minimal number of interpretable cluster explaining user behaviour (also the number of clusters were confirmed using the Silhouette method). As a result, the cluster indicates different kinds of learning strategies that learners deploy when they are facing the MOOC. Three clusters that classify learners according to their interaction sequences patterns and SRL profile were obtained. These clusters are:

- ***Sampling learners (cluster 1):*** They have a low activity in the course. Generally, learners in this group "sample" the course materials and then, leave the course (n = 1,530). Only 7 learners complete the course.
- ***Comprehensive Learners (cluster 2):*** These learners usually follow the path designed by the instructor. They also invest more time watching video-lectures and then try assessments for deeply learning (n = 85). Only 30 learners complete the course.
- ***Targeting Learners (cluster 3):*** They watch fewer video-lectures than comprehensive learners, and focus on completing the assessments, thus being more strategic or goal oriented (n = 420). Only 143 learners complete the course.

We look for statistically differences between clusters 1, 2 and 3 based only in the SRL profile (mean) running t-tests. As a result, no statistically significant differences between comprehensive (cluster 2) and targeting (cluster 3) learners were observed based on the SRL profile. Consequently, we selected these two as groups of interests to explore if we can find differences in the predictors of the grades between them.

(2) Applying Predictive Models. Once we identified the mined sequence patterns, we combined these with self-reported SRL strategies, other traditional self-reported variables such as demographics, intentions, and variables that result from the activity of the learner within the platform, in order to identify which of these variables (fine- and coarse grained) are predictors of learners' success in self-paced MOOCs.

In order to assess whether the variables in Table 2 had statistically significant and independent effects for predicting learners' success, we conducted multiple linear regression analyses and logistic regression analysis. Variables used in the predictive model were selected by means of a stepwise regression, using the 23 predictors. Stepwise regression uses an algorithm to select the best grouping of predictor variables that account for the most variance in the outcome (R^2); this technique is useful in exploratory studies or when testing for associations.

All the predictors are continuous except for gender, employment status, interest in topic, interest in assessment and prior experience, which are dummy-coded binary predictors. Finally, with the self-reported data on SRL strategies as well as the patterns extracted, demographic data about learners, intentions towards the course and activity registered in the course, we built a dataset containing 23 variables that were considered as possible predictors of success. These predictors are presented in Table 2.

Table 2. Predictors classified by categories

Category	Predictors
SRL Strategies	(1a) Goal setting (1b) Strategic planning (1c) Self-evaluation (1d) Task strategies (1e) Elaboration (1f) Help-seeking
Sequence patterns	(2a) Only video-lectures (2b) Only Assessment (2c) Explore (2d) Assessment-try to video-lecture (2e) Video-lecture-complete to assessment-try (2f) Video-lecture to assessment-complete
Demographics	(3a) Age (3b) Gender (3c) Employment status (student) (3d) Employment status (job)
Intentions	(4a) Time commitment (4b) Interest in topic (4c) Interest in assessment (4d) Prior experience
Activity	(5a) Active days (5b) Time spent (minutes) (5c) Number of sessions

4 Results

4.1 Regression analysis of course success

We assessed individual differences between three groups: (1) Comprehensive learners as a group (cluster 2), (2) Targeting learners as a group (cluster 3) and (3) all learners as one group (cluster 1, cluster 2 and cluster 3). For this assessment, we used 23 individual characteristics, encompassing SRL strategies, sequence patterns extracted from the behaviour of the learner with the course content, demographics, intentions and activity with the course resources. Fig. 4 illustrates the results of the regressions, one for each group, with estimated standardized coefficients (sign and magnitude) from each model in each column. Blank entries in Fig. 4 indicate that the corresponding predictor was excluded from the model. These standardized coefficients were obtained after running multiple linear regression and logistic regression. For each group, we have considered grades as a dependent variable. For multiple linear regression, the grades were considered as a continuous variable. For logistic regression, the grades were considered as a binary variable (grade >= 80; grade >= 80 & proportions of video-lectures >= 50%). A number of individual differences emerged for learners who succeed in a MOOC across different set of indicators and depending on the group in which they were classified. For comprehensive learners, the *strategic planning* strategy was associated with success in the course, while *elaboration and help seeking* were the strategies associated with success for targeting learners (grade >= 80; grade >= 80 & proportions of video-lectures >= 50%). Comprehensive learners who performed the sequence patterns *only assessment, explore, and assessment try to video-lecture* while they were facing the course, were more successful (grade >= 80; grade >= 80 & proportions of video-lectures >= 50%). Targeting learners who performed the sequence patterns *only assessment and assessment try to video-lecture* were more successful (grade >= 80), while for the same group the strategy *assessment try to video-lecture* was associated only with success (proportions of video-lectures >= 50%) if learners passed the course and attempted, at least, 50% of video-lectures. Regarding activity indicators, comprehensive learners who spent more *active days and time* in the MOOC were more successful, while targeting learners only *time spent* was associated with success.

To predict the final grade (as continuous), we run a stepwise method. As a result, we obtained 3 models for (1) Comprehensive learners as a group, (2) Targeting learners as a group, and (3) all learners as one group. Table 3 describes the regression models obtained for each group.

Table 3: Summary of the models using multiple linear regressions for the three groups (grade continuous)

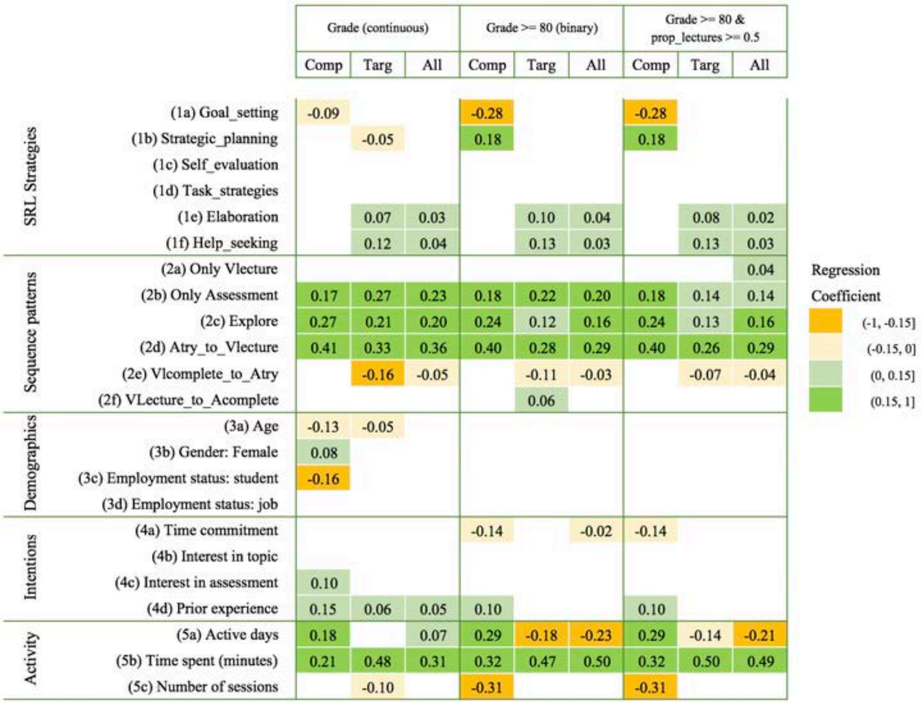
Group	**R^2**	***adj.* R^2**	**df**	**F**	***p***
(1) Comprehensive	0.8296	0.8039	73	32.31	<0.001
(2) Targeting	0.7249	0.7175	408	97.73	<0.001
(3) All	0.8559	0.8552	2026	1202	<0.001

For group (1) Comprehensive learners, the self-reported variable *goal setting*, the sequences patterns *only assessment*, *explore* and *assessment try to video-lecture*, the reported demographics as *young learners*, be *women* and *employment status as student*, the learners' *prior experience* and *interest in assessment* reported, the *active days* and the *time spent* were significant predictors of the final grade. These variables explained 80.39% of the variance in the final grade ($R^2 = .8039$, $F = 32.31$, $p < .001$).

For group (2) Targeting learners the self-reported variables *strategic planning*, *elaboration* and *help seeking*, the sequences patterns *only assessment*, *video-lecture complete to assessment try*, *explore* and *assessment try to video-lecture*, the reported demographics as young learners, the learners' *prior experience*, the *time spent,* and the *number of sessions* were significant predictors of the final grade. These variables explained 72.49% of the variance in the final grade ($R^2 = .7249$, $F = 97.73$, $p < .001$).

For group (3) "All learners as one group", the self-reported variables *elaboration,* and *help seeking,* the sequences patterns *only assessment*, *video-lecture complete to assess-*

ment try, *explore,* and *assessment try to video-lecture*, and the learners' *prior experience* reported, the *active days* and the *time spent* were significant predictors of the final grade. These variables explained 85.5% of the variance in the final grade ($R^2 = .855$, F =1,202, p < .001).



		Grade (continuous)			Grade >= 80 (binary)			Grade >= 80 & prop_lectures >= 0.5		
		Comp	Targ	All	Comp	Targ	All	Comp	Targ	All
SRL Strategies	(1a) Goal_setting	-0.09			-0.28			-0.28		
	(1b) Strategic_planning		-0.05		0.18			0.18		
	(1c) Self_evaluation									
	(1d) Task_strategies									
	(1e) Elaboration		0.07	0.03		0.10	0.04		0.08	0.02
	(1f) Help_seeking		0.12	0.04		0.13	0.03		0.13	0.03
Sequence patterns	(2a) Only Vlecture									0.04
	(2b) Only Assessment	0.17	0.27	0.23	0.18	0.22	0.20	0.18	0.14	0.14
	(2c) Explore	0.27	0.21	0.20	0.24	0.12	0.16	0.24	0.13	0.16
	(2d) Atry_to_Vlecture	0.41	0.33	0.36	0.40	0.28	0.29	0.40	0.26	0.29
	(2e) Vlcomplete_to_Atry		-0.16	-0.05		-0.11	-0.03		-0.07	-0.04
	(2f) VLecture_to_Acomplete					0.06				
Demographics	(3a) Age	-0.13	-0.05							
	(3b) Gender: Female	0.08								
	(3c) Employment status: student	-0.16								
	(3d) Employment status: job									
Intentions	(4a) Time commitment				-0.14		-0.02	-0.14		
	(4b) Interest in topic									
	(4c) Interest in assessment	0.10								
	(4d) Prior experience	0.15	0.06	0.05	0.10			0.10		
Activity	(5a) Active days	0.18		0.07	0.29	-0.18	-0.23	0.29	-0.14	-0.21
	(5b) Time spent (minutes)	0.21	0.48	0.31	0.32	0.47	0.50	0.32	0.50	0.49
	(5c) Number of sessions		-0.10		-0.31			-0.31		

Regression Coefficient

🟧	(-1, -0.15]
🟨	(-0.15, 0]
🟩	(0, 0.15]
🟩	(0.15, 1]

Fig. 4. Individual differences between 3 groups of learners (comprehensive, targeting, all) considering the grade as a continuous and binary variable (grade >= 80; grade >= 80 & proportions of video-lectures >= 50%), examined by SRL strategies, sequence patterns, demographics, intentions and activity. Blank boxes indicate predictor variables that were excluded by variable selection. Colors indicate the sign and magnitude of standardized coefficients. All regression coefficients are significant (p <.001).

The sequence patterns *only assessment*, *explore* and *assessment try to video-lecture,* and the *time spent* were significant positive predictor for the three groups. The magnitude of the standardized coefficient for the predictor *assessment try to video-lecture* for group "Comprehensive" and "All", and the magnitude of the standardized coefficient for the predictor *time spent* for "Targeting" were the highest. It is also worth noting that *video-lecture complete to assessment try* and *employment status as student* were significant negative predictors for "Targeting" and "Comprehensive" respectively.

Finally, an evaluation of the models was performed to analyze the predictive power. The dataset was split in train and test sets (80% for training and 20% for testing) and

10-fold Cross Validation (CV) was used within the training set. The first model to predict continuous grades was evaluated through the Root Mean Square Error (RMSE), while the other models to forecast binary variables were assessed through the accuracy, kappa and the Area Under the Curve (AUC) (see Table 4).

Table 4: Evaluation of the predictive models

Cluster	**Set**	**Grade (continuous)**	**Grade >= 80 (binary)**			**Grade >= 80 & prop_lectures >= 0.5 (binary)**		
		RMSE	*Accuracy*	*Kappa*	*AUC*	*Accuracy*	*Kappa*	*AUC*
All	*CV*	11.30	0.95	0.74	0.98	0.96	0.77	0.98
	Test	11.85	0.95	0.70	0.98	0.95	0.70	0.98
Comprehensive	*CV*	16.62	0.82	0.63	0.84	0.82	0.63	0.84
	Test	11.66	0.94	0.86	0.92	0.94	0.86	0.92
Targeting	*CV*	17.22	0.86	0.70	0.92	0.83	0.63	0.92
	Test	17.86	0.80	0.57	0.90	0.90	0.79	0.92

* *CV* – Cross Validation; AUC – Area Under the Curve

Results show that the predictive power is higher with all learners. This is normal because sampler learners are also included, and their grade is easier to predict given that sampler learners do not do the activities and they fail. As for comprehensive, some differences are encountered between the train and test set. The reason is that there are very few comprehensive learners and data limitations may suppose generalization issues. Nevertheless, the kappa values indicate at least substantial agreement [17] in all cases (in all groups) and AUC values are excellent [22] (excepting the AUC value for comprehensive learners in CV, which can be considered good). These results entail that the new variables related to self-regulated learning and sequence patterns can be useful for predicting grades, together with the well-known activity variables.

5 Conclusions

This paper has presented an exploratory study on the variables that are good predictors of the success (grades) for three groups of learners in a self-paced MOOC: "Comprehensive", "Targeting" and "All" learners. Comprehensive learners are those who follow the course path designed by the teacher. Targeting learners are those who seek for the information required to pass assessments. For both type of learners, we found a group of variables as the most predictive: (1) the self-reported SRL strategies 'goal setting', 'strategic planning', 'elaboration' and 'help seeking'; (2) the activity sequences patterns 'only assessment', 'complete a video-lecture and try an assessment', 'explore the content' and ' try an assessment followed by a video-lecture'; and (3) learners' prior experience, together with the self-reported interest in course assessments, and the number of active days and time spent in the platform.

The variables analysed in these groups were extracted from self-reported SRL strategies, mined interaction sequence patterns, traditional self-reported variables such as demographics, intentions, and variables that result from the activity of the learner within the platform. Multiple linear regression models were obtained for each of the three groups of learners, which are statistically significant at 99,9% level of confidence.

The findings of this study are subject to some limitations due to the nature of data, and methodological choices. First, the study is based on learners' behavioural data automatically collected by the platform, and self-reported data collected from an optional survey. Second, the study sessions are computed considering an inactivity threshold of 45 minutes, and only the interactions of learners with video-lectures and assessment were used to extract interaction sequence patterns.

Future work will expand the study considering (1) week by week analysis instead of per sessions, and (2) considering interaction sequence patterns mined by using other MOOC resources such as forum messages, readings, use of dashboard, access to external resources outside the MOOC, and formative activities. We will also consider exploring different types of courses, those that have a defined start and end date. This, with the aim of finding other factors that affect the predictive power when forecasting grades. The final aim is to better understand how a student reaches the status of comprehensive or targeting.

Acknowledgments

This work was supported by FONDECYT (Chile) under project initiation grant No.11150231, the MOOC-Maker Project (561533-EPP-1-2015-1-ES-EPPKA2-CBHE-JP), the LALA Project (586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), and CONICYT/ DOCTORADO NACIONAL 2016/21160081, the Spanish Ministry of Education, Culture and Sport, under an FPU fellowship (FPU016/00526) and the Spanish Ministry of Economy and Competiveness (Smartlet project, grant number TIN2017-85179-C3-1-R) funded by the Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional (FEDER).

6 References

1. Van der Aalst, W.M.P.: Process mining: data science in action. Springer (2016).
2. Bannert, M.: Promoting self-regulated learning through prompts. Zeitschrift für Pädagogische Psychol. 23, 2, 139–145 (2009).
3. Barba, P.G. de et al.: The role of students' motivation and participation in predicting performance in a MOOC. J. Comput. Assist. Learn. 32, 3, 218–231 (2016).
4. Boekaerts, M.: Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. Learn. Instr. 7, 2, 161–186 (1997).
5. Brinton, C.G. et al.: Mining MOOC Clickstreams: Video-Watching Behavior vs. In-Video Quiz Performance. IEEE Trans. Signal Process. 64, 14, 3677–3692 (2016).
6. Broadbent, J.: Comparing online and blended learner's self-regulated learning strategies and academic performance. Internet High. Educ. 33, 24–32 (2017).
7. Broadbent, J., Poon, W.L.: Self-regulated learning strategies &amp; academic achievement in online

higher education learning environments: A systematic review. Internet High. Educ. 27, 1–13 (2015).

8. Chuang, I., Ho, A.D.: HarvardX and MITx: Four Years of Open Online Courses--Fall 2012-Summer 2016. (2016).
9. Corrin, L. et al.: Using learning analytics to explore help-seeking learner profiles in MOOCs. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. pp. 424–428 (2017).
10. Davis, D. et al.: Activating learning at scale: A review of innovations in online learning strategies. Comput. Educ. (2018).
11. Grainger, B.: Massive open online course (MOOC) report 2013. Univ. London. (2013).
12. Günther, C.W., Rozinat, A.: Disco: Discover Your Processes. BPM. 940, 40–44 (2012).
13. Hood, N. et al.: Context counts: How learners' contexts influence learning in a MOOC. Comput. Educ. 91, 83–91 (2015).
14. Kizilcec, R.F. et al.: Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. Computers & Education. 104, (2017).
15. Kocdar, S. et al.: Measuring Self-Regulation in Self-Paced Open and Distance Learning Environments. Int. Rev. Res. Open Distrib. Learn. 19, 1, (2018).
16. Kovanović, V. et al.: Penetrating the black box of time-on-task estimation. Proc. Fifth Int. Conf. Learn. Anal. Knowl. - LAK '15. October, 184–193 (2015).
17. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics. 159–174 (1977).
18. Littlejohn, A. et al.: Learning in MOOCs: Motivations and self-regulated learning in MOOCs. Internet High. Educ. 29, 40–48 (2016).
19. Maldonado-Mahauad, J. et al.: A Questionnaire for measuring self-regulated learning in Massive Open Online Courses based on a systematic review. Under revision (2018).
20. Maldonado-Mahauad, J. et al.: Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. Computers in Human Behavior 80,179–196 (2018).
21. Maldonado, J.J. et al.: Exploring differences in how learners navigate in MOOCs based on self-regulated learning and learning styles: A process mining approach. In: Computing Conference (CLEI), 2016 XLII Latin American. pp. 1–12 (2016).
22. Mezaour, A.-D.: Filtering web documents for a thematic warehouse case study: eDot a food risk data warehouse (extended). In: Intelligent Information Processing and Web Mining. pp. 269–278 Springer (2005).
23. Moreno-Marcos, P.M. et al.: Analysing the predictive power for anticipating assignment grades in a massive open online course. Behav. Inf. Technol. 0, 0, 1–16 (2018).
24. Pardo, A. et al.: Generating actionable predictive models of academic performance. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. pp. 474–478 (2016).
25. Reich, J.: Rebooting MOOC research. Science (80-.). 347, 6217, 34–35 (2015).
26. Sinha, T. et al.: Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. arXiv Prepr. arXiv1407.7131. (2014).
27. Xu, B., Yang, D.: Motivation classification and grade prediction for MOOCs learners. Comput. Intell. Neurosci. 2016, 4 (2016).
28. You, J.W.: Identifying significant indicators using LMS data to predict course achievement in online learning. Internet High. Educ. 29, (2016).
29. Zhao, C. et al.: Discover learning behavior patterns to predict certification. In: Computer Science & Education (ICCSE), 2016 11th International Conference on. pp. 69–73 (2016).