Understanding Learner's Drop-out in MOOCs

Alya Itani[1], Laurent Brisson[1], and Serge Garlatti[1]

IMT Atlantique, Lab-STICC, UBL
F-29238 Brest, France
http://www.imt-atlantique.fr/fr/personne/{firstname-name}
{firstname.name}@imt-atlantique.fr

Abstract. This paper focuses on anticipating the drop-out among MOOC learners and helping in the identification of the reasons behind this drop-out. The main reasons are those related to course design and learners behavior, according to the requirements of the MOOC provider OpenClassrooms. Two critical business needs are identified in this context. First, the accurate detection of at-risk droppers, which allows sending automated motivational feedback to prevent learners drop-out. Second, the investigation of possible drop-out reasons, which allows making the necessary personalized interventions. To meet these needs, we present a supervised machine learning based drop-out prediction system that uses *Predictive algorithms* (Random Forest and Gradient Boosting) for automated intervention solutions, and *Explicative algorithms* (Logistic Regression, and Decision Tree) for personalized intervention solutions. The performed experimentations cover three main axes; (1) Implementing an enhanced reliable dropout-prediction system that detects at-risk droppers at different specified instants throughout the course. (2) Introducing and testing the effect of advanced features related to the trajectories of learners' engagement with the course (backward jumps, frequent jumps, inactivity time evolution). (3) Offering a preliminary insight on how to use readable classifiers to help determine possible reasons for drop-out. The findings of the mentioned experimental axes prove the viability of reaching the expected intervention strategies.

Keywords: Learning Analytics, Supervised Machine Learning, Massive Open Online Courses, Modeling Drop-out

1 Introduction

Massive Open Online Courses (MOOCs), offer an alternative education method that changed the standards of teaching and learning forever. In the after MOOCs era, education has reformed to become attainable to the whole public at any age, price, country, time, and mean [5]. This elevated ease and unrestricted access to material led to massiveness not only in the scale of participation but also in that of incompletion, commonly known as drop-out [14]. Consequently, a wide investigation on MOOC drop-out rates was provoked. The prevailing research on that subject, revolved generally around anticipating drop-out and studying solutions for preventing or decreasing it among learners [10,16]. Essentially, this is how applications of machine learning techniques for drop-out prediction started taking form. The literature encompasses several intervention strategies for drop-out prevention [3,7,8,11]. Some strategies assert sending automated motivational messages or emails from the prediction system to the spotted learners at-risk. While other strategies assert sending personalized intervention messages either directly to learners or to an intermediary party, usually the teacher. In return, this intermediary, teacher, chooses the necessary intervention to make after analyzing the information offered by the prediction system [7].

OpenClassrooms mainly intend to find the reasons for drop-out among its learners and prevent this drop-out when possible using the appropriate intervention strategy. Therefore, to help OpenClassrooms in meeting their needs, we present a supervised machine learning based drop-out prediction system that uses *Predictive algorithms* (Random Forest and Gradient Boosting) for automated intervention solutions, and *Explicative algorithms* (Logistic Regression, and Decision Tree) for personalized intervention solutions to learners through an intermediary teacher. We summarize our contributions as follows: (1) Proposing a predictive system that can detect at-risk droppers at different instants of the learner's interaction with the course. (2) Introducing and testing new features associated with learners' trajectory of engagement with the course. (3) Deploying the readability of different classifiers to offer suitable intervention strategies for both teachers and learners.

This paper is structured in 6 sections. Section 2 presents related works. Section 3 describes the predictive system. Section 4 presents the experiments and their results. Section 5 offers an exhaustive analysis and discussion of the obtained results. Finally, section 6 concludes the main findings.

2 Related works

The idea of applying learning analytics on MOOCs emerged with the rise of massive raw data from recorded learners activity on various MOOC platforms [15]. At first, lights were mostly shed on exploring and evaluating MOOCs and their low completion rates; often found to be $\leq 13\%$ [1]. Subsequently, researchers' inquisition started orienting toward understanding this immense drop-out among MOOC learners and its causes. They discovered that reasons for MOOC learners drop-out can be very diverse due to its audience heterogeneity. In that context, Khalil and Ebner [8], Colman [3], and Onah et al. [11] all investigated the reasons of this marked drop-out. The most addressed reasons in the literature can be summed up as follows:(1) Lack of intention to complete (ex: material hunters, curious explorers, assessment hater, etc.) (2) Personal circumstances (ex: lack of time, family situations, etc.) (3) Bad MOOC design (ex: inefficient material, high workload, shortage in organization etc.) (4) Deficiency in digital skills. (5) Inaccurate expectations. (6) Bad prior experience.

The investigation of MOOC drop-out and its reasons opened the horizon towards using machine learning techniques to predict drop-out ahead of time and try to prevent it. Initially, studies attempted drop-out predictions considering mono-type contextual features, like forum interactions or video restricted events [12,17]. However, such feature restrictions can restrain the model's predictive potential. Consequently, multi-type feature based prediction models emerged. Kloft et al. [9] proposed a machine learning algorithm that works on clickstream data and other features to identify learners' most active time and its effect on drop-out. Still, studies were mostly restricted to one prediction algorithm. Soon after, learning analytics predictive models became more and more advanced with various tested algorithms, proper feature selection testing and evaluation [18]. In this context, Hlosta et al. [6] present an early at-risk identification upon the absence of legacy information for the case of new courses.

3 Drop-out Predictive System

Figure 1 shows the proposed drop-out predictive system and the analysis process. This system uses the historical traces of learners in a given MOOC to construct models that accurately classify new learners into droppers and completers at some point in their progress. Hence, the prediction target in this problem is of two categorical classes (dropper, completer), and the dataset at hand is a labeled dataset. Therefore, we propose a system based on a supervised machine learning process.

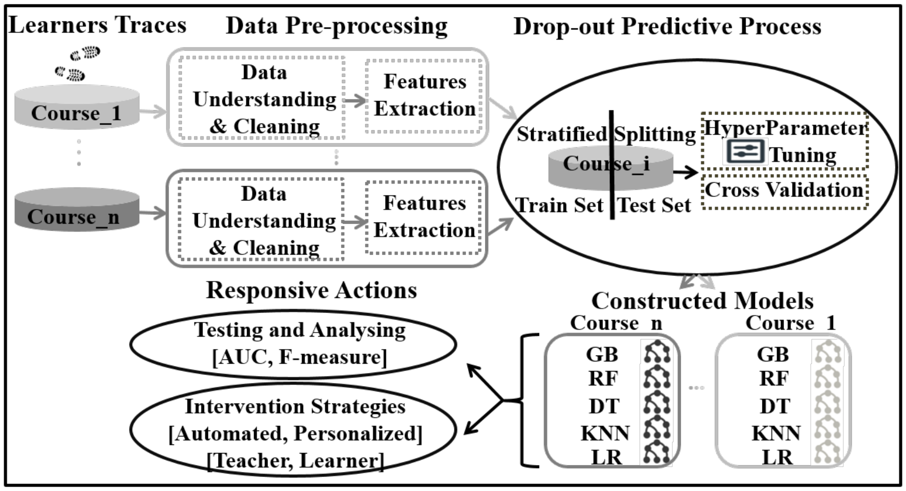

Fig. 1. Phases of the Drop-out Predictive System

Mainly, the process helps in attaining two goals: (1) Offering accurate predictions for the purpose of automated interventions such as motivational messages to learners (2) offering readable and explainable predictions in order to personalize interventions with learners or improve course structure. Here we describe the analysis process that we have implemented. Once the system is configured for a MOOC platform it is easy to automate the process.

Data Preparation In this phase, the collected historical traces undergo the necessary preparatory steps to be used by the classification process. The first step is *data exploration*, which is familiarizing with the data in hold (course structure, number of chapters and activities, important terms like: part, grade, success, etc.). Whereas, the second step is *structure validation and cleaning* of the dataset which is basically fixing any inconsistency in the data (missing entries, redundant entries, duplicates). Lastly, comes the phase of *features selection*, it involves constructing a features matrix to be used as input to the classification process. Typically, the features in this matrix should include a representation of any useful information available in the data.

Classification Process and Models Construction In this phase, the classification process takes as input the constructed features matrix. We construct efficient predictive models that are optimized on different level of their construction. First, we perform a *Stratified Splitting* of the data. This implies partitioning the final features dataset into a training set (60%) and a testing set (40%) while preserving the initial dataset balance of droppers and completers. The testing test is used to compare the performance obtained by the 4 algorithms tested under different experimental conditions. Second, *Grid Search Hyper-parameter Tuning* is performed by searching exhaustively for the best scoring parameters for each model through a manually specified subset of the hyper-parameter space [19]. Third, to avoid overfitting *K-fold Cross Validation* is applied, by randomly partitioning the data into K subsamples. This action is then repeated K times with measuring the performance on each time and then averaging it at the end.

Responsive Actions After the construction of the five classification models, the models are tested and evaluated on testing set. Explicative models are analyzed and the obtained information is sent to teachers for investigation.

4 Case study and Experimentation

In this case-study, we conducted an experiment on a data set from the OpenClassrooms MOOC platform. The interest of the platform's analysts was to predict the drop-out of the platform's premium members by being able to explain the reasons to Mooc's designers. However, one constraint was not to use the demographic and social data of the platform users. We propose advanced features related to learners trajectories of engagement with the course to test their effect on the predictive efficiency and readability of the results.

4.1 The OpenClassrooms Case study

This dataset includes activity traces of 20,142 premium learners within *"Create your Website with HTML"* (HTML5) and *"Understanding the Web"* (Web) courses from October 2015 till October 2016. OpenClassrooms courses have no sessions or weeks, once a course is posted online it is available for students to start following at any time. Also, there is no maximum duration limit for finishing a course (a learner can take months to finish a course). Each course is divided into chapters and each chapter into parts. At the end of each chapter, there are graded multiple choice or peer assessed exercises. Upon completing all exercises, learners are given a final course grade.

Table 1. Statistics of Premium Learners Population in *"Create your Website with HTML"* (HTML5) and *"Understanding the Web"* (Web) courses

Measures	HTML5	Web
Number of Learners	12,114	7,379
Number of Active Learners	11,520	7,160
Completers (%)	4,333 (37.6%)	5,085 (71%)
Droppers (%)	7,187 (62.4%)	2,075 (29%)

Data Description The dataset contains subscription related events (following and un-following a course), course related events (visualization events, completions events, grades), and exercise session events. Table 1, describes the distribution of the different types of learners in our data set. We can notice that dropper rates are low compared to the rates generally observed in MOOCs: HTML5 course has a balance of 40% droppers and 60% completers, whereas Web course has 20% droppers and 80% completers. This is because we are only interested here in the premiums members of the platform: the motivation increases when a payment is involved [4].

Features selection MOOC designers know that learners rarely navigate the course in its planned linear manner. They rather go back and forth creating back and forward jumps. Therefore, we introduce two types of indicators for features selection: descriptive indicators and behavioral indicators. Descriptive indicators describe learner-course interactions. Whereas, behavioral indicators describe learners trajectories of engagement with the course versus the recommended trajectory of the course. The expected worth of behavioral indicators comes from their ability in revealing the flaws or strengths of the MOOC design and content. Indeed, bad MOOC design or inappropriate MOOC content are considered inevitable reasons of dropping (see section 2). Tables 2 and 3 offer a detailed view on both indicator types.

Table 2. Descriptive indicators and corresponding features

	Completed parts of the course
Definition:	An estimate of completed MOOC parts for each learner
Features:	Binary features with values 1: Completed Part 0: Uncompleted part A Part can be either a chapter or an exercise
Purpose:	Permits studying the effect of parts completion on learner's drop-out
	Exercise scores
Definition:	Incorporates the learners' scores on each completed exercise
Features:	Numeric values of grades, 0 denotes an uncompleted exercise
Purpose:	Allows studying the effect of grades on MOOC completion
	Time passed on exercises
Definition:	An estimate of the time passed on each completed MOOC exercise
Features:	Numeric values of time in seconds, 0 denotes an uncompleted exercise
Purpose:	Helps in studying the effect of exercise-invested time on completion

Table 3. Behavioral Indicators and Corresponding Features

	Number of back jumps in a course
Definition:	Number of back jumps performed throughout the course for each learner
Features:	Numeric value, number of performed back jumps
Purpose:	Allows studying the effect of back jumps on course completion
	Most frequent jumps in a course
Definition:	The N most frequent jumps performed by learners in the MOOC
Features:	N binary features with values 0: Jump not made, 1: Jump made
Purpose:	Study the effect of performing the most frequent jumps on completion
	Inactivity time evolution
Definition:	The learner's evolution of inactivity time between two parts of the course
Features:	Numeric value representing a logarithmic scale of time
Purpose:	Study the effect of increase or decrease of inactivity on completion

The final features matrix includes 34 mixed type features (numerical and categorical) alongside one categorical binary target variable, where 0 denotes completion and 1 denotes dropping.

4.2 Experimentation

We consider four main aspects upon the evaluation of the proposed drop-out predictive system:

1. The efficiency of prediction at different instants of the course. In other words, the system is tested for classifying learners into either completers or droppers at different points in the course progression, we test on 25% and 50% of activities. If a user skip some sections of the course, corresponding features are marked as uncompleted (see Table 2). All features after 25/50% of the course are left out so the dataset changes slightly between experiments. If

a user skip the course before 25/50% of the activities are presented, he is considered as a dropper but is not removed from the dataset.

2. The effect of dynamic behavioral indicators on the predictive system's efficiency and readability. The system's performance is tested with behavioral indicators vs. without behavioral indicators.
3. Variety of supervised classification algorithms with hyper-parameter tuning[1]. We test two different types. Explicative ones that are simple readable algorithms including Decision Tree (DT) and Logistic Regression (LR). Aggregated ones that have generally better prediction rates, but are more complex to read, including Gradient Boosting (GB) and Random Forest (RF).
4. Multiple course topics: *"Create your Website with HTML"* (HTML5) and *"Understanding the Web"* (Web).

Tables 4 and 5 compare the performances obtained during the experiments using the F-measure which is a balance between precision and recall [13].

Table 4. *"Create your Website with HTML"* (HTML5) course (35 activities). Performance of 5 classifiers on the F-measure metric.

Percentage of activities completed:		25% (9 activities)		50% (17 activities)	
Behavioral Indicators:		with	without	with	without
Algorithms:	RF	0.76	0.77	0.84	0.85
	GB	0.79	0.77	0.85	0.85
	DT	0.75	0.78	0.85	0.85
	LR	0.74	0.74	0.85	0.85

Table 5. *"Understanding the Web"* (Web) course (23 activities). Performance of 5 classifiers on the F-measure metric.

Percentage of activities completed:		25% (6 activities)		50% (12 activities)	
Behavioral Indicators:		with	without	with	without
Algorithms:	RF	0.26	0.26	0.91	0.91
	GB	0.25	0.25	0.91	0.91
	DT	0.25	0.25	0.91	0.91
	LR	0.46	0.46	0.90	0.91

A first analysis of these results leads us to two findings that we discuss in the next section:

[1] You can access the selected parameters for each model after hyper-parameter tuning by consulting the following link: http://www.laurent-brisson.fr/publication/2018-understanding-learner-dropout-mooc/

- The low impact of behavioral indicators on the prediction performance (section 5.2).
- The disparate impact, depending on the studied MOOC, of the number of activities completed on the performance of predictions (section 5.3).

Other related research that also address post-hoc learning methods (i.e. earning takes place on the same course) exists in the literature with close marked attainments on early drop-out detection. For example, Whitehill et al. [18] obtained a 90.20% AUC in the detection of drop-out averaged over 8 weeks on HawardX MOOCs. In the case of our predictions at 50% of the activities we obtain an AUC around 85% which corresponds to the same order of magnitude even if the use of two different data sets and different variables (they use "*click-streams features* that contains all interaction events between every student and the MOOC courseware") makes the comparison difficult. Additionally, neural network based methods can overcome our outcome in performance measures, but are out of our research scope and objective. In this paper, our goal does not stop at detecting drop-out we rather seek insights and actionable outcomes to help MOOC providers and stakeholders make the right decisions and understand early drop-out (with 25% of activities carried out).

5 Results Discussion

In this section we analyze the results taking into consideration: 1) the readability of models, 2) the effect of behavioral indicators and 3) the effect of the number of completed activities at the time of prediction.

5.1 Models readability

One of the objectives of this experiment is to be able to help decision-makers, in this case the MOOC designers, to improve the structure and content of their course. Here we will compare two types of algorithms, predictive algorithms (Random Forest, Gradient Boosting) and explanatory algorithms (Decision Tree, Logistic Regression). By comparing the results obtained in Tables 4 and 5 we can realize that, in our context, purely predictive algorithms do not really have better results than explanatory algorithms. It would be interesting to know whether this is due to the nature of the studied MOOCS or due to a platform effect but this is beyond the scope of this paper.

Predictive algorithms Although they are not readable at all, these models can still be used to determine which features have the most influence on dropout prediction. In the case of our experiment, the most discriminating variables were the scores obtained at the end of chapter exercises. However, these methods do not indicate the direction of this influence (positive or negative) on drop-out, which is very damaging to understand the context of each influencing feature.

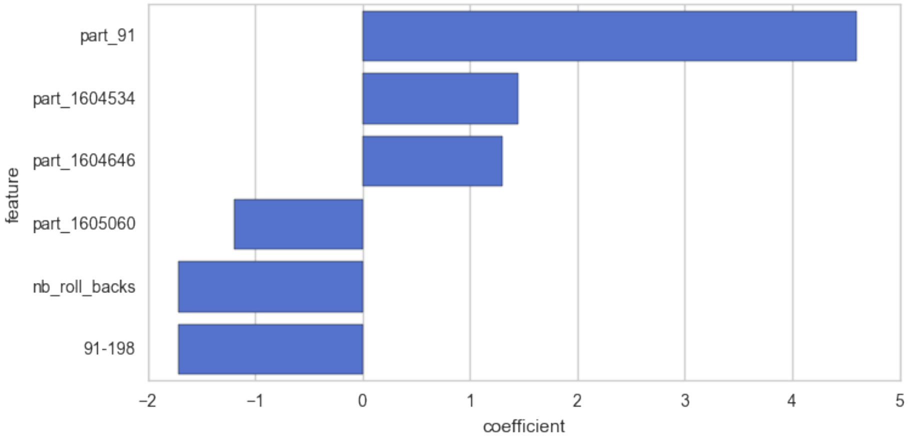

Fig. 2. Logistic Regression coefficients for predicting drop-out in *"Create your Website with HTML"* (HTML5) course with 25% of the activities. Display of the 6 most discriminating features.

Explicative algorithms Logistic regression is capable of demonstrating the influence of features on drop-out along with the direction of each feature's influence. Figure 2 shows the model coefficients for the *"Create your Website with HTML"* (HTML5) course after 25% of the activities have been completed. We can observe the negative impact of completing some activities (for example *part_91*) and the positive impact of some jumps (for example *91-198*) on dropout. To understand this phenomenon it is necessary to know that the activity *part_91* corresponds to the last quiz of the first chapter, and that *part_198* is the following activity which consists in making a bibliographic research. Thus, here Logistic Regression shows that the chance of success increases when learners carry out these two activities in their expected linear manner (one after the other).

Decision trees allow us to understand and measure the impact of each feature on the prediction. In Figure 3 the dropping class nodes are in dark gray and the completion class nodes are in light gray. Each node details the feature condition, total number of samples, dropper samples, completer samples, and %ratio. Furthermore, decision rules can be derived from each tree and can hold important information on the manner droppers behave. For example a rule for HTML5 droppers at 25% of course activities: "If a learner does not jump from *part_91* to *part_198* (which is the recommended progress) $\Longrightarrow$ he is at risk of dropping with %D=81.06% (out of 3364 learners)."

That kind of information can be very helpful in discovering problems related to course material, design, or level of difficulty. However, with inspecting the entire tree, dozens of rules can be derived from each model which is a weakness of decision trees. Here, after pruning, it is easy to observe interesting rules but the trees generated can be very large and become difficult to interpret.

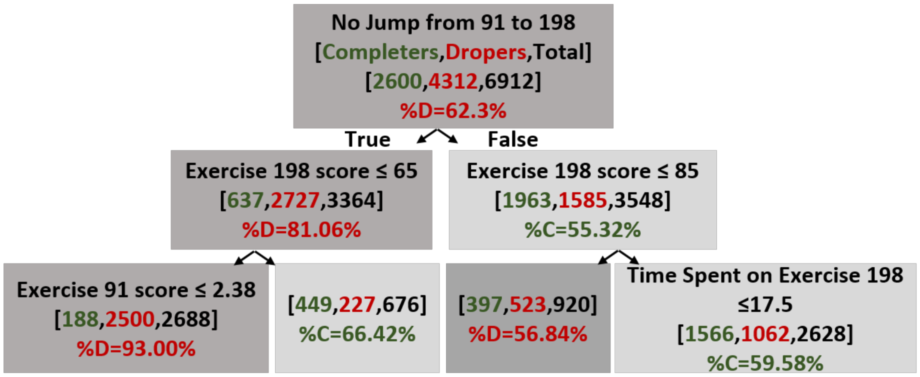

Fig. 3. Decision Tree (pruned after the 2nd branch) for predicting drop-out in *"Create your Website with HTML"* (HTML5) course with 25% of the activities.

5.2 Effect of Behavioral Indicators

Tables 4 and 5 show the impact of behavioral indicators on the classification results: when these indicators increase the predictive performance of classifiers boxes are grayed out. We can notice that the performance increase is very low for the MOOC *"Create your Website with HTML"* (HTML5) while it is negligible for the MOOC *"Understanding the Web"* (Web). However, the results here are not disappointing because they open the doors wide to discussion with the designers of the MOOC. As we saw in section 5.1, behavioral indicators allowed us to understand a phenomenon with the HTML5 course: the risk of drop-out increases with the completion of some activities while it decreases with the completion of a sequence of two activities. This is interpreted by the fact that this course attracts an audience looking for resources (a video explaining a concept, a definition or an exercise to practice) for which there is no commitment in successfully completing the MOOC. We distinguish here 3 types of interesting situations from a business point of view:

- The most discriminating feature is a descriptive indicator: the teacher must ask himself about the relevance of an activity. If it helps in detecting student involvement, this is a good thing. If it identifies a difficulty (related to a tool or concept), an accompanying measure should be put in place.
- The most discriminating feature is a behavioral indicator: this is the ideal case to suggest new routes to students, or to remind students of the importance of following the recommended progression as they move away from it.
- The prediction is bad and so no feature is really relevant: the proposed activities do not anticipate successes and failures. Either there is nothing to do, activities are designed to build student self-confidence, or there is a corrective action to consider to implement constructive alignment [2] if any of the following activities has a high failure rate.

5.3 Effect of Prediction at Different Course Instants (25%, 50%)

Tables 4 and 5 also show the impact of the number of activities considered on the results: we will focus here on predictions at two instants after having carried out 25% then 50% of the activities. Not surprisingly here, the increase in the number of activities considered to make the prediction improves the system's performance. However, what interests us here is the increase in this performance according to the course. For the course *"Create your Website with HTML"* (HTML5), the F-measure increases on average by 8.5 between the two instants (25/50%), while for the course *"Understanding the Web"* (Web) the F-measure increases on average by 61.5, i.e. 7 times more!

This difference can be explained by the presence in the Web course of an activity which takes place between 25% and 50% of their progression. In this example, the quiz in the first chapter was taken into account in the 25% of activities carried out, while the last exercise in the chapter was not included. We illustrate here very well the third situation presented in section 5.2 where the activities carried out are aimed at putting the student in confidence. The fact of making predictions at different stages of completion of activities thus makes it possible to observe a learning dynamic.

6 Conclusion

In this paper we present a supervised machine learning based drop-out prediction system that uses aggregated and explicative type classifiers. The aggregated classifiers can help in accurately detecting at-risk droppers, which allows sending automated motivational feedback to learners. Whereas, explicative classifiers allows the personalized intervention through a teacher. We state the findings according to the three main tested axes: (1) **Readability of explicative models**: Decision Trees and Logistic Regression permit the detailed inspection of the classification process and the effect of features on this classification. They could be hard to interpret by non-experts, but they can be used to send teachers valuable information to analyze and accordingly make personalized interventions. (2) **Dynamic Behavioral Indicators**: Including these indicators enhances slightly the predictive performance of the system, but it noticeably contributes to the readability of the prediction, however their effect depends highly on the studied course and material included. (3) **Prediction at different instants of the course**: the further the instant is, the more material included, the better is the performance of the system. Predicting at different instants can expose activities that are critical for classification. However, a critical activity for classification is not necessarily a critical pedagogical activity, and the current proposed system rather sheds the light on interesting aspects that can aid teachers in uncovering problems in course design and material.

References

1. Y. Belanger and J. Thornton. Bioelectricity: A quantitative approach duke university's first mooc. Technical report, 2013.
2. J. B. Biggs. *Teaching for quality learning at university: What the student does*. McGraw-Hill Education (UK), 2011.
3. D. Colman. Mooc interrupted: Top 10 reasons our readers didn't finish a massive open online course. *Open Culture*, 2013.
4. K. Devlin. Moocs and the myths of dropout rates and certification. *Huff Post College. Retrieved March*, 2:2013, 2013.
5. E. J. Emanuel. Online education: Moocs taken by educated few. *Nature*, 503(7476):342–342, 2013.
6. M. Hlosta, Z. Zdrahal, and J. Zendulka. Ouroboros: early identification of at-risk students without models based on legacy data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 6–15. ACM, 2017.
7. S. M. Jayaprakash, E. W. Moody, E. J. Lauría, J. R. Regan, and J. D. Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.
8. H. Khalil and M. Ebner. Moocs completion rates and possible methods to improve retention - a literature review. In J. Viteli and M. Leikomaa, editors, *Proceedings of EdMedia + Innovate Learning 2014*, pages 1305–1313, Tampere, Finland, June 2014. Association for the Advancement of Computing in Education (AACE).
9. M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. Predicting mooc dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 60–65, 2014.
10. E. Lackner, M. Ebner, and M. Khalil. Moocs as granular systems: design patterns to foster participant activity. *Retrieved September*, 10:2015, 2015.
11. D. F. Onah, J. Sinclair, and R. Boyatt. Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 Proceedings*, pages 5825–5834, 2014.
12. A. Ramesh, D. Goldwasser, B. Huang, H. Daumé III, and L. Getoor. Learning latent engagement patterns of students in online courses. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1272–1278. AAAI Press, 2014.
13. C. J. V. Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.
14. R. Rivard. Measuring the mooc dropout rate. *Inside Higher Ed*, 8:2013, 2013.
15. Y. Tabaa and A. Medouri. Lasym: A learning analytics system for moocs. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 4(5), 2013.
16. M. Wen, D. Yang, and C. Rose. Sentiment analysis in mooc discussion forums: What does it tell us? In *Educational Data Mining 2014*, 2014.
17. M. Wen, D. Yang, and C. P. Rosé. Linguistic reflections of student engagement in massive open online courses. In *ICWSM*, 2014.
18. J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley. Delving deeper into mooc student dropout prediction. *arXiv preprint arXiv:1702.06404*, 2017.
19. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.