Predicting attrition from Massive Open Online Courses in FutureLearn and edX

Ruth Cobos[1], Adriana Wilde[2], and Ed Zaluska[2]

[1] Departamento de Ingeniería Informática, Universidad Autónoma de Madrid, Spain
[2] Electronics and Computer Science, University of Southampton, United Kingdom

ruth.cobos@uam.es orcid.org/0000-0002-3411-3009
a.wilde@soton.ac.uk orcid.org/0000-0002-1684-1539
ejz@ecs.soton.ac.uk

Abstract. There are a number of similarities and differences between FutureLearn MOOCs and those offered by other platforms, such as edX. In this research we compare the results of applying machine learning algorithms to predict course attrition for two case studies using datasets from a selected FutureLearn MOOC and an edX MOOC of comparable structure and themes. For each we have computed a number of attributes in a pre-processing stage from the raw data available in each course. Following this, we applied several machine learning algorithms on the pre-processed data to predict attrition levels for each course. The analysis suggests that the attribute selection varies in each scenario, which also impacts on the behaviour of the predicting algorithms.

Keywords: MOOCs, learning analytics, prediction, attrition, attribute selection, FutureLearn, edX.

1 Introduction

The advances in telecommunications in the last decade, together with an increased accessibility to personal computers and internet-enabled devices, have revolutionised teaching and learning. This increased accessibility has meant that for more than 35 million students, geographical and economical barriers to learning have been overcome by accessing Massive Open Online Courses (MOOCs) offered by more than 500 universities. This is a figure which has doubled from 2014 to 2015, and is expected to continue to increase, given that (according to Class Central [1]) "1800+ free online courses are starting in October 2016; 206 of them are new".

The richness of the diversity of learning with MOOCs provides unprecedented opportunities for study. In tackling this diversity, it helps to understand the principles and affordances given by the FutureLearn platform compared with another well-recognised

FutureLearn data: what we currently have, what we are learning and how it is demonstrating learning in MOOCs. Workshop at the 7th International Learning Analytics and Knowledge Conference. Simon Fraser University, Vancouver, Canada, 13-17 March 2017, p. 1-20.

MOOC provider with similar courses which could be used for a comparative study (such as edX).

Against this background, we investigated whether the inherent similarities and differences between these two different MOOCs platforms (FutureLearn and edX) could influence learner behaviour (assuming all other things are equal) and whether there are any observable factors that can provide an early attrition prediction in either case. This is especially valuable as it could be used to inform interventions designed to improve learners' performance in future courses.

The structure of this paper is as follows: in section 2 (*Positioning FutureLearn and edX*) we discuss theoretical underpinnings and describe the practical organisation of one exemplar FutureLearn course, contrasting it against that of a comparable edX course. In section 3 (*Learning Analytics*) we review related work on learning analytics, which has predominantly been concerned with studying dropout rates and in demonstrating the feasibility of machine learning algorithms for classification and prediction. In section 4, (*Context of the present approach)*, the research questions are specified; whilst the processes followed in addressing them are described in section 5 (*Methodology*) alongside a detailed description of the courses selected (as the context of our study) and other technical details. The results are shown in section 6 (*Analysis and Discussion*), and the insights obtained are summarised in section 7, (*Conclusions and Future Work*), where we also identify further research.

2 Positioning FutureLearn and edX

The emergence of MOOCs is a consequence of the increased interconnectivity of the digital age. When Siemens [2] proposed connectivism as a new theory to sit alongside classical learning theories (of which Piaget's constructivism is an example [3]), pioneer online courses started to be created based on this theory: people learn by making meaningful connections between knowledge, information resources and ideas during the learning process. The key to a successful connectivist course would therefore be the use of a platform which fosters the formation of such connections in a distributed manner. These courses have become known as c-MOOCs, of which the first one was delivered in 2008 by Siemens and Downes [4].

In contrast, other courses were designed to adapt the medium, learning materials and assessments of traditional (instructivist, or cognitive behaviourist [5]) courses so that these could be delivered at scale. Under instructivism, learning is also an active process, but the relationship between teachers and learners is key – the relationship is mediated through specific tasks which are assessed as a measure of the learning process. These MOOCs have become known as x-MOOCs, a term coined by Downes in 2012 to differentiate them from his c-MOOCs. The first x-MOOC was delivered in 2007 though: the *Introduction to Open Education*, by David Wiley from Utah [6].

Noting that there are many similarities as well as important differences between learning-MOOCs and x-MOOCs (summarised in **Table 1**), it is interesting to compare them in practice by analysing case study courses.

Table 1. Summary of similarities and differences between c-MOOCs and x-MOOCs.

Characteristic	c-MOOCs	x-MOOCs
Number of learners	Should scale to large numbers	Should scale to large numbers
Method of delivery	Online	Online
Communication approach	Distributed	Centralised
Related learning theory	Connectivism	Instructivism or behaviourism
Design *primarily* supports the…	… creation of *connections* between learners, resources and ideas	… relationship between teachers and learners, mediated through *task* completion
First MOOC delivered (year)	Connectivism and Connective Knowledge (2008)	Introduction to Open Education (2007)

It is of interest to investigate whether the inherent similarities and differences between these models (which in turns translate in a number of affordances provided by MOOC platforms) may influence learner behaviours. In particular, in this research study we compare the results of applying algorithms for predicting course attrition within two case studies. More specifically, we have selected a FutureLearn MOOC and an edX MOOC, and secured the corresponding datasets for their analysis.

FutureLearn courses are organised in weeks. Each week contains a set of activities, called "steps", each of which has a learning object belonging to a prescribed category. Typical examples of these categories are: videos, articles, exercises, discussions, reflections, quizzes and peer reviews. For each step, learners are able to write comments, each of these in turn can be visibly "liked" (as in mainstream social media platforms) and have replies or follow-up comments. This facility allows learners to build connections amongst the community and with the learning objects presented, as often these comments allow for their personal reflections and expressions of their own understanding (or lack thereof). The use of such architecture reflects FutureLearn's pedagogical underpinnings inspired in social constructivism and Laurillard's conversational framework [7]. As explained before, with this approach, learning is the result of the social interaction between peers, so the platform has been built in order to afford this connectivist characteristic (and continues to be updated with new features that provide such affordances[1]).

[1] A recent innovation is the facility to work in small groups "to come together and reach shared understanding" (`https://www.futurelearn.com/about-futurelearn/our-principles`).

Similarly to FutureLearn courses, edX courses consist of weekly sections, which are composed of one or several "learning sequences". These learning sequences are composed mainly of short videos and exercises, often with the addition of extra educational content such as links to web pages or interactive educational resources. All courses have an online discussion forum where students are able to post and review questions and comments to teaching assistants and other students. Despite offering this facility, the fostering of conversations for co-creation of knowledge does not feature prominently in the guiding principles of their pedagogy. Instead, edX courses can be categorized as x-MOOCs and follow an instructivist approach where the assessment is based on the completion of exercises. The data traces that learners create through their participation in the courses not only allow the institutions to award certification (when all assessment has been satisfactorily completed) but as it is recorded, it has the potential to be analysed further to predict whether the learner is likely to be eligible for a certificate at the end of the course.

Differences of pedagogical approaches aside, both FutureLearn and edX capture data related to learners' participation in their courses. These data traces left behind by participating learners allow the institutions to award certificates (when all assessment has been completed to satisfaction) but also, while they are captured, have the potential to be analysed to predict whether the learner would be eligible for a certificate at the end of the course [8].

In practical terms, the platform data is collated by both MOOC providers and given to the subscribing institutions with a structure that specifically affords the study of behavioural characteristics of the learners in the course (e.g. their graded achievements or the social interactions of the learners). This wealth of data offers great opportunities for collecting learning analytics (discussed in section 3, below), however, there are challenges in aligning the data collected and performing comparison studies not only because of the fundamentally different approaches taken by each MOOC but also important differences in the technical implementations adopted.

3 Learning Analytics

The term *learning analytics* is widely understood as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" [8].

Despite the recent coinage of this term, applying analytics in learning and education has a long tradition, as educators have been interested in issues such as attrition and dropout rates for many years. Attrition is typically understood in its simplest terms as the rate in which the total number of learners in a course diminishes over time, or the number of individual learners who drop out against those who were originally registered. This definition has been long recognised as not being able to capture the dynamic nature of learning [9], as it conflates the failures with the successes in a non-traditional learning path. For example, a student who "drops out" may have just moved to another course, more suitable to their learning needs, or may have just temporarily suspended (to resume their learning at a later date).

However, the simplicity of this definition allows its application in many scenarios, and this is of special interest in the context of Massive Open Online Courses (MOOCs). Their arrival has coincided with the increasing application of learning analytics as a transformative force: the value of analytics has now been extended beyond the merely administrative and can now be used to inform and transform the teaching process, with a potentially-significant impact on learning, assessment processes and scholarly work (as foreseen by Long and Siemens [10]).

The arrival of MOOCs coincided with increased concern with dropout rates in traditional education. Although MOOCs attract many learners, typically only a very small proportion actually completes these courses, following what Doug Clow called a "funnel of participation" [11]. Kizilcec *et al.* [12] acknowledged that high dropout rates have been a central criticism of MOOCs in general and performed a cluster analysis of the disengaged learners in a study that was one of the first to demonstrate the potential of analytics in understanding dropout. Through this work, Kizilcec *et al*. [12] were able to identify prototype learner trajectories (auditing, completing, disengaging, sampling) as an explanation of learner's behaviour, effectively deconstructing attrition.

Despite the limitations of dropout as a stand-alone metric, which has led researchers to question its value as a measure of success both in MOOCs [13, 14], and in the context of blended learning [15] there are still considerable research efforts on reducing overall student attrition [16, 17, 18, 19, 20, 21, 22], as it is a well-understood metric of course engagement that is still useful while an improved metric for success is agreed. Such a metric could include contextual factors such as learner intention.

It is worthwhile noting that the importance of accurate predictive models of attrition or disengagement as studied through MOOC data can also be applied to face-to-face instruction, by making available predictions to teachers so that they can provide timely feedback or take any other suitable action [14],[23].

4 Context of the present approach

As explained earlier, the main motivation for this research study was to investigate whether the inherent similarities and differences between these two different MOOCs platforms could influence learner behaviour and whether there are any observable factors that can provide an early attrition prediction. The authors are researchers from two institutions, each delivering MOOCs under one of the paradigms, hence enabling a comparative study. The institutions are the Universidad Autónoma de Madrid (UAM) and the University of Southampton (UoS).

The UAM became a member of the edX consortium in 2014 and currently offers eight MOOCs at edX [24]. The University of Southampton (UoS), was one of the first FutureLearn partners, joining the consortium in 2013, and currently offers 15 MOOCs at FutureLearn [25].

4.1 Research questions

Against the existing background, we formulated the following research questions to conduct this comparative study:

1. Amongst those attributes that are common to both MOOCs, which are the most valuable with regards to the prediction of attrition?
2. Is the most valuable attribute in predicting attrition for the FutureLearn (FL) MOOC different from the one for the edX MOOC?

In pursuing these questions, it was important to use a well-performing machine-learning algorithm (in terms of the accuracy of the prediction) for both MOOCs. In particular, it was important to establish how soon it is possible to make a reasonably accurate prediction of attrition within each MOOC. More specifically, in what week (out of the total length of the course) are the predictions sufficiently accurate to be useful for each of the case study courses.

5 Methodology

5.1 Course selection

Firstly, we selected suitable case study courses from those available in FutureLearn and edX, delivered by the collaborating institutions (to ensure student datasets would be readily available). The criteria used in the manual selection of these courses included: they should be of a similar discipline or theme in the broadest possible way (i.e. either from STEM subjects or social sciences), and of a similar duration. In the case of there being more than one matching pair, we would give preference to those for which the duration is the longest. If more than one "run" of these courses had data available, we would select those for which the cohorts were the largest.

After applying the above criteria to the courses available (a total of 22: 8 in edX and 14 in FutureLearn), we selected the following: the 7-week-long edX course in Spanish history titled "The Spain of Don Quixote"[2] (Quijote501x) and the 6-week-long in archaeology titled "Archaeology of Portus: Exploring the Lost Harbour of Ancient Rome"[3] (Portus). We refer to these courses in the rest of this paper as the edX MOOC and FL MOOC respectively. Both courses had a certification available to those learners who meet the platform completion criteria, and both have some sort of assessment (exercises and quizzes, respectively).

---

[2] `https://www.edx.org/course/la-espana-de-el-quijote-uamx-quijote501x-0`

[3] Archaeology is regarded as being on the intersection of science and humanities (`https://www.futurelearn.com/courses/portus/4/steps/76822`). However, as the humanities element of the course is history, we felt this discipline is sufficiently close to that in the Quijote 501x course, and that therefore the Portus course would attract learners of not too dissimilar interests and backgrounds.

5.2 Attribute engineering

For both cases, in a pre-processing stage, we computed the value of a number of attributes from raw data available in each context, such as the number of sessions, the number of days, the number of events, total time spent on exercises/quizzes and the number of social interactions in discussions forums). Following this, we applied machine learning techniques to the pre-processed data to predict attrition levels for each dataset (the algorithms mentioned previously in the "Machine learning" subsection). The prediction is performed looking forward to the week ahead, and it becomes more accurate when there is more information available, as would be expected. However, we have established the point by which an early warning could be provided with a reasonably high degree of accuracy for each case. This was interesting to ascertain as there is a clear trade-off between accuracy and timeliness of the prediction: clearly there is less value of a dropout prediction after the student has already left the course, whereas a timely prediction (even if less accurate) could enable an intervention which might help the student to continue).

5.3 Datasets description

The anonymised datasets for each MOOC were processed using an adaptation (of the early stages) of Jo *et al*'s pipeline for expediting learning analytics [25], as follows:

3. datasets were pre-processed and cleaned;
4. attributes used as predictors were extracted; and
5. a number of predictive models were generated.

FL MOOC Dataset.

The FutureLearn MOOC selected has been offered four times to date as shown in Table 2. Rather than aggregating the four datasets, we opted for selecting the offering (run) with the highest number of learners eligible for certification as this would be the least imbalanced dataset of those available (however, due to the "funnel of participation" effect [11], this cannot be completely avoided).

Therefore, in the selected FL MOOC dataset there was data from 8935 enrolled learners, from which 3646 learners were actively involved in the course content. Of all the students, only 1843 engaged as social learners (typically posting comments, but also through "likes" as in social media). A total of 2075 completed at least 50% of the learning activities and thus were eligible to receive a certificate.

Table 2. Statistics of all the offerings (runs) to date of the FutureLearn MOOC on Portus.

Run	Start date	Enrolled	Active learners	Social learners	Eligible for certificate
1	May 2014	7779	4637	1843	2075
2	January 2015	8935	3646	1300	1589

Run	Start date	Enrolled	Active learners	Social learners	Eligible for certificate
3	June 2015	3256	1231	360	417
4	June 2016	5177	2011	751	707

The course runs for 6 weeks, during which a number of learning activities are presented (videos, articles, exercises, discussions, reflections and quizzes as mentioned earlier). The results of the assessment (in quizzes specifically) are shared with the learner (and recorded) but the actual results do not affect the eligibility for the certificate, as this is based on only completion of activities (as explained above).

edX MOOC Dataset.

The edX MOOC selected has been offered three times to date. For consistency, we also selected the offering with the highest number of learners, which was also on its first delivery (February to May 2015): a total of 3530 learners enrolled in the edX-MOOC, from which 1718 students were actively involved in the course content. Of all the students, only 423 engaged in some activity or viewed multimedia content over the last week. A total of 164 obtained a grade of more than 60% and thus received a certificate.

The length of the course is seven weeks. In addition to the discussion forum every week there are multimedia resources, both in text files and in video formats, and practical activities without evaluation. Each week ended with an evaluation activity that is a test of 21-23 questions. Each weekly evaluation contributed 14% of final grade for the course.

Similarly to FutureLearn, edX stores all learners' events. There is one file per day with the events that happened. Each event has a category. The most common ones are: navigation events, video interaction events, assessment interaction events, discussion forum events.

FL MOOC vs edX MOOC.

As presented before, each MOOC platform creates different type of learners' events that are relevant according to the philosophy behind their MOOC approach. As a result, there are a potentially large number of attributes that could be analysed if studying attrition separately for each of these contexts. However, in order to facilitate a meaningful comparison between both approaches, only the intersection of the attributes from the available data was considered. The following is the list of attributes known for both datasets:

- `number_sessions`: total number of sessions in the course. This was important to calculate as neither platform provides such data. In determining a session within the edX MOOC, an inactivity threshold was established, if the elapsed time between two consecutive interactions of the learner exceeds the threshold, these interactions were assumed to have taken place in two separated sessions as the

learner was considered not to have been active during this time. For the FL MOOC the start time of a given activity (step) is only recorded the first time the learner accesses the given step, so the calculation of the inactivity threshold was performed slightly differently (taking into account the finishing time of a previously finished activity instead) but applying similar principles.

- `number_comments`: total number of social interactions (comments and replies) in the course.
- `total_time`: total time invested in the course (inactivity periods aside). More specifically, it is defined as the aggregate of the elapsed time between the access to each problem or exercise and the submission of the corresponding attempted solution (calculated individually in the case of there being several attempts). As before, an inactivity threshold is applied, considering the student not active if the elapsed time between getting the problem and the problem submission exceeds the threshold.
- `time_problems`: total time invested in answering exercises (assessments).

These attributes were calculated for each week and each learner. The aim of the formulation of our predictive models was to detect those learners which are eligible for a certificate. In the case of FutureLearn learners, they need to complete at least 50% of the course activities (regardless of assessment performance), whilst edX learners need to obtain more than 60% marks in the assessments to obtain a certificate (regardless of participation). The dependent attribute in both cases was to detect whether the learner would obtain a certificate.

5.4 Machine learning

In recent years there has been significant increase in development and the availability of data analytics tools. This includes readily available, complex machine learning algorithms, toolkits such as WEKA, and domain-specific packages and libraries for programming languages such as python and R [4] (amongst others inventoried by Slater *et al.* [26]). Tools such as these facilitate the development of dedicated software and faster generation of learning analytics. Specifically in this study, the following machine learning algorithms were compared: generalised boosted regression models, kNN, boosted logistic regression and extreme gradient boosting.

Generalised Boosted regression Models (GBM) The GBM is a boosting algorithm, similar to AdaBoost, which can be used for multi-class regression problems. GBM was first proposed by Freund and Schapire [28], and improved by Friedman [29] and is available in R in the package `gbm`.

[4] In addition to algorithms, the R package `caret` generates Receiving Operator Curves (ROC), e.g. those in Appendix A (**Fig. 7** and **Fig. 7**), and perform Area Under the Curve analyses (AUC).

Weighted k-Nearest Neighbours (kNN) The kNN makes use of simple heuristics of distance (or similarity sampling) to perform the classification [30], and is available in R in the package `kknn`.

Boosted Logistic Regression (LogitBoost) `LogitBoost` is losely related to the Support Vector Machine [31], it is a popular binary predictor that is available in R. Also referred to as LogReg in this paper.

eXtreme Gradient Boosting (XGBoost) Though it is related to the GBM (also a boosting algorithm), this algorithm can generate decision trees which are human-readable models. It has a good performance as it includes an efficient linear model solver and can also exploit parallel computing capabilities [32]. It is available in R in the package `xgboost`.

6 Analysis and Discussion

Using these datasets, we implemented four classification models that have been extensively tested and shown to generate good classification results. We used the machine learning algorithms presented in subsection 5.4.

Due to the content of the MOOCs being organized in weeks, we calculated weekly models of each course, in a similar way as Kloft *et al.* 20. In order to measure the performance of these models we have used the area under the ROC curve (AUC) metric [5]. (See Appendix A for the full set of ROC curves for all the models using both datasets.) These measurements helped us to select the machine learning algorithm that is best suited for each context. Then, taking into account these previous selections, we took the approach of finding out the best attributes for each MOOC.

6.1 FL MOOC Dataset

Firstly, we compare the performance of the four machine learning algorithms mentioned.on the FL MOOC dataset. The two best-performing algorithms (with regards to the AUC metric) are GBM and XGBoost, (although the difference between them is negligible) (see **Fig. 1**).

However, when considering the performance in computational time, XGBoost is faster than GBM by an order of magnitude (during the training phase) as shown in **Fig. 2**. Due to this comparative advantage, we consider the XGBoost as the best algorithm amongst those tested on this dataset.

[5] Note that in this context, execution time per model is not relevant, given that the predictions are not required tocalculated in real time (can be calculated in real time.daily processes, once data is updated). Therefore, a "poor" performance in this metric is much more importantless indicative of the goodness of the model than the accuracy as reported by the AUC metric.

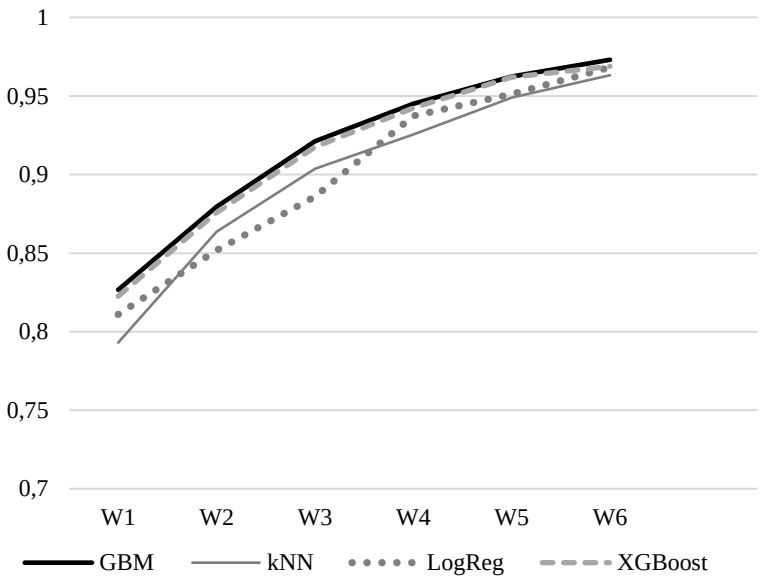


Fig. 1. Performance results for the FL MOOC dataset in terms of AUC metric for the models for each week

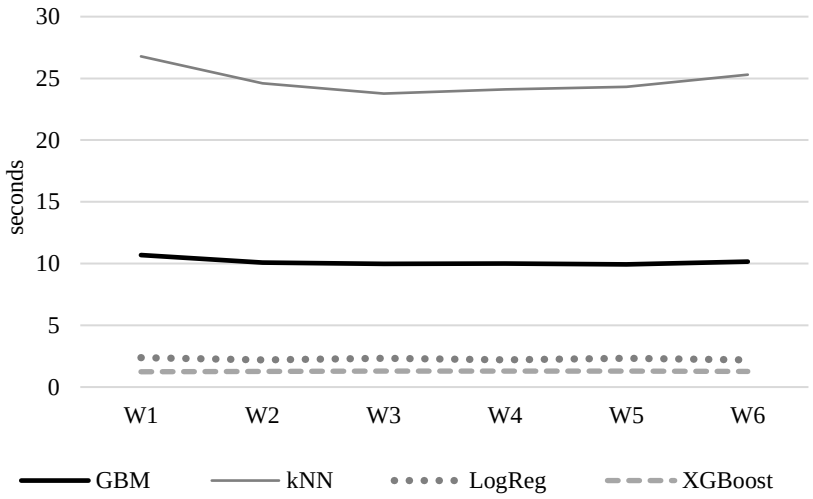


Fig. 2. Training time for the machine learning algorithms benchmarked on the FL MOOC

Once a machine learning algorithm was selected, we studied the varying importance of the attributes throughout the duration course for this algorithm (see **Fig. 3**). This "importance" suggests the predictive value of each attribute at a given week.

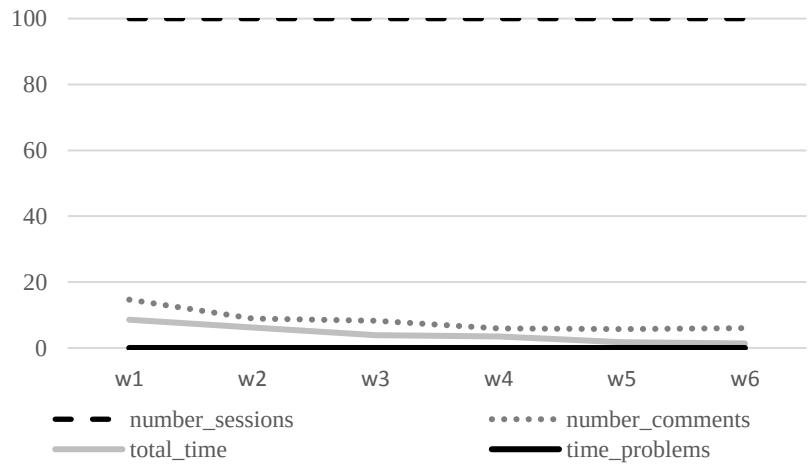


Fig. 3. Evolution of the attribute importance for XGBoost using the FL MOOC dataset

During all the weeks, the most relevant attribute is `number_sessions`; however, the attribute related with social interactions is the second most relevant one.

6.2 edX MOOC Dataset

As before, we first compare the performance results of the four mentioned machine learning algorithms (see **¡Error! No se encuentra el origen de la referencia.**). As in the case of the FL MOOC dataset, the best performing algorithm for the edX MOOC dataset is GBM, though in this case the difference is significant. Finally, we studied the importance of the attributes throughout the course for this algorithm (see **Fig. 5**).

From the start of the course, attributes `number_sessions` and `total_time` are the most valuable for the prediction models. However, from the end of fifth week the most reliable attribute is `time_problems`. We found that in this course, which follows an x-MOOC approach, the attribute related to social interactions (`number_comments`) did not contribute to the prediction.

6.3 Discussion

Each of the machine learning algorithms benchmarked provided good results for both scenarios; however their performance varied in terms of accuracy. GBM is the best one for both approaches from the beginning to the end of the courses; and XGBoost is the second best for the FL MOOC throughout the course and for the edX MOOC after the third week. Based on these results, we selected the XGBoost algorithm for the FL MOOC and the GBM for the edX MOOC.

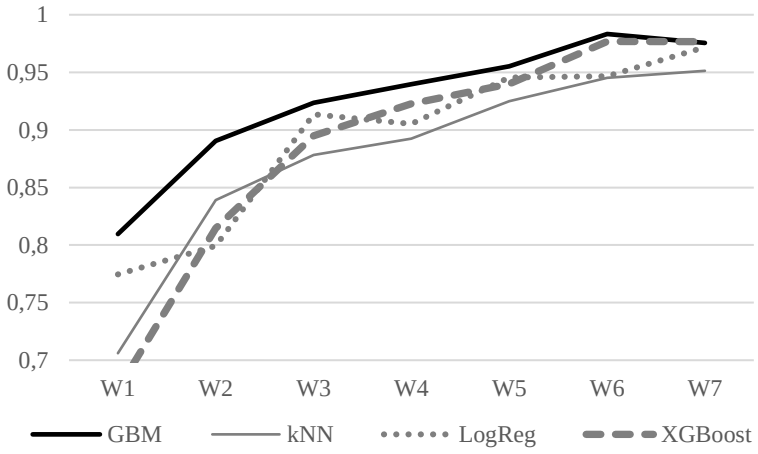


Fig. 4. Performance results for the edX MOOC dataset in terms of AUC metric for the models for each week

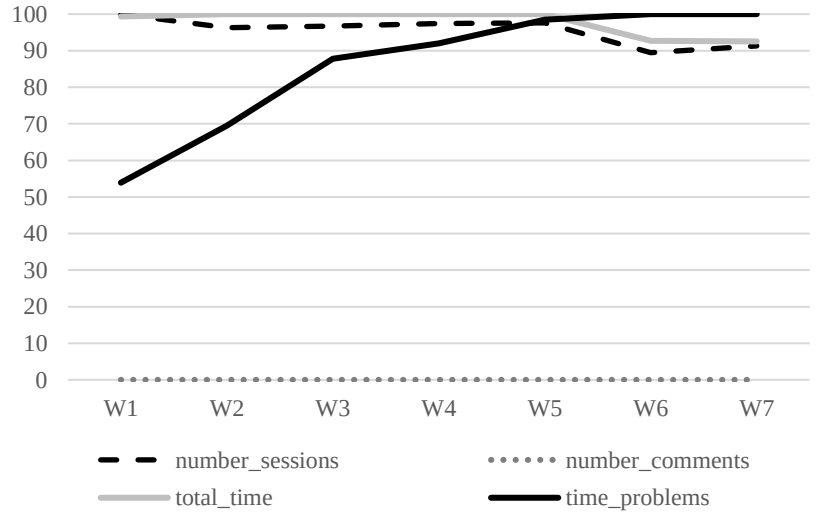


Fig. 5. Evolution of the attribute importance for GBM using the edX MOOC dataset

Once the algorithms were selected, we studied the importance of the attributes in both courses. On the one hand, the most relevant attribute along the duration of the FL MOOC was `number_sessions` and `number_comments` was the second relevant attribute especially during the first weeks of the course. Results confirm that the progression dedicated in the course is the important issue because the most relevant attribute

was the number of session in the course. Moreover, social interactions also have some importance.

On the other hand, the most relevant attributes for the edX MOOC, were `total_time` and `time_problems`. The `total_time` attribute was the most relevant until the fifth week and after that the most relevant one was `time_problems`. These results confirm that in courses such as this, it is important to dedicate time to learning the course content and also to undertake the assessments.

Table 3. Summary of the obtained results connected with the proposed research questions.

Research Question	**FL MOOC**	**edX MOOC**
Most valuable attributes	`number_sessions` `number_comments`	`total_time` `time_problems`
Earliest time for accurate certificate eligibility prediction (Week/Total (%))	3/6 (50%)	3/7 (43%)

Finally, we were interested in knowing how soon it is possible to have a reasonably accurate prediction of attrition. In the case of the FL MOOC, the baseline accuracy of the predictor that classifies learners that do not complete 50% of the course is 0.91. In the case of the edX MOOC, the baseline accuracy of the predictor that classifies non-certificate earners is 0.90.

7 Conclusions and Future Work

A study has been carried out to investigate whether the inherent similarities and differences between the affordances provided by MOOC platforms may influence learner behaviours, such as engagement towards certification or dropout. More specifically, machine learning algorithms were applied to prediction attrition in MOOCs that have been delivered in two different platforms. To this end, we selected an edX MOOC and a FutureLearn MOOC of comparable structure and themes, and selected observable factors that can be used as an early predictor for attrition in these cases.

Common attributes to these cases were identified and the most valuable of these (with regards to the attrition prediction) were used. For both datasets we extracted the following comparable attributes: `number_sessions`: total number of sessions in the course; `number_comments`: total number of social interactions (comments and replies) in the course; `total_time`: total time invested in the course and `time_problems`: total time invested in answering exercises or quizzes (assessments).

Next, we generated several predictive models to detect (for a FL MOOC) that the students could complete at least 50% of the course, and (for an edX MOOC) that they could obtain a grade of more than 60% and hence a certificate. The key attribute in both cases was to predict whether the learner would obtain a certificate.

These predictive models were generated using these four machine learning algorithms: k-nearest neighbours (kNN), gradient boosting machine (GBM), extreme gradient boosting (XGBoost) and boosted logistic regression (logitboost). Due to the content of the MOOCs being organized in weeks, we calculated a model per week of each course.

From those tested, the best machine learning algorithms for both the edX MOOC and the FL MOOC are GBM and XGBoost. However, the relevant attributes were different for each course. In the FL MOOC the most important ones were `number_sessions` and `number_comments`, both related with the connectivism paradigm: as expected, learners who engage more in activities facilitating connections with others and with knowledge itself would do better than those who do not. In contrast, for the edX MOOC the most important attributes were `total_time` and `time_problems`, which is consistent with the pedagogical design of instructivist courses, where learners devoting time to learning activities gain more from these, and therefore do better than learners who do not.

The predictive models offered a reasonably accurate prediction of attrition within each MOOC approach (over 90% accurate prediction, available approximately halfway through the course delivery).

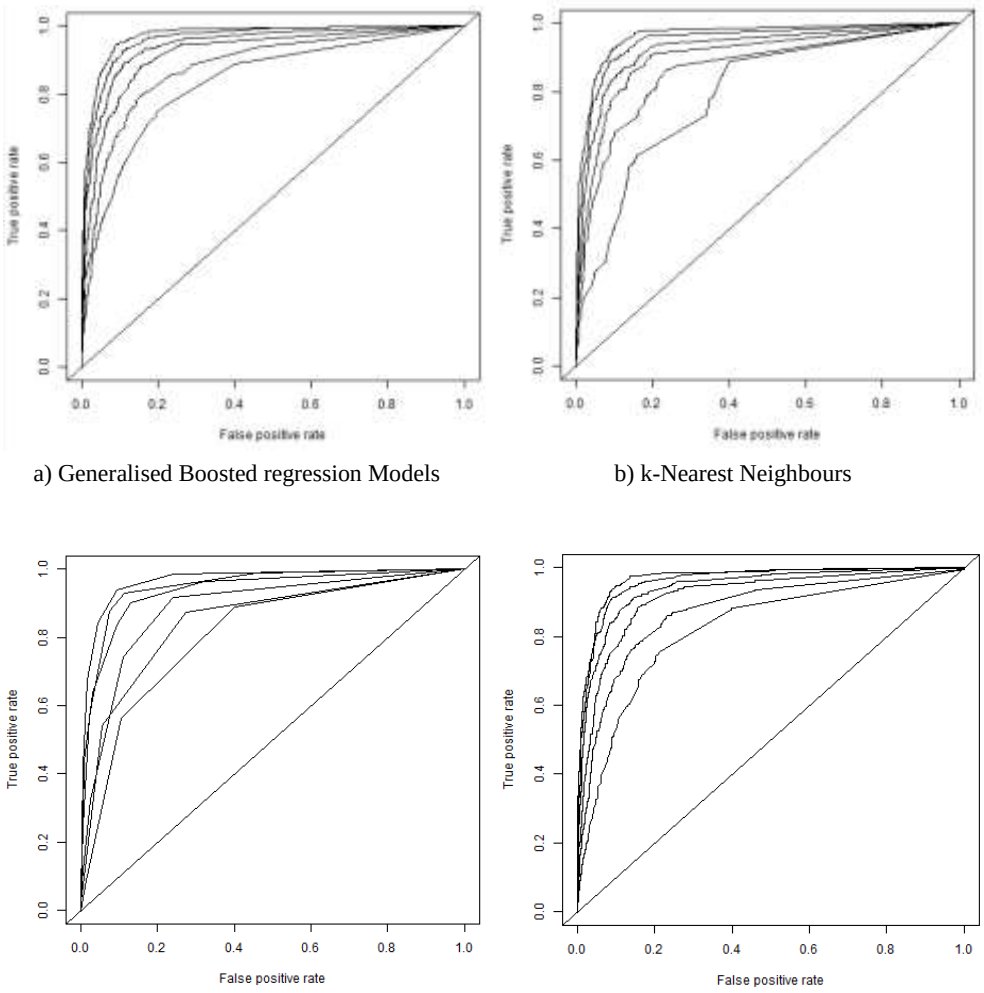
As future work, more case studies could be added to this study. On the one hand, taking into account more deliveries of the studied courses and, on the other hand, including courses from other disciplines. Finally, we are planning the generation of warning systems that may automatically warn the student at risk of not obtaining a certificate. However, more work in understanding attrition and learner dropout models is required still to be able to measure the impact of such interventions.

Acknowledgements

This work has been partially funded by the Madrid Regional Government with grant No. S2013/ICE-2715, the Spanish Ministry of Economy and Competitiveness project "Flexor" (TIN2014-52129-R), and by the Web Science Institute Stimulus Fund 2015/16 Project “The MOOC Observatory Dashboard: Management, analysis and visualisation of MOOC data”. The authors are grateful to Darron Tang and Jasmine Chen from the University of Southampton for their work in this project during their summer internships.

Special thanks are due to the Universidad Autónoma de Madrid and the University of Southampton for the support of this inter-institutional collaboration and facilitating the access of the data of the respective MOOCs on which this study is based.

A Appendix



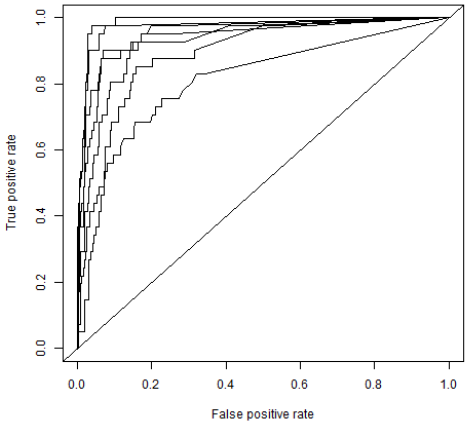
a) Generalised Boosted regression Models

b) k-Nearest Neighbours

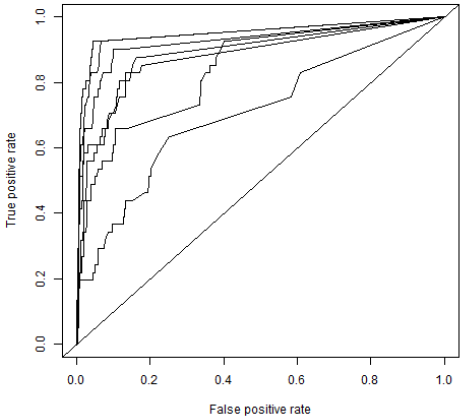
c) Boosted Logistic Regression

d) XGBoost

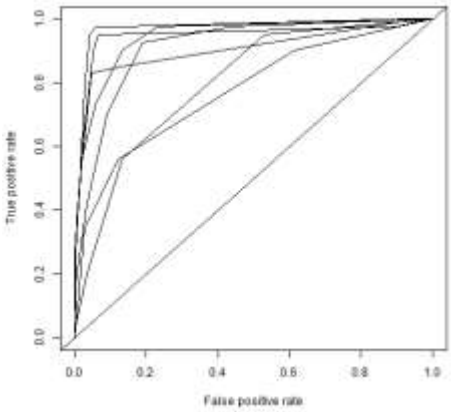
Fig. 6. ROC values for all weeks prediction models varying the algorithm on the FutureLearn MOOC dataset



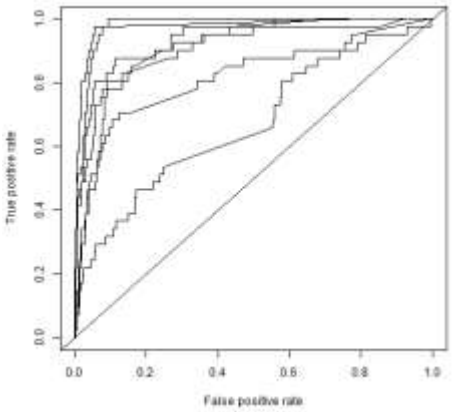
a) Generalised Boosted regression Models



b) k-Nearest Neighbours



c) Boosted Logistic Regression



d) XGBoost

Fig. 7. ROC values for all weeks prediction models varying the algorithm on the edX dataset

References

1. Dhawal Shah. 2016. MOOC aggregator Class Central. Available at `https://www.class-central.com/report/mooc-course-report-october-2016/`
2. George Siemens. 2005. Connectivism: A learning theory for the digital age. *International journal of instructional technology and distance learning*, 2(1) 3–10. `http://er.dut.ac.za/handle/123456789/69.`

3. Richard Fox. 2010. Constructivism Examined. *Journal Oxford Review of Education*. 27(1). 23-35. http://dx.doi.org/10.1080/03054980125310.
4. C. Osvaldo Rodríguez. 2012. MOOCs and the AI-Stanford like courses: Two successful and distinct course formats for massive open online courses. In *European Journal of Open, Distance and E-Learning*, 15(2) 1–13. http://www.eurodl.org/index.php?p=archives&year=2013&halfyear=2&article=516.
5. Abraham Anders. 2015. Theories and Applications of Massive Online Open Courses (MOOCs): The case for hybrid design. In *International Review of Research in Open and Distributed Learning*, 16(6). http://dx.doi.org/10.19173/irrodl.v16i6.2185.
6. Antonio Fini, Andreas Formiconi, Alessandro Giorni, Nuccia Silvana Pirruccello, Elisa Spadavecchia, and Emanuela Zibordi. 2008. IntroOpenEd 2007: an experience on Open Education by a virtual community of teachers. In *Journal of e-Learning and Knowledge Society*. http://www.jelks.org/ojs/index.php/Je-LKS_EN/article/view/266/248.
7. Rebecca Ferguson and Mike Sharples. 2014. Innovative Pedagogy at Massive Scale: Teaching and Learning in MOOCs. In *9th European Conference on Technology Enhanced Learning (EC-TEL 2014)*, Graz, Austria. LNCS 8719:98-111. Springer. http://oro.open.ac.uk/40787/
8. Rebecca Ferguson. 2012. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6). 304–317. http://oro.open.ac.uk/36374.
9. Vincent Tinto and John Cullen. 1975. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, 45(1) 89-125. Retrieved from https://www.jstor.org/stable/1170024.
10. Phil D. Long and George Siemens. Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE review, September/October 2011. http://er.educause.edu/~/media/files/article-downloads/erm1151.pdf.
11. Doug Clow. 2013. MOOCs and the funnel of participation. In *Proceedings of the 3rd Conference on Learning Analytics and Knowledge (LAK2016)*. 8-12 April, Leuven, Belgium. 185-189. http://oro.open.ac.uk/36657/
12. René F Kizilcec, Chris Piech, Emily Schneider. 2013. Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In *Proceedings of the 3rd Conference on Learning Analytics and Knowledge (LAK2016)*. 8-12 April, Leuven, Belgium. 170-179. http://dx.doi.org/10.1145/2460296.2460330.
13. Daphne Koller, Andrew Ng, Chuong Do, and Zhenghao Chen, Retention and intention in massive open online courses: In depth. *Educause Review*, 48(3).

62–63, June 2013. http://er.educause.edu/articles/2013/6/retention-and-intention-in-massive-open-online-courses

14. Adriana Wilde, Ed Zaluska, and David Millard. 2015. Student success on face-to-face instruction and MOOCs. In *Web Science Education: Curriculum, MOOCs and Learning. WEB SCIENCE 2015*. Oxford, UK. http://eprints.soton.ac.uk/377682/
15. Dragan Gašević, Shane Dawson, Tim Rogers, and Danijela Gasevic. 2016. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting learning success. *The Internet and Higher Education*. 28, (2016), 68–84.
16. Carolyn Penstein Rosé, Ryan Carlson, Diyi Yang, Miaomiao Wen, Lauren Resnick, Pam Goldman, and Jennifer Sherer. 2014. Social Factors that Contribute to Attrition in MOOCs. In *Proc. of the 1st ACM Conference on Learning at Scale (L@S)*, Atlanta, 2014. http://dx.doi.org/10.1145/2556325.2567879.
17. Sandeep M. Jayaprakash, Erik W. Moody, Eitel J.M. Lauría, James R. Regan, and Joshua D. Baron. 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics*. 1, 1 (May 2014), 6–47. https://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/3249/4011
18. Hanan Khalil and Martin Ebner. 2014. MOOCs completion rates and possible methods to improve retention –a literature review. In *EdMedia:World Conference on Educational Multimedia, Hypermedia and Telecommunications*, AACE. 1236-1244.
19. Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, Carolyn Rosé. 2015. Exploring the Effect of Confusion in Discussion Forums of Massive Open Online Courses. In *Proc. of the 2nd ACM Conference on Learning at Scale (L@S)*, Vancouver, 2015. 121-130. http://dx.doi.org/10.1145/2724660.2724677.
20. 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).
21. Marius Kloft, Felix Stiehler, Zhilin Zheng, Niels Pinkwart. 2014. Predicting MOOC Dropout over Weeks Using Machine Learning Methods. *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 60–65, October 2014, Doha, Qatar. Retrieved from http://www.aclweb.org/anthology/W/W14/W14-41.pdf.
22. Anne Lamb, Jasha Smilack, Andrew Ho, and Justin Reich. 2015. Addressing Common Analytic Challenges to Randomized Experiments in MOOCs: Attrition and Zero-Inflation. In *Proc. of the 2nd ACM Conference on Learning@ Scale (L@S)*. 21-30. ACM. http://dx.doi.org/10.1145/2724660.2724669.

23. Abelardo Pardo, Jelena Jovanovic, Negin Mirriahi, Shane Dawson, Roberto Martínez-Maldonado, Dragan Gašević. 2016. Generating Actionable Predictive Models of Academic Performance. In *Proceedings of Learning Analytics and Knowledge (LAK2016)*. 25-29April, Edinburgh, United Kingdom. 185-189. http://dx.doi.org/10.1145/2883851.2883870.

24. Iván Claros, Ruth Cobos, Gabriela Sandoval, and Mónica Villanueva. 2015. Creating MOOCs by UAMx: experiences and expectations. *The 3rd European MOOCs Stakeholders Summit (eMOOCs 2015)*: 61-64.

25. Adriana Wilde, Manuel León-Urrutia, Su White. 2016. Tracking collective learner footprints: aggregate analysis of MOOC learner demographics and activity. In *Proceedings of the 9th International Conference of Education, Research and Innovation (iCERi)*. Seville, Spain. November 2016. https://dx.doi/10.21125/iceri.2016.1319

26. Yohan Jo, Gaurav Tomar, Oliver Ferschke, Carolyn Penstein Rosé, and Dragan Gašević. 2016. Pipeline for expediting learning analytics and student support from data in social learning. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge (LAK '16).* 542-543. http://dx.doi.org/10.1145/2883851.2883912.

27. Stefan Slater, Srećko Joksimović, Vitomir Kovanović, Ryan Baker, and Dragan Gašević. 2016. Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*. http://dx.doi.org/10.3102/1076998616666808.

28. Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119-139. Retrieved from http://www.face-rec.org/algorithms/Boosting-Ensemble/decision-theoretic_generalization.pdf.

29. Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, Vol. 29, No. 5 (Oct., 2001), pp. 1189-1232. http://www.jstor.org/stable/2699986

30. J.M. Keller, M.R. Gray and J.A. Givens. 1985. A fuzzy k-Nearest Neighbour algorithm. *IEEE Transactions on Systems, Man and Cybernetics* 15(4) 580-585. http://dx.doi.org/10.1109/TSMC.1985.6313426.

31. Thomas P. Minka. 2003. Algorithms for maximum-likelihood logistic regression http://www.stat.cmu.edu/tr/tr758/tr758.pdf

32. Tianqi Chen, Tong He and Michael Benesty. 2016. Package 'xgboost': Extreme Gradient Boosting. Documentation available in https://cran.r-project.org/web/packages/xgboost/xgboost.pdf