Comparing attrition prediction in FutureLearn and edX MOOCs

Ruth Cobos
Computer Science and Engineering
Universidad Autónoma de Madrid, Spain
ruth.cobos@uam.es

Adriana Wilde
Electronics and Computer Science
University of Southampton
United Kingdom
agw106@ecs.soton.ac.uk

Ed Zaluska
Electronics and Computer Science
University of Southampton
United Kingdom
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.

Author Keywords

MOOCs, predictive model, learning analytics, attribute selection, FutureLearn, edX.

ACM Classification Keywords

• **Applied computing~Education~Interactive learning environments. • Social and professional topics~Informal education** • *Human-centered computing~Collaborative and social computing systems and tools* • Computing methodologies~Feature selection

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, and in tackling this diversity, it helps to understand the principles and affordances given by platforms used by FutureLearn respect to another well-recognised MOOC provider who offer exemplar courses which could be used for a comparative study (such as edX).

Against this background, we investigated whether the inherent similarities and differences between the affordances provided by various MOOCs platforms (FutureLearn and edX respectively) may influence learner behaviours (assuming all other things equal) and whether there is an observable factor that can be used as an early predictor for attrition in either case. This is especially valuable as it could be used to inform interventions intending to improve learners' performance in future courses.

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

POSITIONING FUTURELEARN COURSES

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 have become to be known as c-MOOCs, of which the first one was delivered in 2008 by Siemens and Downes, the latter of whom coined the term [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 became to be known as x-MOOCs, a term coined by Downes in 2012 to differentiate them from his c-MOOCs. The first of these courses was delivered in 2007 though: the *Introduction to Open Education*, by David Wiley from Utah [6].

Characteristic	**c-MOOCs**	**x-MOOCs**
Number of learners	Should scale to large numbers	Should scale to large numbers
Method of delivery	Online	Online
Communication tendencies	Distributed	Centralised
Related learning theory	Connectivism	Instructivism
Design should *primarily* support	Creation of *connections* between learners, resources and ideas	Relationship between teachers and learners, mediated through *task* completion
First MOOC delivered (with year)	Connectivism and Connective Knowledge (2008)	Introduction to Open Education (2007)

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

Noting that there are many similarities as well as important differences between these paradigms of online learning (summarised in Table 1), it is interesting to compare them in practice through the analysis of case study courses. 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, which are called "steps". Each step has a learning object belonging to a prescribed category, such as: videos, articles, exercises, discussions, reflections, quizzes and peer reviews. For each step, learners are able to leave comments, each of these in turn can be visibly "liked" (as in social media platforms) and have replies or follow-up comments, allowing learners to build connections amongst the community and with the presented learning objects, as often these comments allow for their personal reflections and expressions of their own understanding (or lack thereof). This architecture reflects FutureLearn's pedagogical underpinnings inspired in social constructivism and Laurillard's conversational framework [7]. As explained before, under this paradigm, 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).

Similarly to FutureLearn courses, edX courses also 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 html pages or interactive educational resources. All courses have an online discussion forum where students can post and review questions and comments to each other and teaching assistants. edX courses can be categorized as x-MOOCs, falling under the instructivist paradigm 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 completed to satisfaction) but as it is collected, it has the potential to be further analysed to predict whether the learner would be eligible for a certificate at the end of the course.

In both cases, the platform data is collated by the 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 data analytics (discussed in the *Learning Analytics* section, below), however, differences in the technical implementations of already fundamentally different approaches also present additional challenges in aligning the collected data and perform a suitable comparison study.

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 dropout rates for many years [9]. However, the advent of Massive Open Online Courses (MOOCs) 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, informing and transforming teaching, with 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 their 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 performs a cluster analysis of the disengaged learners in a study that was one of the first to showcase the potential of analytics in understanding dropout. Through this work, the authors were able to identify prototypical learner trajectories (auditing, completing, disengaging, and sampling) as an explanation of learners' 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], as it is a well-understood metric of engagement that is still useful while an improved metric for success is defined. 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 providing actionable predictions to teachers so that they can provide timely feedback [14], [19].

Machine learning

In recent years there has also been an explosion of tools for data analytics, with complex machine learning algorithms now readily available. These include toolkits such as WEKA, to domain-specific packages and libraries for programming languages such as python and R (amongst many others inventoried by Slater *et al.* [20]). Tools such as these facilitate the development of dedicated software and the faster generation of learning analytics.

Amongst the large number of machine learning algorithms available, the following are considered in this paper: GBM, kNN, LogReg and XGBoost.

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 Shapire [20], and is available in R in the package `gbm`.

Weighted k-Nearest Neighbours (kNN)

The kNN, makes use of simple heuristics of distance (or similarity sampling) to perform the classification [21], and is available in R in the package `kknn`.

Logistic Regression (LogReg)

Closely related to the Support Vector Machine [22], it is a very popular binary predictor that is available in R in the package `LOGIT`.

eXtreme Gradient Boosting (XGBoost)

Though it is related to the GBM (also a boosting algorithm), it can generating decision trees which are human-readable models together with a good performance as it includes an efficient linear model solver and can exploit parallel computing capabilities [23]. Available in R in the package `xgboost`.

CONTEXT OF THE PRESENT APPROACH

As explained earlier, the main motivation for this research study was to investigate similarities and differences between FutureLearn courses and comparable ones in other platforms such as edX, specifically in relation to their attrition levels. The authors are associated to two institutions, each delivering MOOCs through each platform, which facilitated such a comparative study. The institutions are the University of Southampton (UoS) and the Universidad Autónoma de Madrid (UAM)

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], whilst the UAM became a member of the edX consortium in 2014 and currently offers eight MOOCs at edX [24].

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 predictive attribute for the FL MOOC different from the one for the edX MOOC?

In pursuing these questions, it was important to identify a well-performing machine-learning algorithm (in terms of the accuracy of the prediction) for both MOOCs. Also of interest is 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 for each of the case study courses.

METHODOLOGY

Course selection

We selected suitable case study courses from those available in FutureLearn and edX, delivered by the collaborating institutions (for which datasets were readily available). The criteria used in the selection of these courses included: they needed to belong to the same broad discipline (i.e. either from STEM or social sciences), and have a similar duration. When there was more than one matching pair, we gave preference to those for which the duration was the longest. If more than one "run" of the thus selected courses had available data, we would select those for which the cohorts were the largest.

After applying the above criteria, we selected the FutureLearn course in archaeology[1] titled "Archaeology of Portus: Exploring the Lost Harbour of Ancient Rome" (Portus) and the edX course in Spanish history titled "The Spain of Don Quixote"[2] (Quijote501x). We refer to these courses in the rest of this paper as edX MOOC and FL MOOC respectively. Both had a certification available to those learners who meet the platform completion criteria, and are assessed (via exercises and quizzes, respectively).

Attribute engineering

During pre-processing, we computed the value of a number of attributes from raw data available in each course, such as the time spent on exercises/quizzes, the numbers of sessions, days, events and social interactions in discussions forums. We then applied machine learning algorithms on the pre-processed data to predict attrition levels for each dataset (those mentioned in the "*Machine learning*" subsection). The prediction was performed looking forward to the week ahead, increasing accuracy when more information is available, as expected. However, we were able to establish the point by which an early warning could be given for each case. Note that there is a trade-off between accuracy and timeliness of the prediction: clearly a dropout prediction is of no value once the student has left the course, whereas a timely prediction (even if potentially less accurate) would inform an intervention which, in turn, could prevent the dropout event.

Datasets description

The anonymised datasets for each MOOC were processed them using an adaptation (of the early stages) of Jo *et al*'s pipeline for expediting learning analytics [26], as follows:

1. datasets were pre-processed and cleaned;
2. attributes used as predictors were extracted; and
3. 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 course with the highest number of learners who became eligible for certification as this would be the least imbalanced dataset of those available (however, due to the "funnel of participation" effect [11], this is unavoidable altogether).

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
3	June 2015	3256	1231	360	417
4	June 2016	5177	2011	751	707

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

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.

The course runs for six 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 to the certificate, as this is based on completion of activities instead, as explained above.

edX MOOC Dataset

Conversely, a total of 3530 learners enrolled for 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. Students engaged in a discussion forum and were presented multimedia resources and practical non-assessed activities. Additionally, each week ended with an assessment activity: tests comprising 21 to 23 questions. Each weekly evaluation contributed to 14% of the final course marks.

Similarly to FutureLearn, edX stores all learners' events. There is one file per day with the events that happened.

[1] 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.

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

Each event has a category. The most common events are related to navigation, video interaction, assessment interaction, and discussion forum participation.

FL MOOC vs edX MOOC

Each MOOC platform creates different type of learners' events that is relevant according to the philosophy behind their MOOC approach. As a result, and in order to facilitate a meaningful comparison between both approaches, we needed to look at the intersection of a potentially large number of attributes that could be engineered from the data collected. This leaves us with the following list of attributes known for both datasets:

- `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.
- `time_problems`: total time invested in answering exercises (assessments).

We calculated these attributes 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% in the assessments to earn a certificate (regardless of participation). The dependent attribute in both cases was to detect whether the learner would obtain a certificate.

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. [18]. The machine learning algorithms presented in the subsection *Machine learning* were used, resulting in four classification models which were then extensively tested as follows.

ANALYSIS OF RESULTS AND DISCUSSION

In order to measure the performance of these models we used the area under the ROC[3] curve (AUC) metric[4]. These measurements helped us to select the machine learning algorithm that is best suited per MOOC approach, on the datasets studied. Then, taking into account these previous selections, we have studied what are the best attributes per MOOC approach.

FL MOOC Dataset

We compared the performance results of the four mentioned machine learning algorithms. The two best-performing algorithms (with regards to the AUC metric) in the FL MOOC dataset are GBM and XGBoost, though the difference between them can be considered as negligible (see Figure 1).

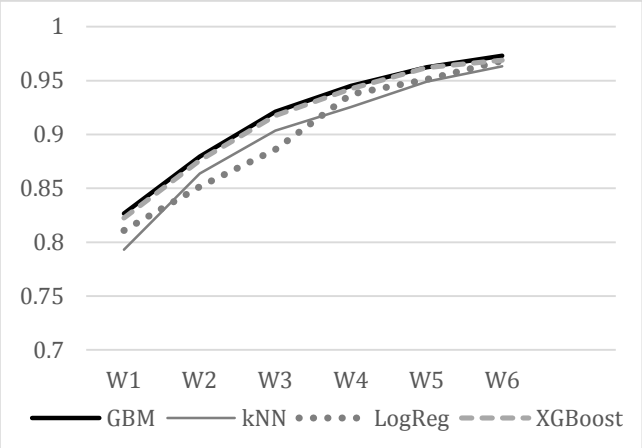


Figure 1. Performance results for the FL MOOC dataset in terms of AUC metric for the models for each week

We then studied the importance of the attributes throughout the course for XGBoost (Figure 2), which refers to the predictive value of each attribute in a given week.

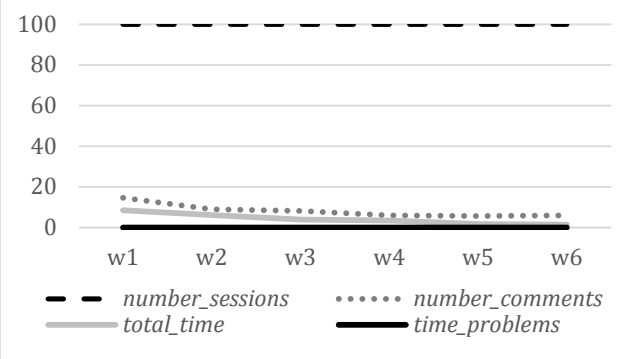


Figure 2. Evolution of attributes importance for XGBoost

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

edX MOOC Dataset

As before, firstly, we compare the performance results of the four mentioned machine learning algorithms (Figure 3).

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 more significative. Finally, we studied the importance of the attributes throughout the course for this algorithm (see Figure 4).

[3] Receiving Operator Curves (ROC) for each model per dataset were generated using the R package `caret` (see Appendix A).

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

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`) was always unimportant for the prediction.

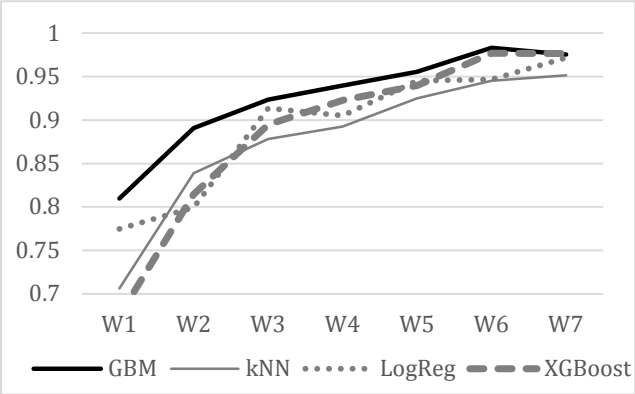


Figure 3. Performance results for the edX MOOC dataset in terms of AUC metric for the models for each week

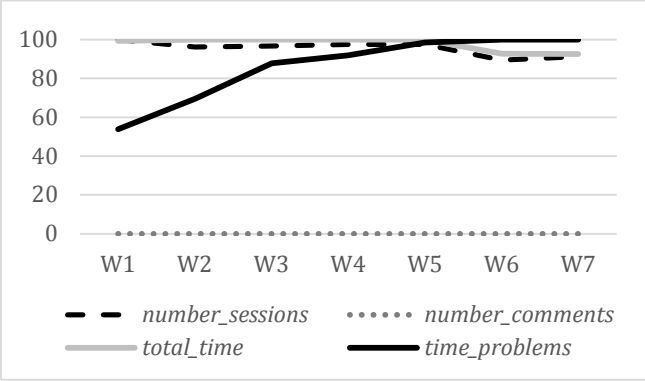


Figure 4. Evolution of attributes importance for GBM

Discussion

Each of the machine learning algorithms benchmarked provided good results for both scenarios; however their performance varied in terms of accuracy over time. GBM is the best algorithm for both scenarios from the beginning to the end of the courses; and the 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.

Once the algorithms were selected, we studied the importance of the attributes in both courses, obtaining important insights. 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 number of session in the course. Moreover, social interactions have some importance, too.

In 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 then the most relevant one was the `time_problems`. These results corroborate that in courses such as this, it is important to dedicate time learning the content of the course and to dedicate time to engage with the assessments.

Research Question	FL MOOC	edX MOOC
Most valuable attributes	`number_sessions` `number_comments`	`total_time` `time_problems`
Week/Total	3/6 (0.5)	3/7 (0.43)

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

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.

CONCLUSIONS AND FUTURE WORK

We have presented a research study where we compare the results of applying machine learning algorithms to predict course attrition in MOOCs on similar courses that have been delivered in different platforms. More specifically, we have selected an edX MOOC and a FutureLearn MOOC of comparable structure and themes.

In this research study, we identified machine learning algorithms that give a good performance (in terms of the accuracy of the prediction) for this study, in which, firstly, we sought the most valuable attributes with regards to the prediction of attrition per MOOC. Secondly, we sought to detect if a reasonably accurate prediction of attrition within each MOOC approach could be done sufficiently early.

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 in the case of a FL MOOC that the students could complete at least 50% of the course, and for an edX MOOC, that they could obtain more than 60% of the grade and a certificate. Therefore, the dependent attribute in both cases was to detect whether the learner would obtain a certificate.

These predictive models where generated with these four machine learning algorithms for implementing the models: k-nearest neighbours (kNN), gradient boosting machine (GBM), extreme gradient boosting (XGBoost) and a logistic regression (LogReg). Due to the content of the MOOCs being organized in weeks, we calculated a model per week of each course. Moreover, we measured the time in training and test phases per model.

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, for these learners, it was important to participate in activities facilitating (and reinforcing) connections with others and with knowledge itself. The most important attributes for the edX MOOC were `total_time` and `time_problems`, resonating with the intuition behind instructivist courses, where learners devoting time to learning activities gain more from these than from the connections with others.

The predictive models offered a reasonably accurate prediction of attrition from the third week onwards (approximately in the middle of the course length).

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 can automatically warn student at risk of not obtained a certificate.

ACKNOWLEDGEMENTS

This work has been funded by the Web Science Institute Pump-priming 2015/16 Project "The MOOC Observatory Dashboard: Management, analysis and visualisation of MOOC data". The authors are grateful to the interns Darron Tang and Jasmine Chen from the University of Southampton for their work in this project. Additionally, Ruth Cobos' contribution 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).

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 Rodriguez. 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. Retrieved from 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.je-lks.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. 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. Retrieved from

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] 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

[17] 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).

[18] 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.

[19] Abelardo Pardo, Jelena Jovanovic, Negin Mirriahi, Shane Dawson, Roberto Martinez-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.

[20] 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.

[21] 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.

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

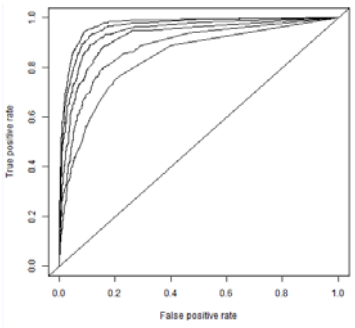
[23] Tianqi Chen, Tong He and Michael Benesty. 2016. Package 'xgboost': Extreme Gradient Boosting. Documentation available in https://cran.r-project.org/web/packages/xgboost/xgboost.pdf

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

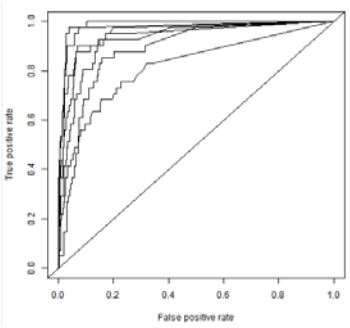
[25] Adriana Wilde, Manuel León-Urrutia, Su White. 2016. Tracking collective learner footprints: aggregate analysis of MOOC learner demographics and activity. In *Proceedings of the 9th International Conference of Education, Research and Innovation (iCERi)*. Seville, Spain. November 2016. https://dx.doi/10.21125/iceri.2016.1319

[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.

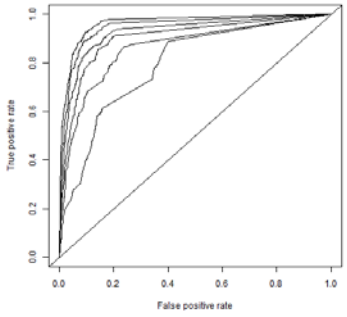
APPENDIX A



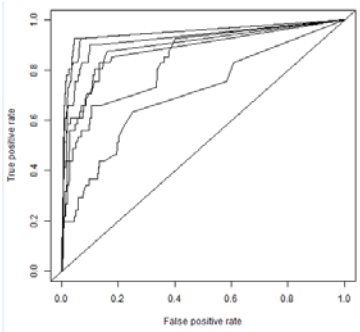
a) GBM on the FL MOOC dataset



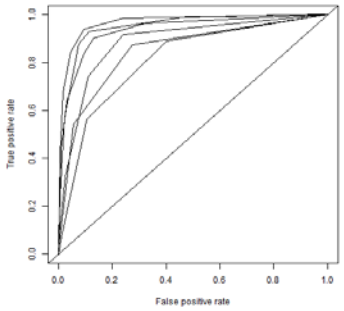
d) GBM on the edX MOOC dataset



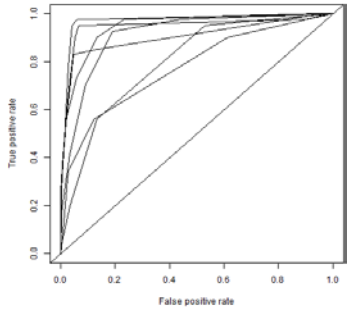
b) kNN on the FL MOOC dataset



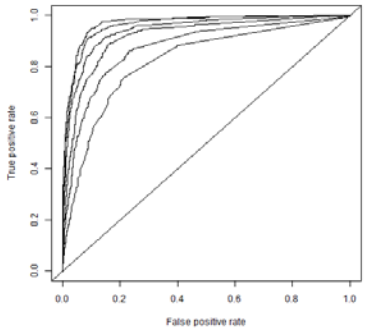
e) kNN on the edX MOOC dataset



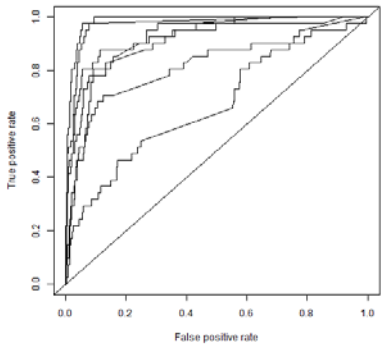
c) LogReg on the FL MOOC dataset



f) LogReg on the edX MOOC dataset



d) XGBoost on the FL MOOC dataset



g) XGBoost on the edX MOOC dataset

Figure 9. ROC values for all weeks prediction models varying the algorithm and the dataset