Early Prediction and Variable Importance of Certificate Accomplishment in a MOOC

José A. Ruipérez-Valiente[a,c], Ruth Cobos[b], Pedro J. Muñoz-Merino[a], Álvaro Andujar[b], Carlos Delgado Kloos[a]

[a] Universidad Carlos III de Madrid. Leganés, Spain
[b] Universidad Autónoma de Madrid. Madrid, Spain
[c] IMDEA Networks Institute. Leganés, Spain

jruipere@it.uc3m.es, ruth.cobos@uam.es, pedmume@it.uc3m.es, alvaro.andujar@estudiante.uam.es, cdk@it.uc3m.es

Abstract. The emergence of MOOCs (Massive Open Online Courses) makes available big amounts of data about students' interaction with online educational platforms. This allows for the possibility of making predictions about future learning outcomes of students based on these interactions. The prediction of certificate accomplishment can enable the early detection of students at risk, in order to perform interventions before it is too late. This study applies different machine learning techniques to predict which students are going to get a certificate during different timeframes. The purpose is to be able to analyze how the quality metrics change when the models have more data available. From the four machine learning techniques applied finally we choose a boosted trees model which provides stability in the prediction over the weeks with good quality metrics. We determine the variables that are most important for the prediction and how they change during the weeks of the course.

Keywords: Educational Data Mining, learning analytics; prediction; machine learning; MOOCs

1 Introduction

MOOCs are courses provided by online platforms that can be accessed by anyone with an Internet connection. These courses might have thousands of students at the same time taking a single MOOC. This massiveness and the fact that each student generates a large amount of events, provides the opportunity to analyze large datasets about the interaction of students with these online educational platforms with the objective of improving the learning process. MOOCs have many positive features and potential to be one of the main possibilities for learning, however many of the problems that were addressed since the beginning have not been resolved yet e.g. the easiness of students to perform academically dishonest behaviors [2].

One of the main identified problems in MOOCs is the high dropout rates. This issue was addressed since edX first MOOC on “Circuits and Electronics” where they reported a completion rate of only 5% with over 155.000 students registered for the course [4]. Most of the studies report very high attrition rates in MOOCs and it is commonly known as one of the main issues. Therefore, the early prediction of success in an educational course can be very important, since if we can predict which students are at risk of not passing the course, then we can take different decisions to try to change it. For example, adaptive systems can be implemented to adapt the contents based on this prediction and we can find in the literature different examples of adaptive systems in education [15].

In this work, we analyze how to predict early on which students are going to earn a certificate and which of them will not, with the purpose of enabling intervention that can alert students that they are in risk of not earning their certificate. This can be especially important for those students that have payed to obtain a verified certificate. In this study, we analyze data from “The Spain of Don Quixote” MOOC offered via edX platform and taught by instructors from Universidad Autónoma de Madrid (Spain, UAM). We use several indicators related to the learning process and different machine learning algorithms to predict which students are going to achieve a certificate. We evaluate and compare the different proposed algorithms and determine which variables are more important regarding the prediction of certification outcome. In addition, we discuss about the best moment for making the prediction and whether we can make accurate prediction with data of only the first weeks.

2 Related Work

One of the main identified problems with MOOCs is the high dropout rates. Recent reviews estimate the average completion rates in MOOCs around the 7% [13]. Regarding the prediction on student absenteeism in MOOC courses, several institutions have developed and designed models that focus on the detection and prediction of student dropout in MOOCs by making use of the indicators obtained from social interactions, the student activity with problems and the navigation within the course. For example, the study by Kloft *et al.* [14] proposes a model based on data from Coursera platform to predict which students are going to dropout the course, and analyzes how the accuracy of the model varies over the weeks; this approach is very similar to ours but our target prediction is certificate accomplishment and some of the variables used are different.

In this direction, The University of George Mason has carried out a task where they bring forward models to predict student performance for a determined assessment activity. It is a real time model that tracks student participation and predicts student performance in the following course evaluation; this model has been tested in Open edX [17]. Technology giant Samsung Electronics has worked towards the objective of being able to predict dropouts in MOOCs by extracting a wide variety of data related to the activity of students and applying different machine learning approaches [19]. The Abdelmalek Essaadi University from Tetouan in Morocco compared the proportion of students who complete the course with those who enrolled and reached the conclusion that some got even a rate lower than 2% [11]. Based on this data, they have created LASyM: A Learning Analytics System for MOOCs, a system based on Phil Hill's behavioral classification [10]. Students were classified in different categories: ghosts, observers, non-completers, passive participants and active participants. The objective was to identify students in risk by using two simple indicators based on interaction and persistence. These indicators are obtained by behavioral analysis and student activity, such as the number of visualized videos, exercises and other course content [20].

One of the most appealing topics in educational research is the prediction of learning outcomes. The study of this topic can help to improve knowledge regarding how students learn and what variables are important, to ultimately be able to improve the learning process of students. The target prediction and learning environment can vary from one study to another. For example we can find traditional settings such as high school education [1], but lately many studies are using data from different types of Virtual Learning Environments (VLEs) such as Intelligent Tutoring Systems (ITSs) [3,12], and more traditional LMS environments [7,16]. Furthermore, these studies target different learning outcomes such as graduation in high school [1], performance in course activities [7], learning gains [18] or end-of-the assessment scores [3].

We can find in the literature different studies that aim at the prediction of learning outcomes of students after interacting with ITS environments [3,12]. MOOC environments are different to ITSs, e.g. the former usually contain more complex and specific exercise players that can provide different features (e.g. about hints) that usually MOOC environments do not have. Additionally, the interaction with videos and the context is different, thus we can expect that the variables that are used to predict learning outcomes in each educational environment differ a bit. We can find also different studies in this direction using MOOC environments, for example towards the prediction of learning gains after the interaction with a Khan Academy instance [18] and also to predict student knowledge status in MOOCs using Open edX [9]. Other works in the literature have approached the prediction of certificate accomplishment using different methods such as LDA [6]. In our work we focus on early prediction of certificate accomplishment, performance of different models and variable importance.

3 Methodology

3.1 Description of the MOOC

UAM offered the first delivery of their MOOCs at edX platform in February 24th 2014 [5] and this dataset belongs to one on these MOOCs, which is entitled “The Spain of Don Quixote” - Quijote501x[1]. A total of 3530 students enrolled in the course; however, only 1718 students were actively involved with any of the course content of which 164 students obtained a grade of over 60% and thus received a cer-

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

tificate. Therefore around 4.65% of the enrolled students earned a certificate, which is a completion rate similar to the ones reported in the literature. It is a 7-week course where every week there are multimedia resources, discussion forums, practical activities without evaluation, and also a final evaluation activity per week. The first and last week (seventh), students were evaluated with a peer review activity. For weeks 2 to 6, they were evaluated with a multiple choice test of 21-23 questions. Each weekly evaluation contributed a 14% to the final grade of the course. For the first three weeks of this course, the evaluation activities deadlines are four weeks after the release date. Then, from the fourth week to the end of the course the evaluation activities deadlines are three weeks after.

3.2 Data collection and selected variables

The main source of data involved the tracking logs of students[2], which contained detailed information regarding all actions performed by the students, a total number of 893.098 events were triggered. We used this file to compute the different variables. The dependent variable is binary and addresses the acquisition of a certificate. We used a total of 11 independent variables to build the model and predict the dependent variable. Two of these variables are related to the progress of students, one regarding the grade achieved in the completed assignments (*problem_progress*) and other regarding the percentage of different video watched from 0 to 100% (*video_progress*). Then, five variables are related to the volume and amount of activity of students, the summation of time invested in problems (*total_problem_time*), the summation of time invested in videos (*total_video_time*), the total time in the course (total_time), the number of sessions (*number_sessions*) and number of events produced by the student (*number_events*). Finally, four variables related to the distribution of the activity with the dispersion of the time invested in each exercise separately (*problem_homogeneity*), dispersion of time in each video separately (*video_homogeneity*), number of days that the student logged in (*number_days*) and dispersion of time invested in each day of the course (*constancy*).

3.3 Method and tools

With the purpose of analyzing early prediction, we divided the data into seven data batches that corresponded with the data available after each week's deadline. We fed the machine learning algorithms with each one of these batches. Therefore, in a practical sense, we implemented a model with all the data just after first week, a model with all the data just after the second week, and so on. This way we were able to analyze the evolution of the performance of each model as the amount of data available increases in the direction of looking for stability and early detection. We used R software and specifically *caret* package and its functions. We implemented four classification models that are random forests (RF) using *randomForest* package, generalized boosted regression modeling (GBM) using *gbm* package, k-nearest neighbours (kNN)

[2] http://edx.readthedocs.io/projects/devdata/en/latest/

using *knn* package, and a logistic regression using *glm* package. The specific steps that we followed for the training and the evaluation of the models are the following:

1. **Training**: We divided the dataset in training and test with a probability of 0.75 and used *train* function of *caret* package to train the four models. As part of the pre-processing we scaled and centered all the numeric variables. We established that the quality metric that we want to maximize is the ROC, and it was estimated through a 10-fold cross validation repeated and averaged three times.
2. **Model selection**: We used each one of the implemented models of the previous step to predict on the data from the test dataset. We evaluated the results over the weeks using F1-score and AUC and selected the best model.
3. **Evaluation**: We evaluated the selected best model in terms of AUC, F1-score, sensitivity, specificity, Cohen's kappa coefficient and accuracy. We analyzed the importance of variables of the selected model as more data becomes available.

4 Results and Discussion

4.1 Training and selection of the model

We partition the data in train (N=1289, 123 certificate earners) and test (N=429, 41 certificate earners) dataset. We train the four models using each one of the seven batches following the method described in Subsection 3.3. Next, we evaluate each one of the models by predicting on the test dataset. Figure 1 represents the quality metrics, i.e. F1-score (on the left) and AUC (on the right), over the weeks for each model.

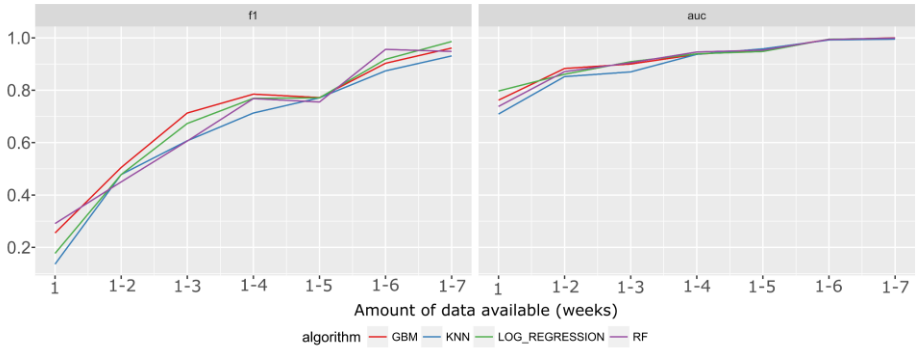

Fig. 1. Evaluation results in the test dataset in terms of F1-score and AUC metrics for the models on each batch of data.

We can see that the performance after the first week in terms of F1-score is a bit higher for the RF model, and afterwards BGM model takes over as the best one. In terms of AUC in the first week log regression is the best, in the second week GBM has the best performance and afterwards the AUC values are very similar. We are looking for a stable model over the weeks, offering always a good performance and specially in the first four weeks, since those are the weeks in which we have chances of sending an early warning to avoid that a student misses the certificate. Considering those premises, in terms of the F1-score and AUC, we consider that the model that provides the best performance for this task is the GBM model, which always performs as the best or second best model over the four first weeks both in terms of F1-score and AUC, offering performance and stability.

4.2 Evaluation results and discussion

In this section we analyze the results when predicting on the test data using the selected GBM model. Figure 2 shows the evolution over the weeks for the GBM model in terms of sensitivity, specificity, F1-score, Cohen's kappa coefficient, AUC and accuracy. Additionally, we have added the baseline accuracy of the predictor that always classifies as non-certificate earners (0.904). We can see how the specificity remains high over the weeks, but the sensitivity is very low at the beginning, we are aware of these results but we think this is the correct approach. We want to minimize false positives, since a false positive implies a student who is not going to receive a certificate and still will not be warned by our system due to the classification created by the prediction model. We are less concerned about students who will get a certificate, but receive a warning regarding they are still in risk of not getting a certificate. F1-score increases in a similar trend than the sensitivity does, since F1-score is the geometric mean of sensitivity and specificity. Additionally kappa coefficient also increases over time, the more that the predictor starts behaving differently than the baseline predictor (being able to detect both true negatives and true positives), the more the kappa coefficient increases.

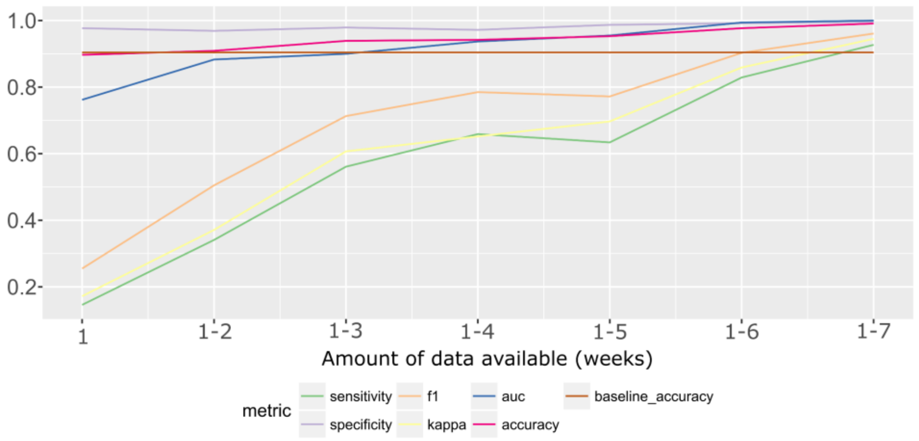

Fig. 2. Evolution of the performance in the test dataset of the selected GBM model over the weeks.

There is a progressive prediction improvement during the first weeks due to the fact that most learners that dropout are likely to do it during the first weeks, therefore activity metrics that show that students are interacting with the platform, are very important during these first weeks. One interesting detail is the effect of the deadlines of week 5, where we can see that in terms of sensitivity, the predictor gets worse and accuracy improves little (this happened also with the rest of the algorithms). Then, after week 6 there is a big improvement in terms of accuracy and sensitivity. These findings suggest that early prediction models should heavily take into account the deadlines of the course, and use them to find the optimal time in which the system should warn students (e.g. just before or after deadlines). These deadlines could even be parametrized into variables and input to the model so that the prediction accuracy improves. By the end of the last week the accuracy is really high (0.991), but the course is finished and the system can no longer send early warnings, that is why we should focus in the first three or four weeks.

We have explored also the influence of the different variables of the GBM model over the weeks. We compute and report the relative variable importance (scaled from 0 to 100) as reported by Friedman [8] of each one of the variables of the model. We plot the variable importance results over the weeks in Figure 3. The results show an interesting trend where there is a lot of difference in the importance of variables during the first three weeks, where it is distributed among many variables, and at the end of the course, where *problem_progress* is with much difference the most important. We can see that after week 3, the most important variable is *problem_progress*, and *total_problem_time* is the second most important one, the rest of the variables have low relative importance. However we can see that during the first weeks the importance is more distributed among the different variables. The most interesting detail is that at the end of the first week, *number_sessions*, *total_problem_time*, *number_events*, *video_homogeneity*, *total_time* and *video_progress* have more importance towards the prediction than *problem_progress*. Additionally, specially at the end of week two but also at the end of week three, some of these variables still have great importance (specially note *number_sessions* at the end of week two). These results seem to indicate that in the first weeks of the course some indicators related to the activity of students with the platform are even more useful to early detect students in risk than the indicator related to the progress in problems, specially at the end of the first week. We believe this is a very interesting finding, since we should mostly focused on warning students during the first weeks, since that is the timeframe where we are still able to help and get them back in track to earn their certificate. We were surprised to see that *number_sessions*, being such a straightforward variable, achieved a high importance during the first weeks. By the end of week three, we can see the variable importance of most of the variables have dropped below the 30% of influence and most of the prediction power comes *problem_progress*. Taking into account that the course analyzed is a xMOOC, this extracted result is coherent. These findings suggest that warning models should probably tune the weight of the variables during the different weeks of the course.

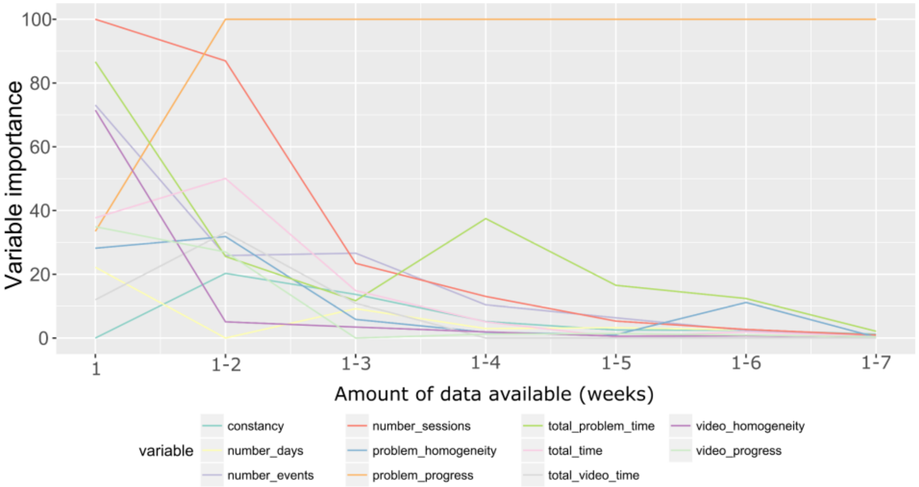

Fig. 3. Evolution of variable importance of the GBM model over the weeks.

5 Conclusions and Future Work

In this study we have approached the implementation of machine learning models with the objective of early predicting which students will fail in the task of accomplishing enough grade to get a certificate, so that interventions (either automatic or by instructors) can be performed before it is too late. Our results suggest that GBM model was the best one in terms of both performance and stability over the first weeks, although the RF and logistic regression models offered also very good performance. Since the performance of the different algorithms might change as more data becomes available, this might raise a question related to whether some algorithms should be used for very early predictions and change to others when the amount of data increases. Following this idea, we found that some of the activity variables had a high importance towards the prediction during the first weeks, whereas after week three the most important indicator was *problem_progress*; this can be of interest implementing an early warning systems for students in risk. We are making a hard class prediction in this work (currently, we select the class with the probability above 0.5), but it would also be possible to follow a soft prediction approach where we get a probability for each class, and depending on this probability we can adapt the type of warning that we are going to send, e.g. a very high probability of not getting a certificate might receive a strong warning whereas a threshold probability can receive a more moderate warning.

The main limitation of the study is that both the training and the test data sets rely on data from a course that is already finished, and we have not been able to compare our results on data from a different course. We believe that the most immediate next step would be to test the model on a second re-run from the same course (i.e. a different delivery of the course) to determine whether these kind of models could at least extrapolate to courses with many similar contents and course structure, e.g. to try to predict using our selected model from this research on the second run of Don Quixote MOOC. A more complex question to ask would be if it is possible to implement a model that can effectively predict on different MOOCs taking into account the different topics, educational resource types and contents.

In the future, more granular analysis would be required to be able to answer such questions, for example, to be able to map the impact of the different learning resources and the structure of the course to the performance of the predictors. This would allow the design of a more general approach that could apply better to different courses. We reported on results after analyzing data after each week of the course, but a more granular data analysis, e.g. each day, could also be tested. Due to the heterogeneity in MOOC learners we would like to add additional variables such as motivation of students or prior education. The final stage of this research could be to perform an A/B experiment in which the treatment group would receive warnings from the system based on a model like this one, and the control group would not; then we would be able to compare if there was improvement in terms of certificate accomplishment rates between the two groups.

6 Acknowledgments

Work partially funded by the Madrid Regional Government with grant No. S2013/ICE-2715, the Spanish Ministry of Economy and Competitiveness projects RESET (TIN2014-53199-C3-1-R) and Flexor (TIN2014-52129-R) and the European Erasmus+ projects MOOC Maker (561533-EPP-1-2015-1-ES-EPPKA2-CBHE-JP) and SHEILA (562080-EPP-1-2015-BE-EPPKA3-PI-FORWARD). This research work was made possible thanks to Universidad Autónoma de Madrid, which provided us with the dataset, and to Prof. Pedro García, who was the instructor of the selected MOOC.

References

1. Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., Addison, K.L.: Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. pp. 93-102. ACM (2015)
2. Alexandron, G., Ruipérez-Valiente, J.A., Chen, Z., Muñoz-Merino, P.J., Pritchard, D.E.: Copying@Scale: Using harvesting accounts for collecting correct answers in a MOOC. Computers & Education 108, 96-114 (2017)
3. Anozie, N., Junker, B.W.: Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press (2006)
4. Breslow, L., Pritchard, D.E., DeBoer, J., Stump, G.S., Ho, A.D., Seaton, D.T.: Studying learning in the worldwide classroom: Research into edX's First MOOC. Research & Practice in Assessment 8 (2013)

5. Claros, I., Cobos, R., Sandoval, G., Villanueva, M.: Creating MOOCs by UAMx: experiences and expectations. The Third European MOOCs Stakeholders Summit (eMOOC 2015) pp. 61-64 (2015)
6. Coleman, C.A., Seaton, D.T., Chuang, I.: Probabilistic use cases: Discovering behavioral patterns for predicting certification. In: Proceedings of the Second (2015) ACM Conference on Learning@Scale. pp. 141-148. ACM (2015)
7. Elbadrawy, A., Studham, R.S., Karypis, G.: Collaborative multi-regression models for predicting students' performance in course activities. In: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge. pp. 103-107. ACM (2015)
8. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189-1232 (2001)
9. Guo, S., Wu, W.: Modeling student learning outcomes in MOOCs
10. Hill, P.: Emerging student patterns in MOOCs: A (revised) graphical view (2013)
11. Jordan, K.: MOOC completion rates: The data. Available at: http://www.katyjordan.com/MOOCproject. [Accessed: 27/08/2014] (2013)
12. Kelly, K., Arroyo, I., Heffernan, N.: Using ITS generated data to predict standardized test scores. In: Educational Data Mining 2013 (2013)
13. Khalil, H., Ebner, M.: MOOCs completion rates and possible methods to improve retention-a literature review. In: World Conference on Educational Multimedia, Hypermedia and Telecommunications. pp. 1305-1313. No. 1 (2014)
14. Kloft, M., Stiehler, F., Zheng, Z., Pinkwart, N.: 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. pp. 60-65 (2014)
15. Muñoz-Merino, P.J., Molina, M.F., Muñoz-Organero, M., Kloos, C.D.: An adaptive and innovative question-driven competition-based intelligent tutoring system for learning. Expert Systems with Applications 39(8), 6932-6948 (2012)
16. Pardo, A., Mirriahi, N., Martinez-Maldonado, R., Jovanovic, J., Dawson, S., Gašević, D.: Generating actionable predictive models of academic performance. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. pp. 474-478. ACM (2016)
17. Ren, Z., Rangwala, H., Johri, A.: Predicting performance on MOOC assessments using multi-regression models. arXiv preprint arXiv:1605.02269 (2016)
18. Ruipérez-Valiente, J.A., Muñoz-Merino, P.J., Kloos, C.D.: A predictive model of learning gains for a video and exercise intensive learning environment. In: International Conference on Artificial Intelligence in Education. pp. 760-763. Springer (2015)
19. Sinha, T., Jermann, P., Li, N., Dillenbourg, P.: Your click decides your fate: Inferring information processing and attrition behavior from MOOC video clickstream interactions. arXiv preprint arXiv:1407.7131 (2014)
20. Tabaa, Y., Medouri, A.: LASyM: A learning analytics system for MOOCs. International Journal of Advanced Computer Science and Applications (IJACSA) 4(5) (2013)