Using a Predictive Analytics Process to identify droppers in MOOCS

Alya Itani[1] Laurent Brisson[1] Issam Rebai[1] Serge Garlatti[1]

[1] IMT Atlantique, Lab-STICC,UBL
F-29238 Brest, France

{alya.itani,laurent.brisson,issam.rebai,Serge.Garlatti}@imt-atlantique.fr

Résumé

Cette communication traite du problème de l'abandon des apprenants dans les MOOCs d'une part pour l'anticiper et d'autre part pour en trouver les causes. Les causes étudiées sont ici liées à la conception du cours et aux comportements des apprenants (demande d'OpenClassrooms, OC). Deux besoins opérationnels critiques ont été identifiés : (1) la détection fine des apprenants décrocheurs, notamment pour leur envoyer des messages de motivation automatisés ; (2) l'étude des causes possibles des abandons, pour intervenir humainement de façon personnalisée. Pour répondre à ces besoins, nous utilisons des classificateurs (apprentissage automatique) de types prédictif et explicatifs. Cet article présente le processus de prédiction que nous avons réalisé.

Mots Clef

Prédiction, learning analytics, MOOCs, Drop-out.

Abstract

This paper focuses on anticipating the drop-out among MOOC learners and finding the reasons. The main excavated reasons are those related to course design and learners behaviors - OpenClassrooms (OC) requirements. Two critical business needs are identified : (1) the accurate detection of at-risk droppers to send automated motivational feedback ; (2) the investigation of possible drop-out reasons to personalize interventions. To meet these needs, we deploy both predictive and explicative types of machine learning classifiers. This article presents the achieved prediction process.

Keywords

Learning Analytics, Supervised ML, MOOCs, Drop-out.

1 Introduction

MOOCs offer an alternative education method that changed the standards of teaching and learning forever. Education has reformed to become attainable to the whole public at any age, price, country, time, and mean [1]. This elevated ease and unrestricted access to material led to massiveness not only in the scale of participation but also in that of incompletion, commonly known as drop-out [2]. Consequently, a wide investigation on MOOC drop-outs rates was provoked. In that context, Khalil and Ebner [3], Colman [4], and Onah et al. [5] all investigated the reasons of this marked drop-out. The most addressed reasons can be summed up as follows :(1) Lack of intention to complete (2) Personal circumstances (3) Bad MOOC design (4) Deficiency in digital skills. (5) Inaccurate expectations. (6) Bad prior experience. In our case-study, we are interested in examining the reasons of drop-out related to course design and learners behaviors as a preference of the concerned MOOC provider OC.

2 The Drop-out Prediction Process

The investigation of MOOC drop-out and its reasons opened the horizon towards using machine learning techniques to predict drop-out ahead of time and try to prevent it. Initially, studies attempted drop-out predictions considering mono-type contextual features, like forum interactions or video restricted events [6, 7].

The principle objective of our predictive process is to classify learners of a MOOC into at-risk droppers and completers at a certain instant throughout the course. Hence, the prediction target in this problem is of two categorical classes (dropper, completer), and the dataset at hand is a labeled dataset. Therefore, we propose a process based on a supervised machine learning solution. Mainly, the process helps in attaining two goals : (1) Offering highly accurate predictions for the purpose of automated interventions (2) offering readable and explainable predictions for the purpose of personalized interventions. Figure 1 illustrates the prediction process detailed in sections 2.1, and 2.2.

2.1 Data Understanding and Preparation

The OpenClassrooms dataset includes activity traces of 190,000 learners within 10 courses in various domains. Each course is divided into chapters and each chapter into parts. At the end of each chapter, there are multiple choice or peer assessed exercises. The available dataset includes : **I) User demographics** with anonymous ids ; **II) Subscription actions** premium subscription, events of following and un-following course ; **III) Course related ac-**

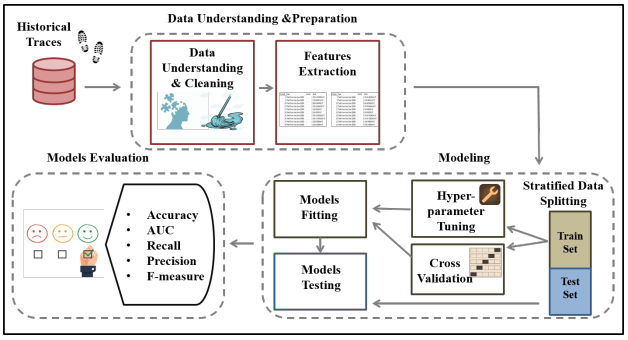

FIGURE 1 – Phases of the Drop-out Prediction Process

tions course visualization, course-parts completion, users course grades ; **IV) Exercise related actions** information on exercises and user exercise sessions.
The structure of the data is verified and the necessary cleaning is performed. After that, suitable indicators and features for analysis are computed. Two types of indicators are introduced for features selection : regular statistical indicators and the dynamic behavioral indicators. Regular indicators are simply computed statistics on learner-course interactions. Whereas, behavioral indicators demonstrate learners trajectories of engagement with the course versus the recommended trajectory of the course.

2.2 Predictive Modeling

For the modeling phase, two types of machine learning classifiers are used : (1) highly accurate predictive classifiers, whose structure analysis does not provide any insight on how the model works. We use of them Random Forest (RF) and Gradient Boosting (GB). (2) explicative classifiers whose structural analysis provides an understanding of how the model works. We use of them Decision Trees (DT), Logistic Regression (LR), and K Nearest Neighbors (KNN). The steps of the prediction process are :
Stratified Data Splitting (*S*) is the first step (figure 1), which is partitioning the final features and target dataset into a training/validation set (60%) and a testing set (40%) while preserving the balance of droppers and completers within the initial dataset (stratification).
Hyperparameter Tuning (*T*) is choosing the right parameters enhances the performance of models. Thus, we tune the parameters of all five models with grid searching. Grid search applies an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm [8]. Moreover, we include a step of scaling and normalization of the features matrix before tuning, fitting, and testing the models to avoid such flaw.
Cross Validation (*V*) to avoid overfitting the models [8], we apply k-fold cross-validation that randomly partitions the data into K subsamples (stratification included to the sampling). One of the K subsamples is retained for testing and the rest of the subsamples are used for training. This action is then repeated K times, known as folds. The performance is measured on each fold and then averaged at the end.

2.3 Experimentation & Evaluation

The overall metrics used for evaluation are : Accuracy, Precision, Recall, F-measure, ROC curve, and AUC. AUC is used for hyper-parameter tuning, since it suits best the available balance of data. Whereas, the classifiers evaluation is based on both AUC and F-measure, since F-measure represents a critical balance between precision and recall.
Our experimentations allow us to verify that applying S-T-V process has a positive effect on the performance of all classifiers in general and on explicative classifiers (DT, KNN, LR) in particular and that including behavioral indicators slightly increased the predictive power of all classifiers mainly KNN. Upon testing, all STV fitted classifiers, with RF dominating, succeeded in delivering accurate predictions of at risk learners at the end of the second chapter of the course. Thus, satisfying the need of the MOOC providers in the prospect of sending automated motivational feedback to at risk spotted learners. Moreover, we were able to attain the awaited readability on predictions using the DT classifier.

Références

[1] E. J. Emanuel, "Online education : Moocs taken by educated few," *Nature*, vol. 503, no. 7476, pp. 342–342, 2013.

[2] R. Rivard, "Measuring the mooc dropout rate," *Inside Higher Ed*, vol. 8, p. 2013, 2013.

[3] H. Khalil and M. Ebner, "Moocs completion rates and possible methods to improve retention-a literature review," in *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, no. 1, 2014, pp. 1305–1313.

[4] D. Colman, "Mooc interrupted : Top 10 reasons our readers didnâĂŹt finish a massive open online course," *Open Culture*, 2013.

[5] D. F. Onah, J. Sinclair, and R. Boyatt, "Dropout rates of massive open online courses : behavioural patterns," *EDULEARN14 Proceedings*, pp. 5825–5834, 2014.

[6] M. Wen, D. Yang, and C. P. Rosé, "Linguistic reflections of student engagement in massive open online courses." in *ICWSM*, 2014.

[7] A. Ramesh, D. Goldwasser, B. Huang, H. Daumé III, and L. Getoor, "Learning latent engagement patterns of students in online courses," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, 2014, pp. 1272–1278.

[8] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.