MOOC Dropout Prediction: How to Measure Accuracy?

Jacob Whitehill
Worcester Polytechnic Institute
Worcester, MA, USA
jrwhitehill@wpi.edu

Kiran Mohan
Worcester Polytechnic Institute
Worcester, MA, USA
kmohan@wpi.edu

Daniel Seaton
Harvard University
Cambridge, MA, USA
daniel_seaton@harvard.edu

Yigal Rosen
Harvard University
yigal_rosen@harvard.edu

Dustin Tingley
Harvard University
dtingley@gov.harvard.edu

ABSTRACT

In order to obtain reliable accuracy estimates for automatic MOOC dropout predictors, it is important to train and test them in a manner consistent with how they will be used in practice. Yet most prior research on MOOC dropout prediction has measured test accuracy on the same course used for training, which can lead to overly optimistic accuracy estimates. In order to understand better how accuracy is affected by the training+testing regime, we compared the accuracy of a standard dropout prediction architecture (clickstream features + logistic regression) across 4 different training paradigms. Results suggest that (1) training and testing on the same course ("post-hoc") can significantly overestimate accuracy. Moreover, (2) training dropout classifiers using proxy labels based on students' *persistence* – which are available *before* a MOOC finishes – is surprisingly competitive with post-hoc training (87.33% v. 90.20% AUC averaged over 8 weeks of 40 HarvardX MOOCs) and can support real-time MOOC interventions.

INTRODUCTION

Within the fields of learning analytics and educational data mining, the possibility of creating automatic MOOC "dropout detectors" has generated considerable interest within the past few years. Such detectors could facilitate automated interventions designed to improve the persistence and performance of those MOOC learners who are at-risk of dropping out. Existing research on dropout prediction (see Table 1) has varied across several dimensions including the **features** (clickstream, social network measures, etc.) used for prediction as well as the machine learning **classifier** (logistic regression, survival analysis models, neural networks, etc.) used for training and testing.

More subtle, but important for real-time deployment in an actual course, is the **training paradigm** that describes how the *source* of the training data – e.g., the same course, a prior instance of the same course, or a different course altogether – relates to the *target use* of the classifier once it has been trained. To-date, most research on MOOC dropout prediction has focused on training and testing on data sampled from the same MOOC, likely because it is simplest to implement in simulation. Live interventions, on the other hand, require dropout predictors that are operational near the *start* of a MOOC when students can still benefit from receiving an intervention. But producing such a predictor can be difficult because the target values which indicate whether each student dropped out or completed the MOOC, and which are usually required by supervised learning algorithms, become available only at the *end* of a MOOC – at which point any intervention is moot.

In this paper, we seek to fill a methodological gap in MOOC dropout prediction research and to investigate how the training paradigm affects the accuracy of the trained predictors. In particular, we conduct machine learning experiments on 40 HarvardX MOOCs using a standard architecture – clickstream features classified by L_2-regularized logistic regression – and compare performance across 4 different training paradigms as well as 2 simple baseline prediction approaches. Our greater goal is to increase awareness of the trade-offs across different training paradigms in terms of accuracy, ease of implementation, and applicability to new courses, so that researchers and practitioners can make better decisions about how to implement real-time MOOC interventions.

DATASET

The analyses in this paper are based on data from 40 MOOCs from HarvardX; due to space constraints, these courses are summarized in Table 2 of [16]. The binary **target labels** used for training and testing were whether (1) or not (0) each student accrued enough points during the MOOC to earn a certificate. The grade threshold for certification differed across the MOOCs but is typically around 70%. Note that, starting in late 2015, some HarvardX MOOCs required students to pay a fee (to have

L@S 2017, April 20 - 21, 2017, Cambridge, MA, USA

ISBN 978-1-4503-4450-0/17/04...$15.00
DOI: http://dx.doi.org/10.1145/3051457.3053974

Survey of Prior Research on MOOC Dropout Prediction

Study	#MOOCs	Features	Architecture	Training paradigm
Balakrishnan & Coetzee [1]	1	Clickstream	HMM + SVM	Same course
Boyer & Veeramachaneni [2]	3	Clickstream	TL+LR	Different offering In-situ
Coleman et al. [3]	1	Clickstream	LDA+LR	Same course
Crossley et al. [4]	1	Clickstream; NLP	DFA	Same course
Fei & Yeung [5]	2	Clickstream	RNN	Same course
He et al. [7]	2	Clickstream	Smoothed LR	Different offering
Jiang et al. [8]	1	Social network; grades	LR	Same course
Kizilcec et al. [9, 6]	20	Clickstream	LR	Different course Same course
Kloft et al. [10]	1	Clickstream	SVM	Same course
Koedinger et al. [11]	1	Clickstream; grades	LR	Same course
Robinson et al. [12]	1	Survey; NLP	LR	Same course
Rose et al. [19, 13]	1	Forum; social network	SA	Same course
Stein & Allione [14]	1	Clickstream; survey	SA	Same course
Taylor et al. [15]	1	Clickstream	LR	Same course
Whitehill et al. [17]	10	Clickstream	LR	Different course
Xing et al. [18]	1	Clickstream; social network	PCA+{BN,DT}	Same course
Ye & Biswas [20]	1	Clickstream	LR	Same course
Our work	40	Clickstream	{LR, DNN}	Same course In-situ Different course

Table 1. Survey of prior literature on MOOC dropout prediction. For the architecture, we use the following abbreviations: Bayesian network (BN), decision tree (DT), deep neural network (DNN), discriminant function analysis (DFA), hidden Markov model (HMM), latent Dirichlet allocation (LDA), logistic regression (LR), principal component analysis (PCA), recurrent neural network (RNN), support vector machine (SVM), survival analysis (SA), and transfer learning (TL). Architecture $a+b$ means methods a and b were used in conjunction; $\{a,b\}$ means that a or b were used as alternatives. Note that our full paper [16] describes our experiments using DNN.

their identity verified) in order to receive a certificate. For these courses, we still considered the target label for a student to be 1 as long as her/his point total exceeded the verification threshold – in other words, we ignored whether or not the student paid the fee.

EXPERIMENTS

We conducted experiments to compare the accuracies of different MOOC dropout predictors. All the predictors that we trained were based on clickstream **features**; these are computed from the clickstream log describing interaction events between the student and the MOOC courseware, e.g., answers to quiz questions, play/pause/rewind events on lecture videos, etc. When predicting at week w whether a learner will drop out, clickstream features only up until w are extracted. For the **classifier**, we used L_2-regularized logistic regression and optimized the regularization strength. See [16] for more details on the feature representation and classifier design.

The key independent variable that we manipulated in our experiments was the **training paradigm**. Specifically, for each week w of each of the 40 MOOCs, we trained one classifier for each of the following paradigms:

1. *Train on same course (post-hoc)*: When predicting which students from course c will drop out, train using features and target labels from the exact same course c. **Assumptions**: since target labels for c become available only *after* c has ended, this approach essentially would require either that (a) the practitioner go "back in time" to when the MOOC first started, or (b) that a new MOOC with the exact same distribution of students (demographics, prior knowledge, etc.) and with the exact same content and structure is offered in the future, and that no exogenous factors cause students to behave differently during the later incarnation of the course.

2. *Train on other course from same field*: When predicting which students from course c will drop out, train using features and target labels from a different course c' that has already completed, and for which the target labels are thus already available. Although it is difficult to know *which* prior course should be used for training, a reasonable choice is a different course from within the same discipline (social sciences, humanities, etc.). We chose to use the *largest* such course in order to maximize the number of data available for training.

3. *Train on many other courses*: When predicting which students will drop out, train using features and target labels from *many* different courses (not necessarily within the same discipline). Specifically, for each course c, we trained a dropout classifier from each of the 39 other courses (recall that our dataset includes 40 MOOCs) and then averaged the classifiers' hyperplanes together.

4. *Train using proxy labels (in-situ)*: When predicting at week w which students from course c will drop out, train using *proxy labels* corresponding to whether each student *persisted* – i.e., interacted with the MOOC courseware at least once – within the previous week $w-1$ (see Figure 1). This approach (similar to [2]) can be implemented for any MOOC. Because it does not require "seeing into the

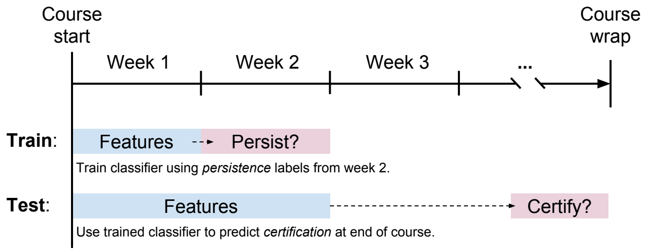
Course start
Week 1
Week 2
Week 3
...
Course wrap
Train:
Features
Persist?
Train classifier using persistence labels from week 2.
Test:
Features
Certify?
Use trained classifier to predict certification at end of course.

Figure 1. Schematic representing how to train using proxy labels (in-situ): At each week w, proxy labels of whether the student *persisted* during week $w-1$ are used to predict whether or not the student will certify or drop out by the end of the course.

future" to obtain target labels, it can deliver a dropout predictor that can be deployed during a *live* MOOC, not just after it has finished. There is, however, a mismatch between the *training* target labels (based on persistence) and the *testing* target labels (based on certification), and this mismatch may degrade performance.

Baseline Approaches

In order to gauge how much "added value" is brought by machine learning approaches to MOOC dropout prediction that utilize detailed clickstream information, we also assessed the accuracy of two simple baseline heuristics. *Baseline 1* uses only demographic information – consisting of self-reported year of birth, continent of origin (Africa, North America, etc.), level of education (primary/elementary school, high/secondary school, college, etc.), and gender – to make predictions. This information is available for each student as soon as she/he registers for the course. We also compared against an even simpler *Baseline 2* which requires no machine learning (and hence no training data) at all; rather, the predictor makes predictions based on the number of days since the student last interacted with the courseware. This variable alone has previously been shown to be highly predictive of dropout [17, 9].

When to Measure Accuracy

When using MOOC dropout detectors to facilitate real-time interventions, it is more useful to be able to predict early in the course, rather than later, which students will eventually drop out. Moreover, near the end of the MOOC, some students may have already accrued enough points to earn a certificate. A dropout detector that predicts that such students will not drop out is not so much "predicting" these students' future performance as it is reporting what they have already achieved; hence, accuracy statistics computed over such students may overestimate the performance of the predictor. For both these reasons, we decided to compute accuracy over all weeks of each MOOC between the course start date (which we call $T_{0\%}$, when instruction begins) and the earliest date by which students could possibly have earned enough points to earn a certificate ($T_{100\%}$).

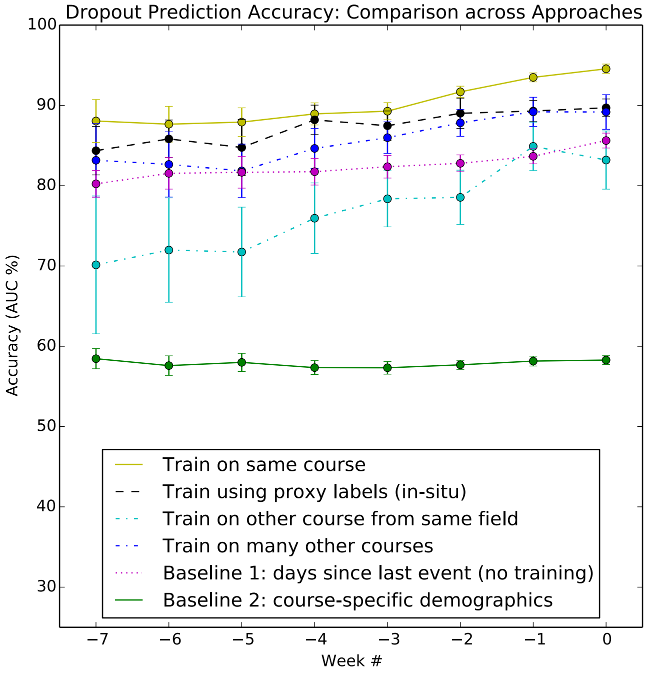

Figure 2. Results comparing different approaches.

RESULTS

Dropout prediction accuracy (AUC) for all training approaches are shown in Figure 2. The horizontal axis indexes the week in the MOOC, where week 0 is defined for each course to be $T_{100\%}$. (Week -3 corresponds to 3 weeks prior to $T_{100\%}$, etc.) Error bars show standard errors. Since the lengths of the courses varied, the accuracy statistics for each week w were computed using only those MOOCs for which data were available at week w.

The most accurate training paradigm was *Train on same course (post-hoc)*, which is the predominant training paradigm used in the dropout prediction literature. It achieved an accuracy (averaged over all 8 weeks, and all MOOCs within each week) of 90.20%. As explained above, the assumptions made by this training paradigm are unrealistic for most intervention scenarios.

Perhaps more surprising is that the second most accurate approach was *Train using proxy labels (in-situ)*. This approach does not require any MOOC – similar or dissimilar – to have been offered previously. Despite the inherent mismatch between the labels of persistence (did the student participate during the previous week?) and labels of certification (will the student earn enough points to earn a certificate?), this approach attained an accuracy (averaged over 40 MOOCs and 8 weeks) of 87.33%.

The third most accurate approach was *Train on many other courses*, with an accuracy of 85.56%. This training paradigm attained a higher accuracy than did *Train on other course from same field* (76.85%), suggesting that, if it is not possible to exploit course-specific structure via either *Train on same course (post-hoc)* or *Train using proxy labels (in-situ)*, then it is better to harness prior data from a large variety of courses than from just a single course (even from within the same discipline).

Comparison to Baseline Approaches: *Baseline 1* achieved an average prediction accuracy of 58.85%, suggesting that only a small amount of information about dropout is contained in the demographics. *Baseline 2* performed remarkably well: It attained an average dropout prediction accuracy of 82.45%, which corroborates previous findings [17, 9] that this variable is highly salient for prediction. Nonetheless, *Baseline 2* was still substantially less accurate than either *Train on same course (post-hoc)* or *Train using proxy labels (in-situ)*, suggesting that harnessing more detailed clickstream features does bring a substantial accuracy boost.

SUMMARY

We explored, on data from 40 HarvardX MOOCs, how MOOC dropout prediction accuracy varies as a function of the **training paradigm**. Results suggest that accuracy estimates obtained by training on the same course (post-hoc) as the target course for deployment – which is generally not possible in real-world intervention scenarios – can be significantly overly optimistic. In addition, we explored a new training paradigm, similar to *in-situ* training [2], based on the idea of *proxy labels* – labels that approximate the quantity of interest (dropout versus certification) but that can be collected *before* a course has completed. Surprisingly, the accuracy of this approach – which is suitable for live interventions in a large variety of MOOCs – is very similar to when classifiers are trained *post-hoc* on courses that have already finished.

REFERENCES

1. G. Balakrishnan and D. Coetzee. Predicting student retention in massive open online courses using hidden markov models. Technical report, UC Berkeley, 2013.
2. S. Boyer and K. Veeramachaneni. Transfer learning for predictive models in massive open online courses. In *International Conference on Artificial Intelligence in Education*, 2015.
3. C. Coleman, D. Seaton, and I. Chuang. Probabilistic use cases: Discovering behavioral patterns for predicting certification. In *Learning at Scale*, 2015.
4. S. Crossley, L. Paquette, M. Dascalu, D. S. McNamara, and R. S. Baker. Combining click-stream data with nlp tools to better understand MOOC completion. In *Learning Analytics & Knowledge*, pages 6–14. ACM, 2016.
5. M. Fei and D.-Y. Yeung. Temporal models for predicting student dropout in massive open online courses. In *International Conference on Data Mining Workshop (ICDMW)*, 2015.
6. S. Halawa, D. Greene, and J. Mitchell. Dropout prediction in MOOCs using learner activity features. In *European MOOC Summit*, 2014.
7. J. He, J. Bailey, Benjamin, I. Rubinstein, and R. Zhang. Identifying at-risk students in massive open online courses. In *AAAI*, 2015.
8. S. Jiang, A. Williams, K. Schenke, M. Warschauer, and D. O'Dowd. Predicting MOOC performance with week 1 behavior. In *Educational Data Mining*, 2014.
9. R. Kizilcec and S. Halawa. Attrition and achievement gaps in online learning. In *Learning at Scale*, 2015.
10. M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 60–65, 2014.
11. K. R. Koedinger, J. Kim, J. Z. Jia, E. A. McLaughlin, and N. L. Bier. Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Learning at Scale*, 2015.
12. C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach. Forecasting student achievement in MOOCs with natural language processing. In *Learning Analytics & Knowledge*, 2016.
13. C. P. Rosé, R. Carlson, D. Yang, M. Wen, L. Resnick, P. Goldman, and J. Sherer. Social factors that contribute to attrition in MOOCs. In *Learning at Scale*, pages 197–198. ACM, 2014.
14. R. Stein and G. Allione. Mass attrition: An analysis of drop out from a principles of microeconomics MOOC. *PIER Working Paper*, 14(031), 2014.
15. C. Taylor, K. Veeramachaneni, and U.-M. O'Reilly. Likely to stop? Predicting stopout in massive open online courses. *arXiv*, 2014. `http://arxiv.org/abs/1408.3382`.
16. J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley. Delving deeper into MOOC student dropout prediction. *arXiv*, 2017. `http://arxiv.org/abs/1702.06404`.
17. J. Whitehill, J. Williams, G. Lopez, C. Coleman, and J. Reich. Beyond prediction: Toward automatic intervention to reduce MOOC student stopout. In *Educational Data Mining*, 2015.
18. W. Xing, X. Chen, J. Stein, and M. Marcinkowski. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Computers in Human Behavior*, 58:119–129, 2016.
19. D. Yang, T. Sinha, D. Adamson, and C. P. Rose. "Turn on, tune in, drop out": Anticipating student dropouts in massive open online courses. In *NIPS Workshop on Data-Driven Education*, 2014.
20. C. Ye and G. Biswas. Early prediction of student dropout and performance in MOOCs using higher granularity temporal information. *Journal of Learning Analytics*, 1(3), 2014.