Dropout Prediction in MOOCs using Learners' Study Habits Features

Han Wan Jun Ding Xiaopeng Gao
School of Computer Science and Engineering
Beihang University
Beijing, China
+86-10-82338059
{wanhan, dingjun, gxp}@buaa.edu.cn

David Pritchard
Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA, 02139
617-253-6812
dpritch@mit.edu

ABSTRACT
Many educators have been alarmed by the high dropout rates in MOOC. There are various factors, such as lack of satisfaction or attribution, may lead learners to drop out. Educational interventions targeting such risk may help reduce dropout rates. The primary task of intervention design requires the ability to predict dropouts accurately and early enough to deliver timely intervention. In this paper, we present a dropout predictor that uses student activity features and then we add learners' study habits features to improve the accuracy. Our models achieved an average AUC (receiver operating characteristic area-under-the-curve) as high as 0.838 (if lacking study habits is 0.795) when predicting one week in advance. The model with learners' study habits features attained average increase in AUC of 0.03, 0.06, 0.08 and 0.05 in different cohorts (passive collaborator, wiki contributor, forum contributor, and fully collaborative).

Keywords
MOOC, dropout prediction, study habits

1. INTRODUCTION
One way to solve the high dropout rates in MOOC is to deliver timely intervention by predicting the dropout probability. Some researchers focused on extracting features of learners' study activities (such as resource accessing) from MOOCs' log, and then building machine learning models. Balakrishnan [1] used the discrete single stream HMMs model to predict whether a student would dropout or not. [2] tried to establish an extensible real-time predicting model, which is fit for any different courses. Loya [3] demonstrated that who executed their learning process on schedule has greater probability to finish the course in MOOCs. Liang J [4] predicted a student's dropout state 10 days later with 3 months' data into four typical machine learning models (LR/SVM/GBDT/RF).

Taylor C. [5] used the dataset of 6.002x: Circuits and Electronics taught in Fall of 2012 on edX, includes course information and students' activity data. In addition to the common simple features, they produced some complex, multi-layered interpretive features, and then used them as the input of predicting models. They divided the students into four groups according to their participation: *passive collaborator* are those learners never actively participated in either the forum or the Wiki, they just view the resources, but did not have contributions; *wiki contributor* are those learners generated Wiki content, but never posted in the forum; *forum contributor* are those learners posted in the forum, but never actively participated in the Wiki; *fully collaborative* are those learners actively participated by generating Wiki content and posting in the forum. Their results shown that if the sample size of the students group is small (especial for wiki contributor, forum contributor and fully collaborative), the predicting accuracy is relative low.

In our work, we focus on extracting more important features of learners' study habits features to improve the accuracy of predicting models, particularly for the small sample size group.

2. PREDICTION PROBLEM DEFINITION
Our data obtained from the 2014 instance of the introductory physics MOOC 8.MReV through the edX platform. We considered defining the dropout point as the time slice (week) a learner fails to submit any further assignments or problems / exam.

The instructor could use the data from week 1 to the current week i to make predictions. The model will predict existing learner dropout during week (i + 1) to week 16. For example, current week is week 7, and we use the logging data from week 1 to week 7 to predict the learners' performance at week 12 with *lead* equals to 4 and *lag* equals to 7.

3. FEATURES ENGINEERIN
Table 1. Self-proposed covariates

	NAME	Definition
x1	stopout	Whether the student continue submit problem
x2	total_duration	Total time spent on all resources
x3	number_forum_posts	Number of forum posts
x4	number_wiki_posts	Number of wiki posts
x5	average_length_forum_post	Average length of forum posts
x6	number_distinct_problems_submitted	Number of distinct problems attempted
x7	number_submissions	Number of submissions
x8	number_distinct_problems_submitted_correct	Number of distinct correct problems
x9	average_number_submissions	Average number of submissions per problem (x7 / x6)
x10	observed_event_duration_per_correct_problem	Total time spent /number of distinct correct problems (x2 / x8)
x11	submissions_per_correct_problem	Number of problems attempted / number of correct problems (x6 / x8)
x12	average_time_to_solve_problem	Average time between first and last problem submissions for each problem (average(max(submission.timestamp) -min(submission.timestamp) for each problem in a week))
x13	observed_event_variance	Variance of a student's observed event timestamps
x14	number_collaborations	Total number of collaborations (x3 + x4)
x15	max_observed_event_duration	Duration of longest observed event
x16	total_lecture_duration	Total time spent on lecture resources
x17	total_book_duration	Total time spent on book resources
x18	total_wiki_duration	Total time spent on wiki resources

We extracted 18 self-proposed features, 7 crowd-proposed features (according to Taylor's work [5]) and 6 study habits related behavioral features on a per-learner basis, these features are list in table 1, table 2 and table 3. And then these features are

assembled from different weeks as separate variables to build predictive models.

Table 2. Crow-proposed covariates

	NAME	Definition
x201	number_forum_responses	Number of forum responses
x202	average_number_of_submissions_percentile	A student's average number of submissions / the average of all the students' submission
x203	average_number_of_submissions_percent	A student's average number of submissions / maximum average number of submissions
x204	pset_grade	Number of the week's homework problems answered correctly / number of that week's homework problems
x205	pset_grade_overtime	Difference in grade between current pset grade and average of student's past pset grade
x206	correct_submissions_percent	Percentage of the total submissions that were correct (x 8 / x 7)
x207	average_predeadline_submission_time	Average time between a problem submission and problem due date over each submission

Table 3. Study habits related behavioral features

	NAME	Definition
x301	problem_finish_percent _pre_start24h	The number of problem learner finished correctly in the first 24h after the problem issued
x302	problem_finish_percent_pre_deadline24h	The number of problem learner finished correctly in the last 24h before the problem due
x303	time_first_visit	Min(time_first_problem_get, time_first_html etext_access) - project_issue_time
x304	time_till_first_check	Average of all problem the time between problem_first_check and problem_first_get
x305	study_before_submit	Total book duration before problem submit + total video duration before problem submit
x306	discussion_duration_after_incorrect_submit	Total discussion duration after incorrect submitssion

4. RESULTS

As shown in figure 1, for all learners, our models achieved an average AUC as high as 0.838 (and lacking study habits features is 0.795) when predicting one week in advance.

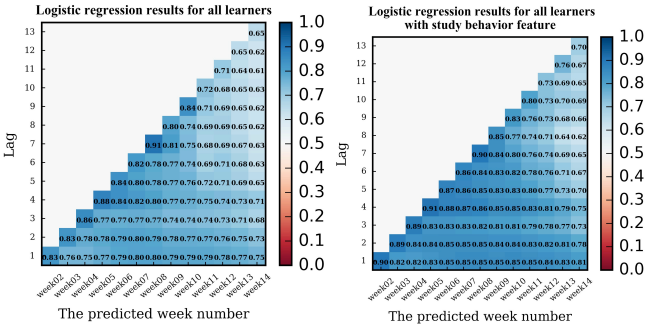

Figure 1. Heatmap for the logistic regression dropout prediction problem

From feature importance analysis as shown in figure 2, the study habits related behavioral features (x301-306) had played more important roles in the dropout prediction. Top features that had the most predictive power including *problem_finish_percent_pre_deadline24h*, *study_before_submit*, and *time_first_visit*.

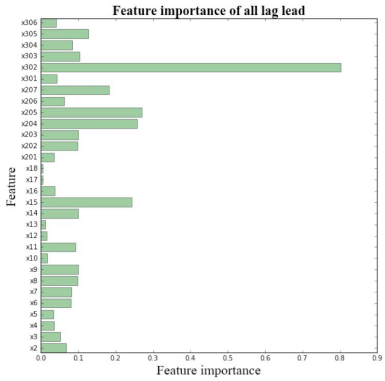

Figure 2. Feature importance

With new features related to study habits, the AUC of our predicting improved (figure 3), especially for the small sample size group (wiki / forum contributor and fully collaborative).

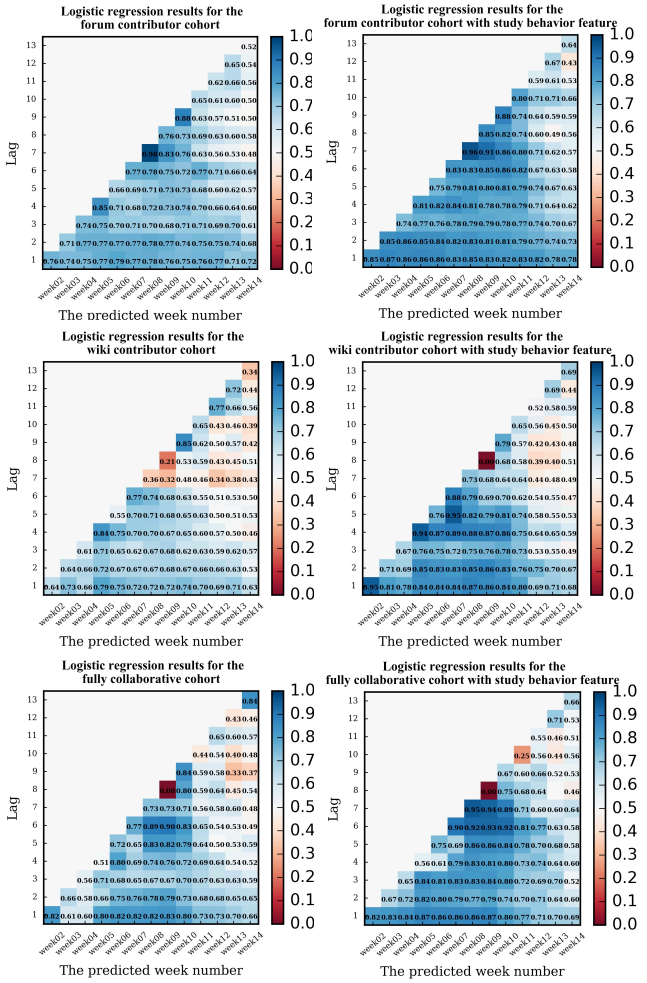

Figure 2. Heatmap for the logistic regression dropout prediction problem for three groups

In the future, we will try to using improved predictor each week within the course progress to deliver the intervention into small private online course.

5. REFERENCES

[1] Balakrishnan, G., & Coetzee, D. (2013). Predicting student retention in massive open online courses using hidden markov models. *Electrical Engineering and Computer Sciences University of California at Berkeley*.

[2] Whitehill, J., Williams, J. J., Lopez, G., Coleman, C. A., & Reich, J. (2015). Beyond prediction: First steps toward automatic intervention in MOOC student stopout.

[3] Loya, A., Gopal, A., Shukla, I., Jermann, P., & Tormey, R. (2015). Conscientious behaviour, flexibility and learning in massive open on-line courses. *Procedia-Social and Behavioral Sciences*, *191*, 519-525.

[4] Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2016, April). Big Data Application in Education: Dropout Prediction in Edx MOOCs. In *Multimedia Big Data (BigMM), 2016 IEEE Second International Conference on* (pp. 440-443). IEEE.

[5] Taylor, C., Veeramachaneni, K., & O'Reilly, U. M. (2014). Likely to stop? Predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382*.