Predicting MOOC Performance with Week 1 Behavior

Suhang Jiang
School of Education
University of California, Irvine
Irvine, CA 92697
suhangj@uci.edu

Adrienne E.Williams
School of Biological Sciences
University of California, Irvine
Irvine, CA 92697
adriw@uci.edu

Katerina Schenke
School of Education
University of California, Irvine
Irvine, CA 92697
kschenke@uci.edu

Mark Warschauer
School of Education
University of California, Irvine
Irvine, CA 92697
markw@uci.edu

Diane O'Dowd
School of Biological Sciences
University of California, Irvine
Irvine, CA 92697
dkodowd@uci.edu

ABSTRACT
Prior studies on Massive Open Online Courses (MOOCs) suggest that there is a significant decrease in student participation after one week of instruction [8]. This paper uses a combination of students' Week 1 assignment performance and social interaction within the MOOC to predict their final performance in the course. The study also examines the role external incentives in final MOOC performance. Using logistic regression as a classifier, we are able to predict the probability of students earning certificates for completion of the MOOC, as well as the type of certificate (i.e. Distinction and Normal) earned, with high accuracy.

Keywords

MOOC; Performance Prediction; Incentive

1. INTRODUCTION
In less than two years, Massive Open Online Courses (MOOCs) have attracted millions of learners to enroll in courses such as the History of Chinese Architecture, Information Visualization, and Healthcare Innovation and Entrepreneurship. These online courses provide a diverse set of learners with opportunities to engage in lifelong learning. Institutions of higher education are in a unique position concerning MOOCs; many universities such as MIT and Stanford supply the content of MOOCs, and many university students are encouraged to take MOOCs to expand their existing knowledge base or access courses that are not available at their home institution. However, against a background of thriving enrollment, the completion rate of MOOCs is staggeringly low [3]. Previous research indicates that fewer than 7% of learners who enroll in a MOOC actually complete it [6]. Of even more concern is that the significant decrease in participation usually takes place by the second week of the course [8]. If institutions of higher education want to take advantage of the potential of MOOCs in student learning, it is important to understand what variables affect the completion of these courses as well as at what kinds of interventions can be designed to encourage persistence. As such, this paper uses learner's behavior in the first week of a Biology MOOC to predict performance at the end of the course.

2. RELATED WORK
A number of studies have explored students' behaviors and learner's engagement patterns in MOOCs to understand issues of persistence. In one study, Balakrishnan [1] looked at student retention in a MOOC offered by UC Berkeley. He used variables related to the (1) cumulative time students spent in watching video, (2) the number of posts viewed in the forum, (3) the number of posts created on the forum, and (4) the amount of time spent on the progress page. Using these variable and Hidden Markov Models, Balakrishnan was able to predict students' likelihood to drop out of the MOOC. These measures of student engagement are likely factors in predicting student retention in MOOCs. While learner engagement seems to be a core component in answering the question of persistence in MOOCs, there still remain questions about other variables involved in understanding the complex problem of MOOC persistence. Patterns of student behavior in MOOCs can tell us something about the types of activities that are known to be engaging. For example, Kizilcec and his colleagues [4] identified four prototypical engagement patterns in a MOOC that consisted of watching videos and taking quizzes. These patterns were: (1) students who completed the majority of assessment, (2) students who engaged mainly in terms of watching videos, (3) students who did assessment at the beginning of the course, and (4) students who only watched videos for one or two assessment periods. The completing category of students is respectively 27%, 8% and 5% in the three sampled courses.

In addition to thinking about individual characteristics that predict persistence of MOOCs, an understanding of situated theories of learning is useful in examining learner persistence in MOOCs [5]. Situated learning theory posits that knowledge resides primarily in social interactions and only secondarily in the individual [2]. In the context of MOOCs, the social interactions learners have with one another via social networks could offer additional explanations for persistence in MOOCs that transcend individual learner characteristics. The role of social networking in MOOCs has been explored in a few MOOC studies. In an initial study, Yang and her colleagues [7] modeled how students' social positioning predicts their dropout using survival analysis. Nevertheless, little research integrates both the predictors of academic assessment and social interaction in modeling students' performance.

Finally, the larger context of incentives to complete a MOOC adds yet another dimension to our understanding of learner persistence. In traditional university environments, grades provide learners with a huge incentive to attend class and engage with the material. Such an incentive does not exist in MOOCs, where completion commonly results in a digital badge or a certificate that often contains little value. Often times, learners enroll in MOOCs out of intrinsic interest in the material and not because the MOOC is required for their undergraduate or graduate degree. The current study is uniquely situated within a context in which a subsample of the learners are incentivized to participate in the MOOC by their University. The study will examine how students who receive additional external incentives performed differently than the general population.

3. DATASET

We are analyzing an online course titled "The Preparation for Introductory Biology," offered by professors from University of California, Irvine (UCI) and hosted in Coursera. The duration of the course was four weeks, and comprised of three units. Each unit was composed of short videos, three to four quizzes, and three to four peer assessments. These quizzes were in the multiple-choice format with automatic, immediate feedback. Additionally, learners were able to re-take each quiz with new questions up to three times in order to improve their scores. This course offered two study tracks: a Basic track, which was based on the performance of ten quizzes and resulted in a normal certificate, and a Scholars track, which included peer assessments and resulted in a distinguished certificate. Students did not pre-select a track, but rather were considered to have completed the Scholars track if they fulfilled the requirements.

The online course was created with a primary goal of preparing incoming freshman for the onsite Bio 93 course at UCI. The failure rate for this onsite course is 15%, which results in a proportion of undergraduates having to retake the course in the following term. At UCI, students must obtain a Mathematics SAT score of at least 550 (out of 800) to be eligible for the Biological Sciences major. Students who enter UCI with a score below 550 enter the Undeclared major and must instead pass a full year of biology and chemistry before being eligible to enroll as a Biological Sciences major. In this particular year, Undeclared major students were incentivized to enroll and complete the Preparation for Introductory Biology MOOC. If Undeclared major students successfully completed the MOOC with Distinction, they would be able to enter the major after just one term instead of waiting a year. This provides us with a unique opportunity to investigate the effects of providing an incentive to a group of learners enrolled in a MOOC.

The data mainly come from the SQL file exported from the Coursera database, which include time-stamped datasets of learners' assignment (quiz and peer assessment) performance, scores, forum activities, and final performance. Another source of data come from UCI registrar, which records students who are enrolled in the onsite Bio 93 course. A reported 37,933 students signed up for the course and 551 students obtained Distinguish certificate for completing the Scholars track while 1,971 students obtained Normal certificate for completing the Basics track. Of the students who signed up, 35,411 students did not complete the course. 232 UCI Biology students and 172 Undeclared major students were identified to have signed up for the online course at the time of the study; we are still waiting to confirm a small proportion of these students.

4. DATA ANALYSIS

Our analysis consists of two logistic regression models, which predict student performance at the end of the course.

4.1 Feature Set

The predicted variable for the first model is the certificate the learner gets, i.e. a Distinction certificate or a Normal certificate. The predicted variable for the second model is whether the students get a Normal certificate or do not complete the MOOC, and thus receive no certificate.

The first predictor is the average quiz score learners obtained in the first week of the course. There are four quizzes in Unit 1 and the quiz score ranged from 0 to 6.

The second predictor is the number of peer assessments students completed in Week 1. We did not include the scores that students received from their peer assessors because those scores were not available until the second or the third week of the MOOC and therefore were not thought to affect MOOC persistence at week one.

The third predictor is learners' social network degree in the first week, which measures the level of social integration. The social network degree measures the local centrality of learners in the online learning community. It is calculated as the number of edges to which the node is connected. In this scenario, we treat learners as nodes and making comments to another learner's post is regarded as a directed edge from the commenter to the poster. The degree value is the number of connections that each learner has. Learners who did not participate in forums are assigned with 0 for their social network degree.

The fourth predictor is whether or not a learner is an incoming UCI Undeclared major student. This subgroup of students will go on to take the Bio 93 onsite course and have received external incentive to participate in the online course. We identified students as Undeclared by matching their school email addresses with Coursera accounts.

4.2 Logistic Regression

Two logistic regression models were run separately. The first logistic regression model predicts the type of certificate a learner received (Distinction or Normal certificates). The second logistic regression predicts whether students receive Normal certificates, or none at all. In the first model, the number of peer assessments taken in Week 1 is a strong predictor for achieving Distinction. For every unit increase in the number of peer assessments taken in Unit 1, the odds of getting Distinction are over 7 times larger than getting Normal, holding other predictors constant. Learners who are more active and well connected in the forum in the first week are more likely to receive Distinction than Normal certification, holding the number of peer assessments taken constant. For students taking the same number of peer assessments, the odds of UCI Undeclared major students getting Distinction are 89.4% higher than the rest of the learners. Table 1 indicates the odds ratio for the five predictors in the two logistic regression models.

Table 1 Odds Ratio

Odds Ratio	Class	
Variables	Distinction vs Normal	Normal vs None
Average Quiz Score	--	2.416***
Number of Peer Assessment	8.745**	1.054
Social Network Degree	1.192*	1.123
UCI Undeclared Major	1.894*	2.282**

Note. *** p<0.001 ** p<0.01 * p<0.05 .p< 0.1

In the second model, the average quiz scores in Unit 1 strongly predicted whether learners get Normal certificate or none. Learners' activity in the forum is no longer statistically significant in the predictive model. The odds of a UCI Undeclared major students getting Normal certificates are 2.282 times non-UCI-Undeclared major students, holding the average quiz score, the number of peer assessments completed and social network degree at fixed values.

Tenfold cross-validation was employed to estimate the predictive model. The first model predicting Distinction and Normal certificate earners achieved 92.6% accuracy. The second model predicting Normal certificate and no certification earners achieved 79.6% accuracy. Table 2 shows the evaluation of the predictive models.

Table 2 Model Evaluation

Evaluation	Modelling Distinction and Normal earners	Modelling Normal and None earners
Accuracy	0.926	0.796
ROC Area	0.947	0.851
Precision Positive	0.779	0.703
Precision Negative	0.978	0.911
Recall Positive	0.924	0.907
Recall Negative	0.927	0.713
Measure Positive	0.846	0.792
Measure Negative	0.952	0.800

5. DISCUSSION AND FUTURE WORK

The models indicate that assignment performance in Week 1 is a strong predictor of students' performance at the end of the course. The degree of social integration in the learning community in Week 1 is positively correlated with the achievement of Distinction certificates. Students with external incentive are more likely to complete the course compared to students in general, even in comparison with students who have similar backgrounds. However, this research is limited because we cannot control more variables that influence students' performance in the MOOC. Future research should focus on how to increase students' social integration and interaction in the online learning community, as these factors have been shown to influence student participation in MOOCs. More investigation into providing external incentive and increasing course relevance for the target audience is still needed. It is also worth exploring the effectiveness of the integration of online education and traditional face-to-face education in more depth.

Additionally, more experimentation and research into the relationship between quality of online courses and students' engagement and performance is recommended. Research from disciplines such as Education, Human Computer Interaction, and Computer Science should collaborate and redesign online education. The low engagement and completion rates reflect the existent opportunities for the improvement of online education.

6. ACKNOWLEDGMENTS

We are very indebted to the Digital Learning Lab, University of California, Irvine.

7. REFERENCES

1. Balakrishnan, G. and Coetzee, D. 2013.Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models. http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-109.pdf.
2. Greeno, J.G.1989.A perspective on thinking. *American Psychologist 44*, 2 (1989), 134–141.
3. Ho, A.D., Reich, J., Nesterko, S.O., et al. 2014. *HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013*. Social Science Research Network, Rochester, NY.
4. Kizilcec, R.F., Piech, C., and Schneider, E. 2013. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, ACM (2013), 170–179.
5. Lave, J. and Wenger, E. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.
6. Parr, Chris. 2013. New study of low MOOC completion rates | Inside Higher Ed. *Inside Higher Ed*. http://www.insidehighered.com/news/2013/05/10/new-study-low-mooc-completion-rates?utm_source=feedly.
7. Yang, D., Sinha, T., Adamson, D., and Rose, C.P.2013. Anticipating student dropouts in Massive Open Online Courses.
8. Hill, P. 2013. Emerging Student Patterns in MOOCs: A Graphical View. *e-Literate*. http://mfeldstein.com/emerging_student_patterns_in_moocs_graphical_view/.