Exploring Initial Dropout

Pavankumar Gurazada[*1], *Moutusy Maity* [1]

[1]*Indian Institute of Management, Lucknow (Noida Campus)*

9/16/2018

1 Introduction

Massive Open Online Courses (MOOCs) have received widespread attention since their launch in 2012. Since then, MOOCs evolved from being largely free to access to a pay-for-certification model. Several local universities and governments have also stepped in to the fray by offering several of their courses online, where learners might get a certificate based on their performance in the course for a fee.

In this document we focus on defining a research hypothesis concerning the drop-out rates of learners in MOOCs based on preliminary analysis of a large data set of about 500,000 learners who enrolled for 16 Harvard and MIT MOOCs for the period 2012-2013.

2 Dropout rate

For the purpose of our research we define drop-out rate as the fraction of learners who enroll for a course but do not finish with a certification. We note that drop-out rates are deal-breakers for government efforts like Swayam (India) where policy initiatives like upskilling the labor force might get derailed by high drop-out rates. Similar is the case of small private online courses offered to corporates by universities, where high initial drop-out rate can be devastating. Following this intuition, a central theme of our proposed research is the initial drop-out rate, i.e., the number of learners who register for a course but don't even watch a single video. Surprisingly, there is miniscule prior research on this peculiar phenomenon that, in our opinion, shares conceptual roots with the long-standing research tradition on the intention – behavior gap observed among consumers. On an average, the initial drop-out rate on the Harvard and MIT MOOCs is about 50%, which is disconcerting in the context of public policy initiatives. For-profit ventures are fine with high initial drop-out since they are a non-paying segment; for them initial drop-out is an artifact of low exit barriers (and is probably a boon since resources can be diverted instead to paying learners).

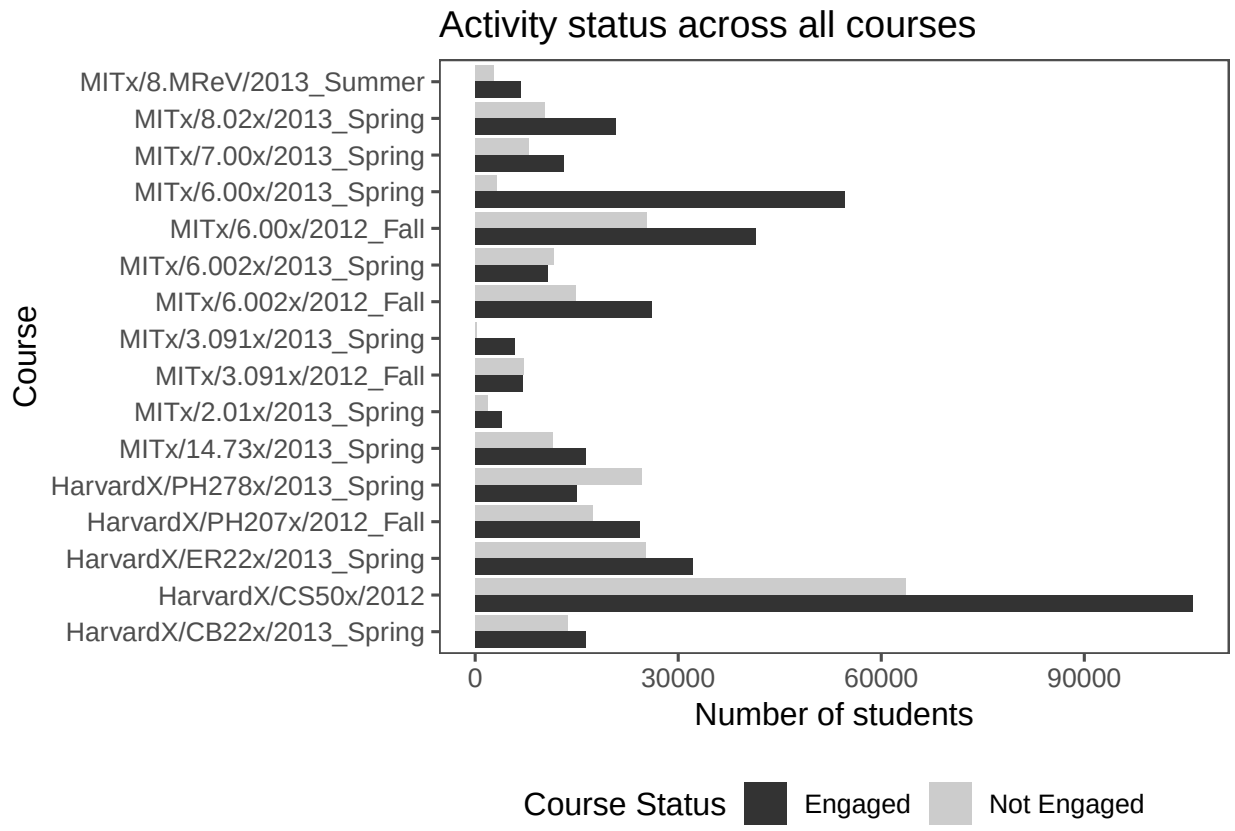
*Corresponding author: efpm04015@iiml.ac.in

2.1 Prior research

As the initial euphoria on MOOCs settled down, significant research interest is being focused (predominantly in the machine learning community) on peculiar aspects of learner behavior in these courses. While this area is still nascent, most early results make use of access to fine-grained data on learner behavior captured via clickstreams. A defining feature of MOOCs is the noticeably low certification rate (usually less than 4%) across courses and providers. There are two arguments proposed by scholars to explain this observation. First, there is a wide gamut of learners who enroll in MOOCs and looking only at certification rate of a course would not do justice to the utility gained by a learner from a MOOC. Second, even though the number of certifications is less in terms of percentages, absolute numbers are still many multiples of the number of students who complete a typical university course. This justifies the investments made into creating and promoting MOOCs. MOOCs are designed to have low barriers to both entry and exit unlike traditional education leading to a wide range of learner behavior. We note that most research attention has been focused on predicting the grades earned by learners who earn certificates based on the clickstream data. This focus on the paying segment is understandable since MOOC platforms (e.g., Coursera, edX) have limited resources that are exhausted once this segment is catered to.

2.2 Initial drop-out

To characterize initial drop-out on MOOCs, we define a learner as 'engaged' if they register and view at least one video from the course (it follows that those who browsed more than one video and those who earned certificates are also classified as 'engaged'). Similarly, a learner is classified as 'not engaged' if they register but never turn up, i.e., they drop-out even before starting the course. Figure 1 summarizes the distribution of learners into these two categories across the 16 categories. We note from Figure 1 that for every course, the 'not engaged' category is a worryingly high number. In order to probe this finding further, we attempted to predict initial drop-out based on the data available on the registered learners using logistic regression and random forests. The engagement status was used as the outcome (coded 1 for 'engaged' and 0 for 'not engaged) while age, gender, country and date of registration were used as predictors.



As an illustration we summarize the distribution of the learners who engage and those who do not as a function of their age and education in Figure 2 (a) and (b).

Following these descriptive measures we arrive at the following research question: RQ: Why do a large proportion of learners who voluntarily register for a MOOC not watch a single video (i.e., why is the initial drop-out rate so high?)

Our initial hypothesis is that date of registration, i.e., whether the learner registered early or late (relative to the fixed course start date) might be strongly predictive of initial dropout. We suspect that learners who register early might have done so in a moment of enthusiasm that dissipates by the time the course eventually starts. We incorporate this information into our models by adding 2 variables indicating whether the learner registered before or after the course launch, and how far the registration date was from the launch date. Similarly, we expect learner demographics to be reflective of the infrastructure available to a learner in order to successfully engage with a MOOC. For e.g., a learner from United States is expected to have access to better internet facilities compared to a learner from Rwanda. In considering these simple features, our hope is that if we can produce a model strongly predictive of initial drop-out, we will need no further exploration and can propose relevant interventions to lower initial drop-out. Since we want our models to be

strongly predictive of drop-out, we attach prime importance to misclassified learners. In particular, we want the number of learners who were misclassified as non drop-outs to be minimum. To enforce this we scored our models using the Cohen's Kappa score and selected the final model parameters and metrics based on 10-fold cross validation with 3 repeats. To account for the class imbalance we observe in Figure 1, we follow prior literature and employ the Synthetic Minority Over-sampling Technique (SMOTE) to form the training set (chosen as 80% of the entire data set). Insert Figure 3 here As can be seen from the values of Cohen's Kappa in Figure 3, the model with just the age, education, gender and registered date is far from being predictive of initial drop-out. For two of the courses, random forests (and logit) do a great job in predicting drop-out using the features we have considered, but we did not observe similar accuracy on the test data. Overall, random forests do improve predictive power over logistic regression, but not by a large amount. In sum, we conclude that key predictors of initial drop-out are missing from the data collected in the Harvard and MIT MOOCs. Given the current data set, we cannot crystallize key observable features that result in high initial drop-out. We note that there are other features in the data set relevant to learners who engaged with the data, e.g., number of chapters accessed, number of video play events that might predict eventual certification.