Why don't they show up?

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

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

October 02, 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. For e.g., 78 million learners took part in MOOCs in 2017, with the proportion of participants paying for courses increasing over previous years (Peters, 2018). Several local universities and governments have also stepped in to the fray by offering several of their courses online, where participants might get a certificate based on their performance in the course for a fee. Over the past couple of years MOOCs have even evolved into public policy initiatives for e.g., Swayam - a collaborative effort between the Government of India and several top universities in India (Bast, 2018) - with the aim of upskilling the labor force.

In this paper, we focus on measuring and predicting the *drop-out rate*, i.e., the fraction of participants 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 which 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*,

[*]Corresponding author: efpm04015@iiml.ac.in

measured as the number of participants 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 (Sheeran, 2002). The rest of the paper is organized as follows. The first section presents a review of existing literature related to the drop-out rates in MOOCs, and formulates the central hypothesis on predictors of initial drop-out. The second section presents the details and results from predictive models. The paper concludes with a discussion on the implications of this study and a proposal for further research.

2 Analyzing initial dropout rate

Dropout rates in MOOCs have been a widely researched area within the machine learning community. Much of this research relies on access to fine-grained data on participant behavior captured via clickstreams. We present an overview of key themes that emerge from related research in the next sub section.

2.1 Related research

As the initial euphoria on MOOCs settles down, significant research interest is being focused on peculiar aspects of participant behavior in these courses (Kross & Guo, 2018). A defining feature of MOOCs is the noticeably low certification rate (usually less than 4%) across courses and providers (Onah, Sinclair, & Boyatt, 2014). There are two arguments proposed by scholars to explain this observation. First, there is a wide gamut of participants who enroll in MOOCs and looking only at certification rate of a course would not do justice to the utility gained by a participant from a MOOC. For e.g., Belanger & Thornton (2013) show that the utility of participating in a MOOC goes beyond the attainment of certification and encompasses a quest to understand a subject, fun convenience or even exploration of a new learning medium. 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 (Kizilcec & Halawa, 2015). This observation is often cited as justification for the investments made into creating and promoting MOOCs. We submit that since MOOCs are designed to have low barriers to both entry and exit, a wide range of participant behavior can be observed.

Researchers have tackled dropout by modeling it as the dependent variable predicted by observable aspects of a participants behavior relying on access to clickstream data (Whitehill, Mohan, Seaton, Rosen, & Tingley, 2017). However, most research attention has been focused on predicting the grades earned by participants who earn certificates. This focus on the paying segment is understandable since for-profit MOOC platforms (e.g., Coursera, edX) have limited resources that are exhausted once this segment is catered to.

2.2 Data

Our research is based on the analysis of a large data set of $476,532$ participants who enrolled for 16 Harvard and MIT MOOCs for the period 2012-2013.

2.2.1 Descriptive statistics

[Figure 1 about here.]

As Figure 1 indicates, most of the participants are male. Also, USA and Asia contribute most to the number of participants.

To characterize initial drop-out on MOOCs, we define a participant 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 participant is classified as 'not engaged' if they register but never turn up, i.e., they drop-out even before starting the course. Figure 2 summarizes the distribution of participants into these two categories across the 16 courses.

[Figure 2 about here.]

We note from Figure 2 that for every course, the 'not engaged' category presents a significant challenge. On an average, the initial drop-out rate on the Harvard and MIT MOOCs is about 50%, which is disconcerting in the context of our earlier discussion on public policy initiatives.

Following these descriptive measures we arrive at the following research question:

RQ: Why do a large proportion of participants 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 participant registered early or late (relative to the fixed course start date) might be strongly predictive of initial dropout. We argue that participants 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 participant registered before or after the course launch, and how far the registration date was from the launch date. Similarly, we expect participant demographics to be reflective of the infrastructure available in order to successfully engage with a MOOC. For e.g., a participant from United States is expected to have access to better internet facilities compared to a participant from Rwanda.

Since we want our models to be strongly predictive of drop-out, we attach prime importance to misclassified participants. In particular, we want the number of participants 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 Oversampling Technique (SMOTE) to form the training set (chosen as 80% of the entire data set).

In the next section, we extend the discussion in this section by predicting initial dropout based on the data available on the registered learners using 4 algorithms - Logistic Regression, Gradient Boosting, Neural Networks 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.

2.3 Modeling initial dropout

random forest, xgboost and mlp here

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.

References

Bast, F. (2018). Learning on the go. *The Hindu*. Retrieved from https://www.thehindu.com/education/learning-on-the-go/article24418318.ece

Belanger, Y., & Thornton, J. (2013). Bioelectricity: A quantitative approach duke university's first mooc.

Kizilcec, R. F., & Halawa, S. (2015). Attrition and achievement gaps in online learning. In *Proceedings of the second (2015) acm conference on learning@ scale* (pp. 57–66). ACM.

Kross, S., & Guo, P. J. (2018). Students, systems, and interactions: Synthesizing the first four years of learning@ scale and charting the future. In *Proceedings of the fifth annual acm conference on learning at scale* (p. 2). ACM.

Onah, D. F., Sinclair, J., & Boyatt, R. (2014). Dropout rates of massive open online courses: Behavioural patterns. *EDULEARN14 Proceedings*, 5825–5834.

Peters, D. (2018). MOOCs are not dead, but evolving. *University Affairs*. Retrieved from https://www.universityaffairs.ca/news/news-article/moocs-not-dead-evolving/

Sheeran, P. (2002). Intention—behavior relations: A conceptual and empirical review. *European Review of Social Psychology*, *12*(1), 1–36.

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). Delving deeper into mooc student dropout prediction. *arXiv Preprint arXiv:1702.06404*.

List of Figures

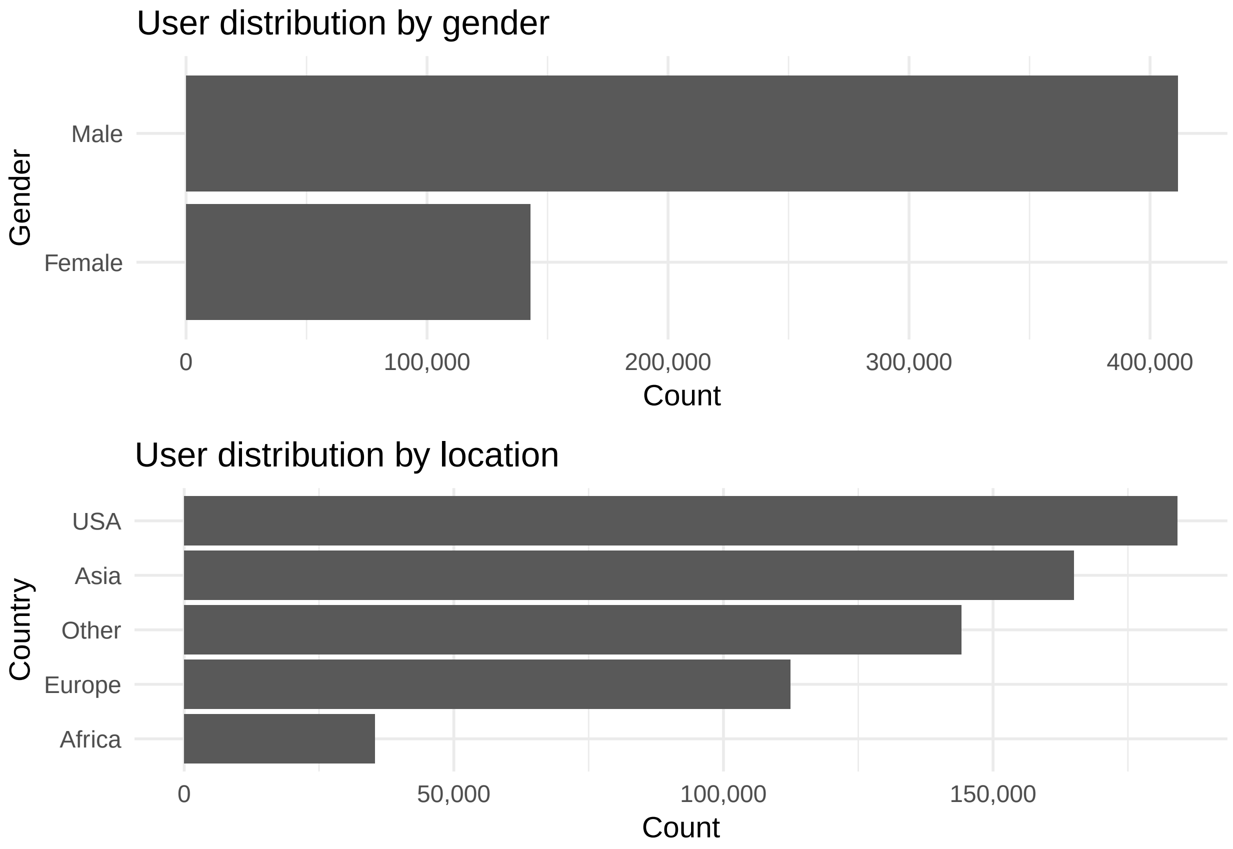

Figure 1: Descriptive statistics

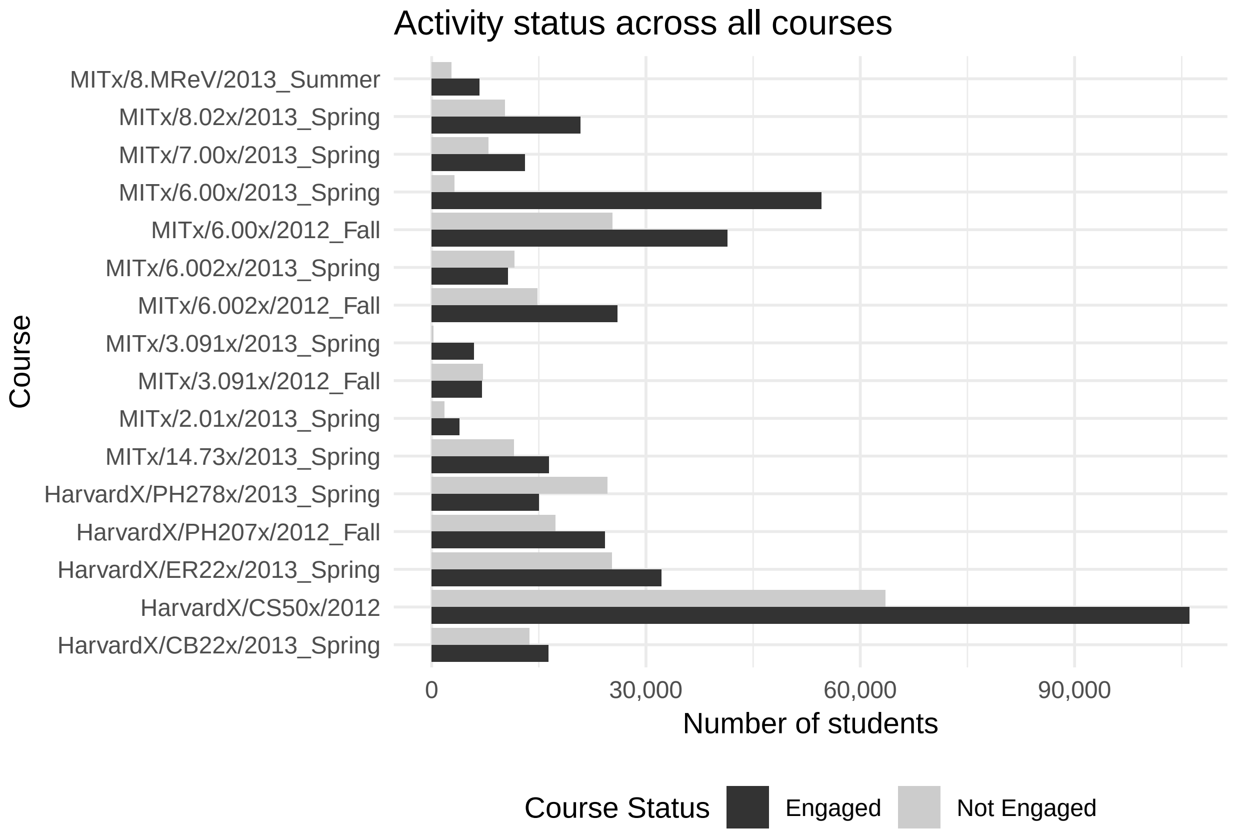

Figure 2: Activity Status