Why Don't They Show Up?

October 05, 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 with the aim of upskilling the labor force. A good example of such a effort is Swayam - a collaborative effort between the Government of India and several top universities in India (Bast, 2018).

In this paper, we focus on measuring and predicting the *drop-out rate*, i.e., the fraction of participants who enroll for a MOOC 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*, measured as the number of participants who register for a course but don't even watch a single video.

The rest of the paper is organized as follows. Section 2 presents a review of existing literature, and our central hypotheses on predictors of initial drop-out. Section 3 presents the details of our predictive models and results. The paper concludes with a discussion on the implications of this study and a proposal for further research.

2 Analyzing initial dropout

Dropout rates in MOOCs are 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 prior 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 they cater to this segment.

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.

[Figure 1 about here.]

As the descriptive statistics in Figure 1 indicate, 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.]

2.3 Predicting initial dropout

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. While there is little research on the drivers of initial dropout, some scholars have pointed out the unique aspects of MOOCs as one of the prime reasons for dropout. For e.g., in an analysis of factors that predict completion of a course, Yang, Sinha, Adamson, & Rosé (2013) argue that MOOCs have a unique development history. Starting from a small participant base upon announcement, new cohorts join in week after week. Consequently, the authors argue that if supportive communities do not evolve as the course progresses, participants might feel overwhelmed and drop-out.

In the context of initial dropout, we argue that there are two factors that might lead early registrants to drop out. First, the growth spurts and attrition that characterize the run-up to start of a course make comunity formation difficult (Yang et al., 2013). Second, participants who register early might have done so in a moment of enthusiasm that dissipates by the time the course eventually starts. Following these arguments, we expect that date of registration, i.e., whether the participant registered early or late (relative to the course start date) might be strongly predictive of initial dropout. 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.

Following the descriptive measures and the arguments presented in this section, the research questions we explore are:

RQ1: Early registration is strongly predictive of initial drop-out

RQ2: The country of residence of a participant is strongly predictive of initial drop-out

3 Modeling initial dropout

In this section, we present predictive models for initial drop-out using 3 methods - Logistic Regression, Gradient Boosting and Neural Networks.

3.1 Preprocessing

Engagement status was used as the label (coded 1 for ‘engaged’ and 0 for ‘not engaged) to be predicted in all our models. The features used to predict the engagement status were extracted from age, gender, country and date of registration. Following the discussion in section 2.3, we added a new variable `joined_early_or_late` by subtracting the start date of a course from the date of registration by the participant. We expect this variable to be the key predictor in our model. The other key variables included were age, gender and country (one-hot encoded). Finally, we excluded the participants who registered for more than one course (this allows us to justify the independent samples assumption underlying the models). The final sample size on which the models were evaluated is $n = 316,917$.

3.2 Model execution

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 AUC of the ROC curve as the metric and selected the final model hyperparameters 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).

Logistic regression was executed using L2-regularization with the regularization strength as the hyperparameter. For Gradient Boosting, we use the `xgBoost` algorithm (Chen & Guestrin, 2016), with depth of the trees and the number of estimators tuned as hyperparameters during training (we performed a random grid search over a larger parameter space to arrive at a shortlist). The neural networks were composed of a sequence of fully connected layers with the number of units per layer and the number of layers tuned during training. The final model comprised 9 fully connected layers, with 32 units in each layer (total number of trainable parameters $= 4,417$).

3.3 Results

[Figure 3 about here.]

As can be inferred from 3, Gradient Boosting and Neural Networks perform much better than Logistic Regression. However, since the Gradient Boosting model had a higher accuracy on the test set (76.1%) we choose this as our final model.

4 Discussion

The results presented in Section 3 indicate that the Gradient Boosted Model is a good model for the data set. In order to address the two research questions posed in Section 2, we now move to probe the relative importance of the factors in the predictive ability of the model on the data set. Relative importance of a feature is computed by averaging (across all trees) the number of times this feature is selected for the split weighted by the improvement to the resulting model (Elith, Leathwick, & Hastie, 2008). Figure 4 shows the relative importance of the top 5 features that contribute most to the predictive ability of the final Gradient Boosting Model on the test set.

[Figure 4 about here.]

Figure 4 indicates that the most important feature to the prediction of initial drop-out is whether the participant joined early (validating RQ1). Further, contrary to our expectation summarized in RQ2, our results indicate that education and country are not on the same scale of importance as early registration. This is suprising since we expect participants with higher education levels to be more disciplined in terms of attending and completing MOOCs (given that they have attended and completed courses within their formal education). The emergence of age as a stronger predictor is a surprising finding we wish to explore in further research.

In sum, we conclude that early registration is a key predictor of initial drop-out. This has important implications for the marketing and execution of MOOCs. In practise, we observe that little attention is paid to monitoring of participants before the course begins and our results indicate that neglecting this phase has a devastating effect on the drop-out rate. Our results indicate that early registrants need to be watched carefully and nudged to begin the course. We submit that this is paramount for government initiatives, and we strongly advocate the allocation of resources and attention to participants who have registered early.

Since this study is one of the first to probe initial drop-out, it is limited by the analysis of the secondary data. However, by using powerful machine learning methods, we are able to crystallize key features decicion makers should be aware of while launching MOOCs. In the next phase of research, we wish to undertake primary research to understand the decision-making process of MOOC drop-outs and incorporate the findings from this research into better predictive models of initial dropout.

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.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813.

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/

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

Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 nips data-driven education workshop* (Vol. 11, p. 14).

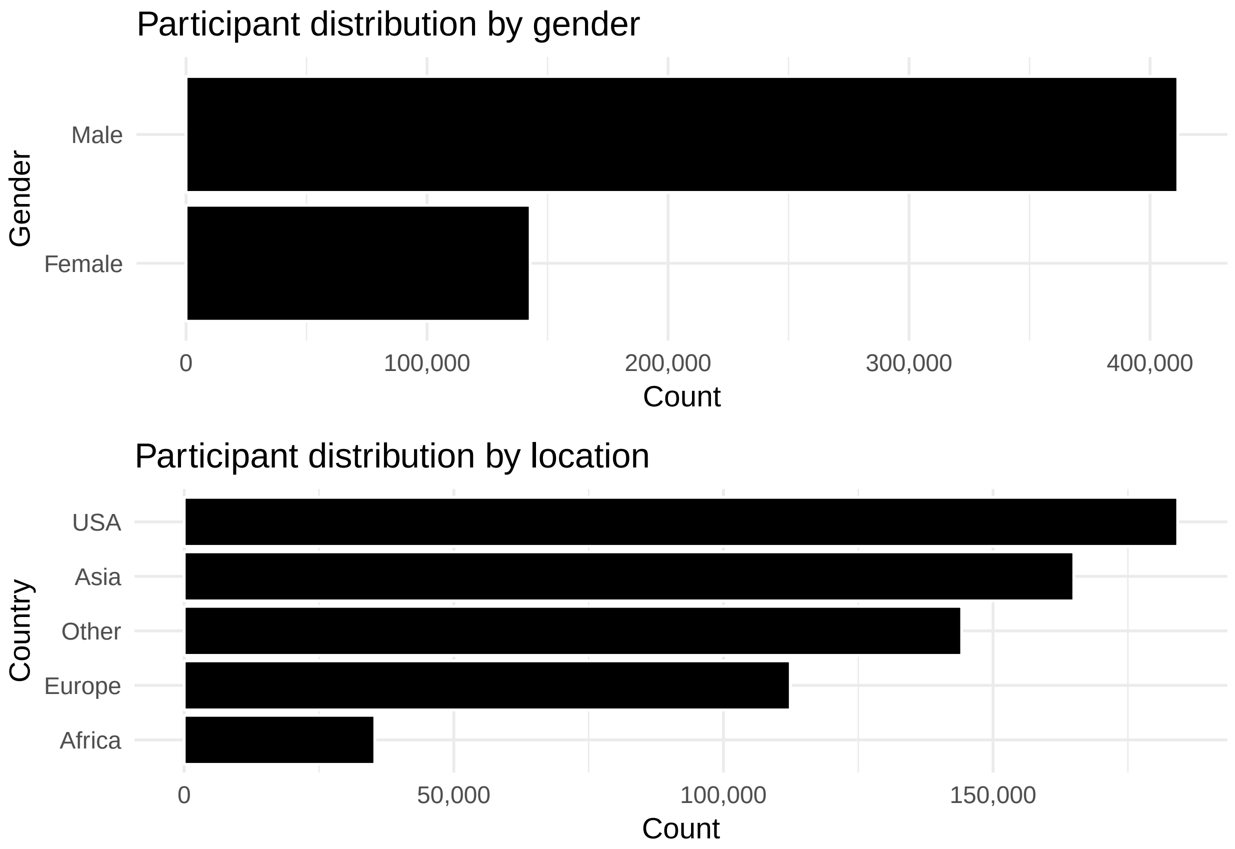

Figure 1: Descriptive statistics

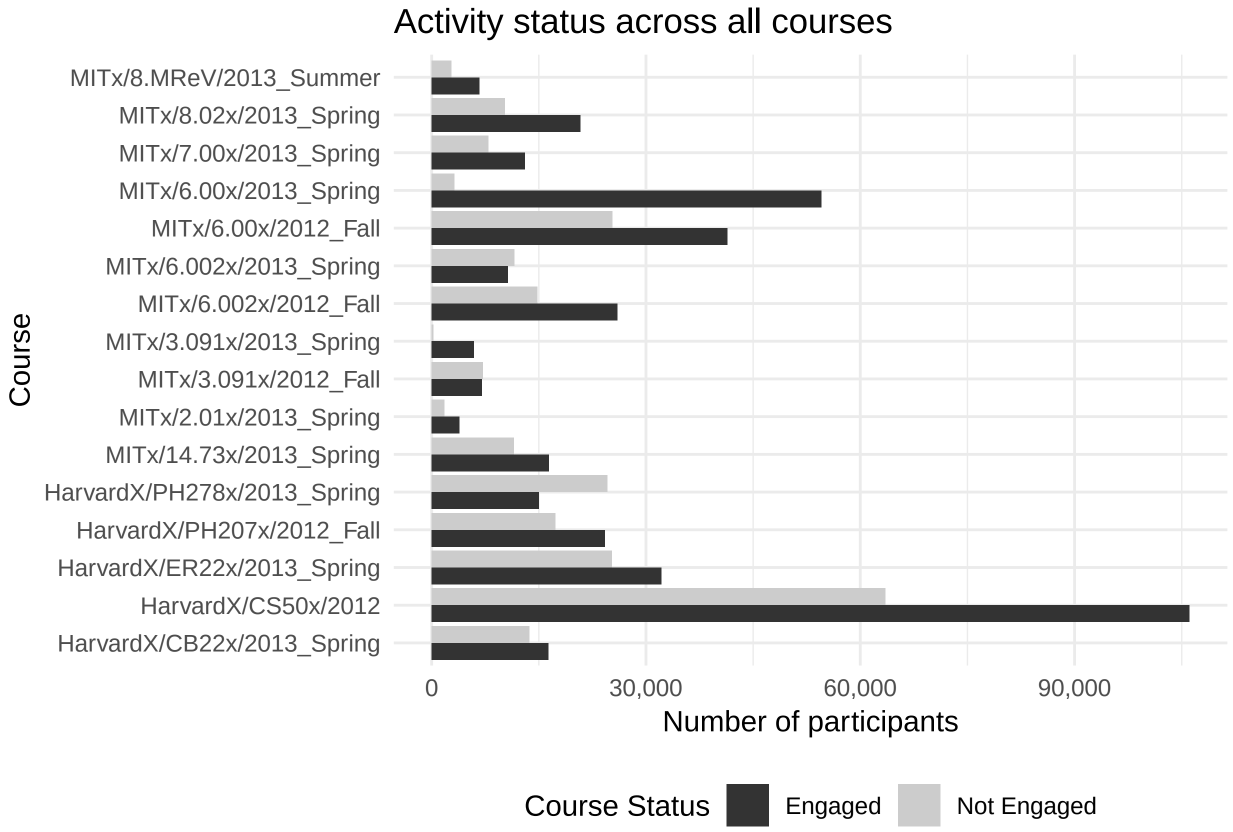

Figure 2: Activity Status

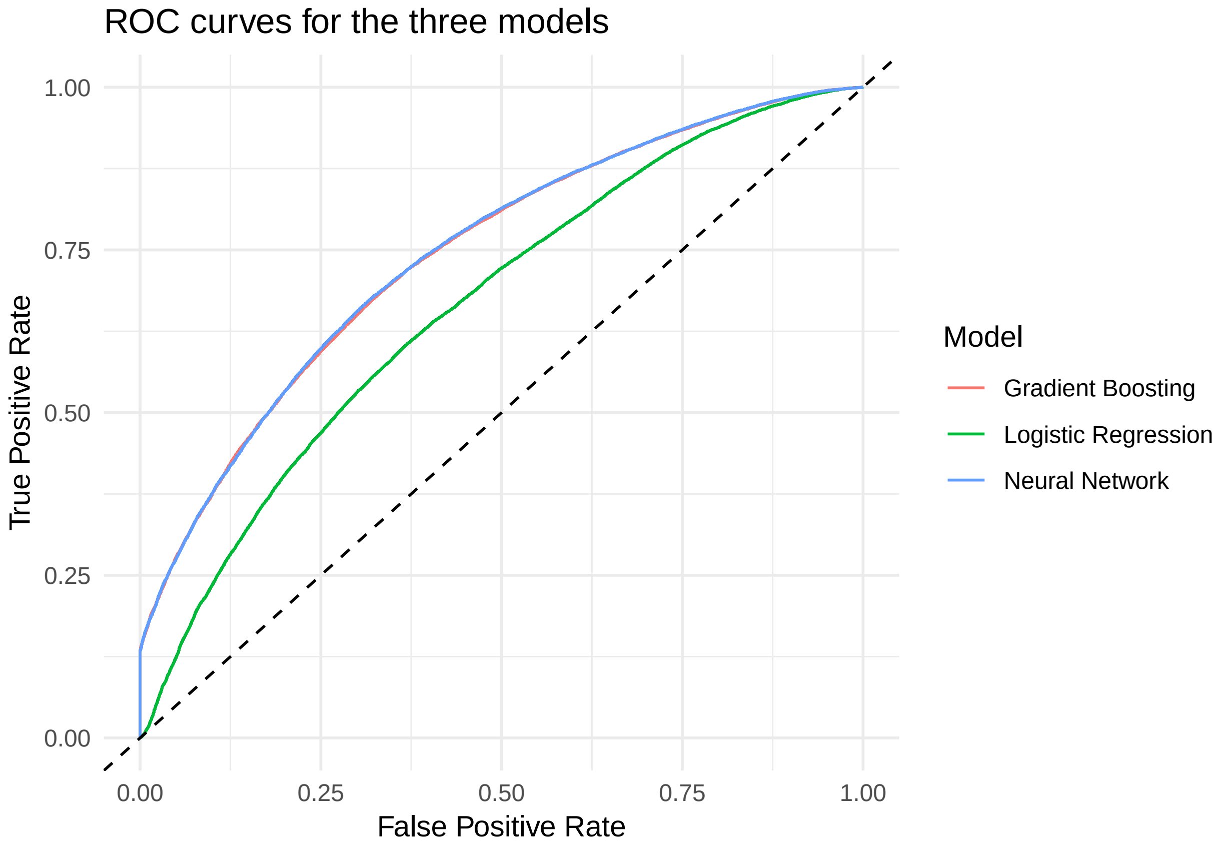

Figure 3: ROC curves for the 3 model fits

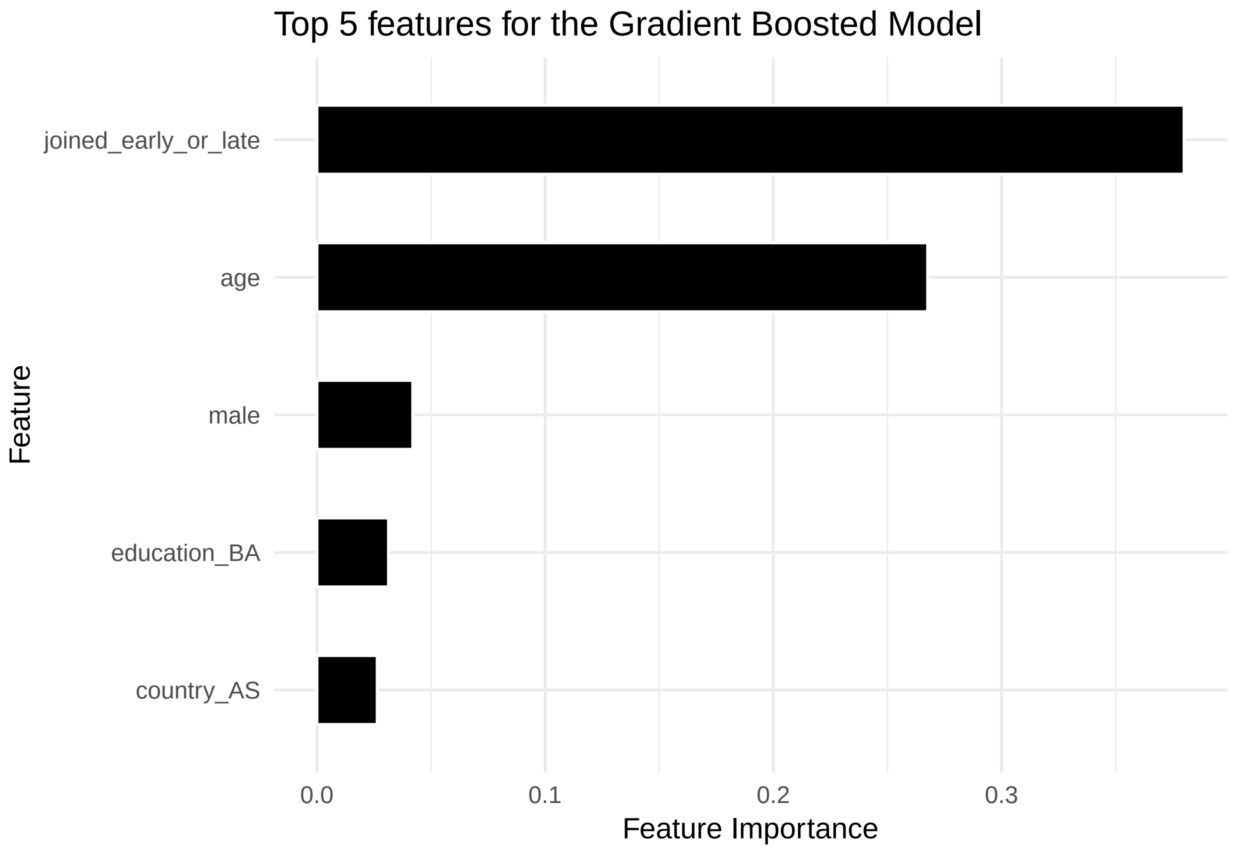

Figure 4: Feature importance plot of the Gradient Boosted Model