**Introduction**

Founded by Harvard University and MIT in 2012, **edX** is an online learning destination and MOOC provider. Last year, a joint research team from Harvard and MIT released a report, *HarvardX and MITx: Four Years of Open Online Courses* and presented their findings about the evolution of the MOOC by using many descriptive statistics. After reading their report, I started to wonder **what kind of MOOC classes are most popular** and **what are some characteristics of the course that can motivate MOOC participants to finish the whole course** instead of just accessing some of the course contents.

**Research question**

In this project, I explored two questions about edX by using a linear regression model and a logistic regression model:

1. What kind of classes attract most participants?
2. What kind of classes lead to the highest probability of completing the whole course?

**Data description**

Attached with the report, this dataset is so far **the largest survey of MOOCs**. It contains data of 290 edX courses, about 4.5 million participants, and different variables about the course. For this project, I only include variables that are useful to answer my research questions.

Outcome variables:

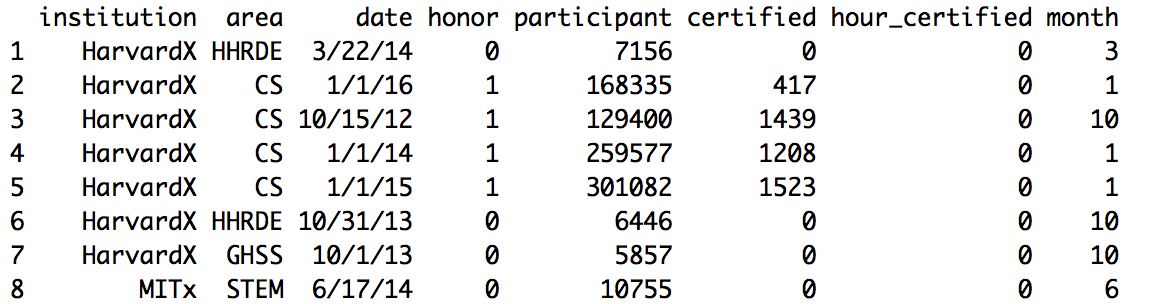
1. participants – number of participants (any registrant that has accessed course content in any chapter is called a participant)
2. perc\_cert – the percentage of participants who completed the whole course
   1. There’s no variable in the original dataset that represents people who finished the whole course. For this analysis, I subset courses that offer honor code certification and used the percentage of participants who got certificates as the variable of the percentage of participants who completed the whole course.
   2. edX offered two types of certificates: honor code certificates and verified certificates. As long as the participant has successfully completed a course he or she can get the honor code **certificate for free without verifying the identity**. So I assume every participant who completes the course gets certified if the course offers the honor code certification.
   3. I’ll talk about limitations later.

Predictors:

1. institution - Harvard/MIT
2. area – curricular areas
   1. GHSS = Government, Health, and Social Science
   2. HHRDE = Humanities, History, Design, Religion, and Education
   3. STEM = Science, Technology, Engineering, and Mathematics
   4. CS = Computer Science
3. honor – if the course offers honor code certification
   1. As explained in the last section, not all courses offer both honor code certification and verified certification
4. season – when the course launch
   1. Spring: February, March, April
   2. Summer: May, June, July
   3. Fall: August, September, October
   4. Winter: November, December, January
   5. The original dataset gives the exact date when the course launch, but that will be too specific to be a predictor. So I extracted the month information from the date and tried to use month as a predictor. However, after EDA, I saw varied numbers of cases for each level of month. Using month as the predictor will probably lead to an overfitting, so I lumped together several levels of month and created the predictor “season”.
5. hour\_certified - median hours used to get certified; an indicator of the workload of the course

Data errors:

There are eight courses in the dataset where the median hours of certified is 0, so I looked specifically at these eight cases.



We can see from the above table that the first, sixth, seventh, eighth rows have 0 median hour\_certified because there were no participants who ever got certified. Therefore, for these four rows, it’s not a data error problem. However, I still decided to remove these four observations because I intended to include hour\_certified in my model as an indicator of the workload of the course while these four courses don’t have indicators of workload. For the other four observations, I can’t think of any reasons to explain why the median hour\_certified is 0, so I treated them as data errors and removed them. However, it’s worth noticing that these four observations are all Harvard CS classes.

**Model 1: What kind of classes attract most participants?**

Selected Exploratory Data Analysis

Model Selection

Interpretation and discussion of results

Holding all else constant, the median number of participants increases by a factor of 1.89 (95%CI: 1.39, 2.64) if a course offers honor code certification. Different from verified certificates, honor code certificates are very attractive to participants because they are free of charge and easy to get (they don’t require ID verification). It’s a great way for participants to prove to their employers and schoolmates that they’ve gained new skills outside the workplace/school. People are encouraged to participate if the course offers honor code certifications.

Holding all else constant, the median number of participants increases by a factor of 1.38 (95% CI: 1.24, 1.54) if the median hours needed to get certified is doubled. Perhaps, courses that require more efforts are more helpful for MOOC participants, so more participants take those courses.

Holding all else constant, compared to a CS course, the median number of participants of a HHRDE course decreases by a factor of 0.44 (95% CI: 0.27, 0.73) and the median number of participants of a STEM course decreases by a factor of 0.37 (95% CI: 0.23, 0.58). The difference of median numbers of participants between a CS course and a GHSS course isn’t statistically significant (p value: 0.14431). From this result, we can see that CS courses and GHSS (Government, Health, and Social Science) courses are more popular than courses in other areas on edX.

To sum up and answer my research question, on edX, GHSS and CS courses that offer honor code certifications and require longer hours of efforts attract more participants.

Limitations and the scope of inferences

The R squared of my model is only 0.2419, so my model only explains about 24% of variations. There probably are other variables out there but not included in this dataset that are more useful for predicting the number of participants. Also, I used the median hours needed to get certified for each course as an indicator of the workload in this analysis. This probably is not so accurate for courses that have relatively fewer participants. While some courses in this dataset have more than ten thousand participants, some courses only have around 400 participants. Another limitation is that compared with other categories, CS has relatively fewer cases which might lead to overfitting problems. I chose to make CS a separate group to see the significant popularity CS courses have in this analysis, but we shall be cautious of possible overfitting problems.

This analysis uses dataset of edX, so it has some implications for other MOOC platforms, but very limited. Insights gained from this analysis can be useful for other MOOC platforms. For example, offering honor code certification is a great way to attract more participants. If a MOOC platform wants to attract more participants, it should consider to promote its courses by offering free certifications. However, some results are limited in the scope of edX. For example, the reason why CS and GHSS (Government, Health, and Social Science) courses attract most MOOC participants on edX can probably be that edX offers more high-quality courses in GHSS and CS than other areas. We need to be careful when making inference for MOOC platforms besides edX.

**Model 2: What kind of classes lead to the highest probability of completing the whole course?**

In my dataset, instead of giving 0 or 1 as the outcome of completing the course, it gives the percentage of those who completed the course (who earned the certificates). Therefore, in this analysis, I used binomial logistic regression model, which is a little different from the bernoulli logistic regression that we learned in class. While these two logistic regression models share similar analyzing methods, binned residual plots no longer apply for binomial logistic regression. To do EDA for continuous variables, I’ll use scatterplot instead.

Interpretation and discussion of results

Holding all else constant, compared with courses offered by Harvard, for courses offered by MIT, the odds of completing the course are multiplied by a factor of 0.77 (95% CI: 0.76, 0.78). This finding is interesting, but hard to think of why.

Holding all else constant, compared with CS courses, the odds of completing GHSS courses are multiplied by a factor of 1.01 (95% CI: 1.00, 1.03), the odds of completing HHRDE courses are multiplied by a factor of 0.95 (95% CI: 0.93, 0.96) and the odds of completing STEM courses are multiplied by a factor of 0.56 (95% CI: 0.55, 0.57). Therefore, CS, GHSS and HHRDE have similar completion rates while STEM has a way lower completion rate, probably because STEM courses are more difficult for self-learners.

Holding all else constant, when increase hour\_certified by one hour, the odds of completing the course are multiplied by a factor of 1.00 (95% CI: 1.00, 1.00). This indicates that the probability of completing the course doesn’t change with the workload of the course.

To sum up and answer my research question, on edX, STEM courses have a significant lower possibility for participants to stick with the course and finish it and MIT courses have a lower possibility for participants to complete the course than Harvard courses. Median hours needed to get certified don’t seem to influence the possibility of completion.

Limitations and the scope of inferences

First, according to the residual plots, my model in this analysis doesn’t explain a group of points well, although I’ve tried a bunch of transformations. So we shall be skeptical about the conclusion drawn from this model. Another major limitation in this analysis is that, I assume all those completed the whole course will earn an honor code certificate for themselves, which is not necessarily true for every single participant of edX. Furthermore, I subset courses that offer honor code certificates to do this analysis. This might lead to some unknown bias in the model. Lastly, like in the first model, I used the median hours needed to get certified for each course as an indicator of the workload in this analysis. This probably is not so accurate for courses that have relatively fewer participants.

This analysis has many limitations even within the scope of edX platform, so I don’t think it’s appropriate to do inference for other MOOC platforms. However, MOOC platforms can be inspired by insights like STEM courses might have a lower possibility of finishing the course and further investigate these possible situations.