Investigating Learner Disengagement IN Massive Open Online Courses(MOOCs)

By : Sarvesh MEENOWA

Supervised by: Dr Matthieu Cisel

Bachelor of Data Science by design

January 8, 2022

Abstract

This paper focuses on the evolution of learning engagement patterns and learner profiles in the Effectuation MOOC. We saw bystanders accounting for the majority of the audience.We also found that the statistical relationships between course completions and video consumption were influenced by socioeconomic status, country of residence, to some extent the age groups and the type of learners, suggesting a dependence on the structure or context of the course.

Keywords: Engagement,completion, MOOC, socioeconomic, HDI

Contents

List of Figures

List of Tables

1 Introduction

The great diversity of enrollees regardless of their socioeconomic status, sociocultural context, motivations, or behaviors, of massive open online courses(MOOCs) is indubitably one of its most impressive consequences. Their patterns of engagement are as disparate as their backgrounds, the simple comparison of completers and discontinuers is not necessarily accurate in describing the diversity of the situation. In general, a substantial proportion of enrollees may still account for a considerable portion of course participation, even if they do not finish the course. Although these topics have received significant coverage by researchers and practitioners, few studies have focused on the long-term progression of these patterns of learning engagement in a particular course. These questions are pertinent to both course creators who are keen to grasp the unfolding dynamics and tailor the course structure accordingly, and to researchers who wish to identify trends on a more holistic scale. In this paper, we analyzed a MOOC that has been organized on Canvas before being on Coursera, the MOOC Effectuation, to try to address the question of the evolution of learners' profiles and course dynamics over time. To what extent have engagement patterns and registrants profiles evolved across iterations, and most importantly, how has the relationship between learners' behavior and profiles evolved over time?

2 Methods

2.1 Source of data and Course description

The case studies we analyzed in this paper are a five weeks long entrepreneurship course called Effectuation (Professor Philippe Silberzahn, EMLYON Business School). It was hosted by a MOOC agency that used the open-source LMS Canvas from Instructure.

It was necessary to submit a peer-evaluated mid-term assignment and to pass an exam to earn the certificate. In both courses, new course material including quizzes and half a dozen of short videos were made available every week. Variations among iterations were minor. Course designers estimated that completing the course required fifteen to twenty-five hours. Student activity reports, grade books and survey responses were downloaded from the platform. Regarding video, consumption, we used a proxy as we considered that the video had been viewed when the page where it was, embedded was opened, regardless of the number of times this page was loaded. We manually removed from, subsequent analyses the videos that were not part of the course strictly speaking, such as weekly introductions or, tutorials. The global activity of the course was defined from the video perspective as the total number of views, without taking into account multiple views, and from the quiz perspective as the total number of submissions, without taking into account multiple submissions. Participants were asked to fill in a survey at the beginning of the course; response rates ranged from 40 percent, to 60 percent of enrollees. IP addresses were not collected; all available data on countries of residence come from these, surveys; the Human Development Index of these countries were retrieved from U.N data. In both

courses, the students who could gain credits by completing the course were excluded from our analyses since they, were not strictly speak ing following a self-directed learning approach. They represented a significant contingent, in the case of MOOC2.

2.2 Feature engineering

We have data for three iterations, therefore we patch together all the iterations by performing full-joins. Participants were categorized based on their level of engagement: those who obtained a certificate were called "completers", those who submitted at least one quiz or assignment but did not complete, the course was referred to as "disengaging learners"; those who did not submit any quiz or assignment was, referred to as "auditing learners" if they had viewed at least 10 percent of available course videos, and bystanders, if they fell below this threshold.A new variable "learner" is created based on described criterias representing the 4 types of learner. The latter allows us to compare the learners' disengagement via survival analysis.

We also created videos decile and video status(1 or 0), if the decile is less than 10, then the status is 1 else if the videos decile is 10, the status is 0. This is done in order to prepare the data so that they are fit for survival analysis. Another indicator built was age group, we divided age into 3 groups; 0-30, 31-50 and 51-80 to be used in logistic regression.

The whole analysis was carried out with the open-source statistical software R 4.0.3.

3 Results

3.1 Describing behavior in the courses

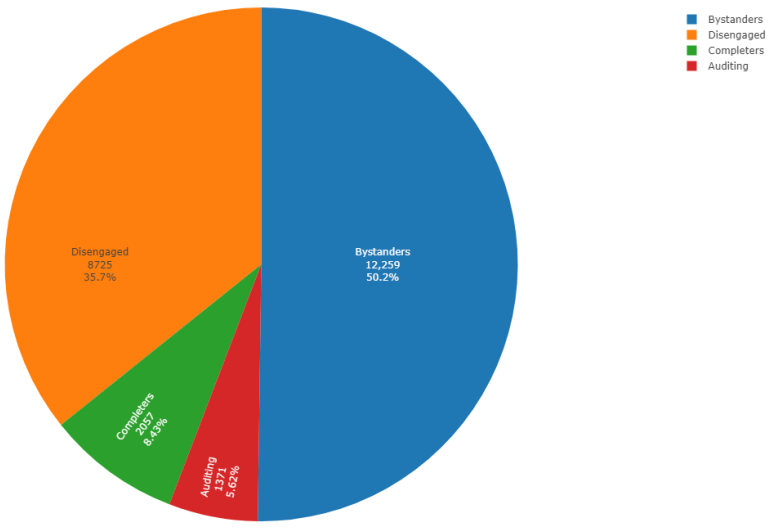

Figure 1: Proportion of the type of learners : Disengaging learners, auditing learners, bystanders,and completers.

Learners were categorized according to their level of engagement. Figure 1 represents the 4 types of learners; Auditing, Disengaged Learners, Bystanders and Completers. After merging all the three iterations of the effectuation database, there were 24412 participants. Bystanders represent the maximum proportion of learners, out of 24412 participants, around 50% or 12259 learners did not complete at least one assignment or quiz and they fell below the threshold of at least 10 percent of available course videos. The second category being disengaged represent slightly more than one-third of the participants at 35.7%, which means 8725 learners submitted at least one quiz or assignment but did not complete the course. The second least proportion of learners, the completers account for slightly more than 8% of the participants, which translates to 2057 learners who have completed the course. Lastly, auditing learners depict the least of the proportion of students at a little more than 5%, which means 1371 learners did not submit any assignment or quiz but have watched at least ten percent of the available course videos.

3.2 Linear models to identify factors on video consumption

-	t	df	95% CI	p-value	Mean(Male)	Mean(Female)
Equal Variance not assumed	-3.5	6247	-1.6 - -0.5	***	15.6	16.6
Equal Variance assumed	-3.5	9929	-1.6 - -0.5	***	15.6	16.6

Table 1: Two Sample independent t-test between the number of videos and gender with 95% confidence interval(CI) assuming both equal and unequal variances.
p-value $<$ 0.05: * / p-value $<$ 0.01: ** / p-value $<$ 0.001: *

In order to compare the number of videos watched between genders, from Table 1, a two-sample or unpaired t-test is performed since the means from two independent groups are being compared. In both of the assumed cases of equal and unequal variances, we can see that on average females watch more videos than males(p-value $<$ 0.001).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HDI	2	1197320	598660	6836	$<2.2^{-16}$
Residuals	28373	2484640	88		

Table 2: One-way ANOVA between number of videos and HDI(Human Development Index).

We also investigated if the countries' HDI levels affected the number of videos watched. To do so, one-way ANOVA is performed rather than a t-test since the categorical factor HDI has 3 levels(low, intermediate and very high). From 2, we can see that there is a difference in the means of the number of videos watched by the the different levels of HDI($F_{2,28373} = 6836$,p-value $< 2.2^{-16}$).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	2251	2251	13.4	<0.001
HDI	2	102869	51434	307.1	$<2.2^{-16}$
Gender:HDI	2	1259	629	3.8	0.02
Residuals	9831	1646367	168		

Table 3: Two-way ANOVA with interaction parameter(Gender*HDI).

We refined our linear model by accounting for the interaction parameter of gender*HDI(Table 3). We can see that there was an effect of gender and HDI on the number of videos watched($F_{2,1259}$=3.8, p-value =0.02), however since the interaction effect is significant(p-value = 0.02), the effect of the HDI of the country of residence cannot be generalised for both males and females together. Further tests must be performed such as Tukey's HSD (honestly significant difference) post hoc test in order to know between which groups of gender and HDI the results are significant.

	term	contrast	diff	conf.low	conf.high	adj.p.value
1	Gender:HDI	une femme:B-un homme:B	2.0	-1.1	5.1	0.5
2	Gender:HDI	un homme:I-un homme:B	5.6	3.5	7.7	$<2.2^{-16}$
3	Gender:HDI	une femme:I-un homme:B	3.0	0.3	5.6	<0.05
4	Gender:HDI	un homme:TH-un homme:B	9.5	8.2	10.8	$<2.2^{-16}$
5	Gender:HDI	une femme:TH-un homme:B	9.5	8.1	10.9	$<2.2^{-16}$
6	Gender:HDI	un homme:I-une femme:B	3.6	0.2	7.0	<0.05
7	Gender:HDI	une femme:I-une femme:B	1.0	-2.8	4.7	0.9
8	Gender:HDI	un homme:TH-une femme:B	7.5	4.5	10.4	$<2.2^{-16}$
9	Gender:HDI	une femme:TH-une femme:B	7.5	4.5	10.5	$<2.2^{-16}$
10	Gender:HDI	une femme:I-un homme:I	-2.6	-5.6	0.3	0.1
11	Gender:HDI	un homme:TH-un homme:I	3.9	2.1	5.7	$<2.2^{-16}$
12	Gender:HDI	une femme:TH-un homme:I	3.9	2.1	5.8	$<2.2^{-16}$
13	Gender:HDI	un homme:TH-une femme:I	6.5	4.1	8.9	$<2.2^{-16}$
14	Gender:HDI	une femme:TH-une femme:I	6.6	4.1	9.0	$<2.2^{-16}$
15	Gender:HDI	une femme:TH-un homme:TH	0.1	-0.8	1.0	1.0

Table 4: Tukey HSD (Honest Significant Test) post hoc test on the interaction between gender and HDI(Human Development Index). B:(low HDI), I:(Intermediate HDI),TH:(Very high HDI).

We performed Tukey HSD post hoc test on the interaction between gender and HDI(Table 4). We can see 15 different pairwise comparisons of the interaction between gender and HDI. Both males and females from intermediate and very high HDI countries watch more videos on average compared to males from low HDI countries (p-value $< 2.2^{-16}$). We also see similar results when we compare males and females from intermediate and high HDI countries, they consume more videos than females of low HDI countries(p-value $< 2.2^{-16}$), with the exception of having no significant difference between females of intermediate and low HDI countries. Likewise, males and females from very high HDI countries consume more videos than intermediate HDI countries(p-value $< 2.2^{-16}$). However, there was no evidence to suggest a difference in the number of videos consumed between the genders in countries with similar HDI.

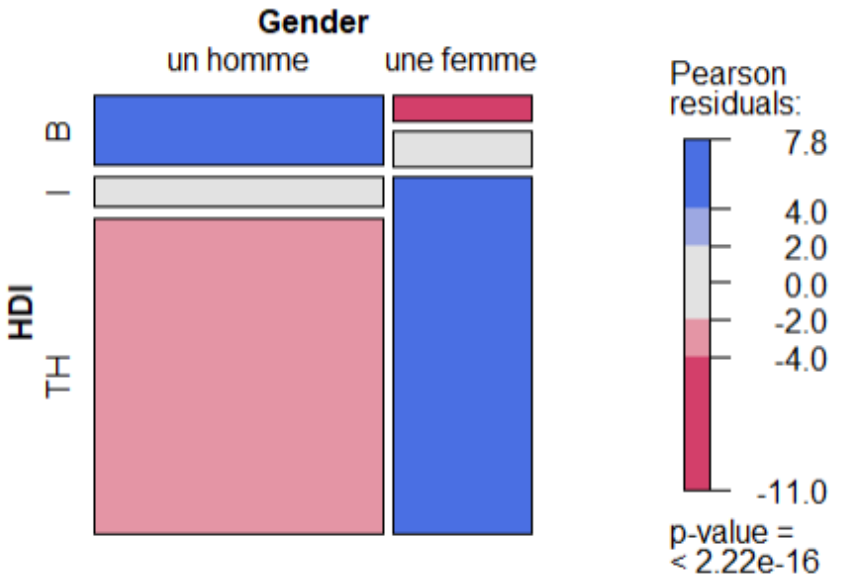

Figure 2: Mosaic plot to assess collinearity between HDI and gender.

In order to assess collinearity between our independent variables gender and HDI, we produce a mosaic plot(Figure 2) which is a visual representation of a contingency table for a chi-squared test. The units are in standard deviations, so a residual greater than 2 or less than -2 represents a significant departure from independence. There we can see that males from low HDI and very high HDI are dependent(p-value $< 2.2^{-16}$) and females and HDI(low,intermediate and very high) are also dependent (p-value $<2.2^{-16}$).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	1676	1676	10.0	**
HDI	2	90880	45440	270.3	***
Socioeconomic Status	6	6437	1072	6.4	***
Residuals	9380	1576694	168.		

Table 5: Three-way ANOVA with Gender, HDI and socioeconomic status as explanatory variable
p-value $<$ 0.05: * / p-value $<$ 0.01: ** / p-value $<$ 0.001: *$**

We then expanded our linear model to add socioeconomic status as another explanatory variable(Table 5), we saw that indeed the video consumption is influenced by socioeconomic status of the registrants($F_{6,9380}$=6.4,p-value $<$ 0.001). However, we must perform Tukey HSD test for instance in order to see the pairwise differences between learners of different socioeconomic status.

contrast	diff	conf.low	conf.high	adj.p.value
Others - Lower management jobs	-0.8	-4.2	2.6	
Employees - Lower management jobs	-1.4	-4.6	1.8	
Students - Lower management jobs	-2.2	-5.3	0.8	
Higher management jobs - Lower management jobs	1.0	-0.5	2.6	*
Jobseekers -Lower management jobs	2.6	0.7	4.4	***
Higher management jobs - others	-0.2	-2.0	1.6	
Employees - others	-0.6	-2.7	1.6	
Jobseekers - other	1.3	-0.7	3.4	
Student - Others	-1.4	-3.3	0.5	
Employees - Higher management jobs	-0.4	-1.8	1.0	
Jobseekers - Higher management jobs	1.5	0.2	2.8	**
Jobseekers - Employees	1.9	0.2	3.6	*
Students - Employees	-0.8	-2.3	0.7	
Students - Jobseekers	-2.7	-4.1	-1.3	***

Table 6: Tukey HSD pairwise differences between the socioeconomic status of the learners
p-value $<$ 0.05: * / p-value $<$ 0.01: ** / p-value $<$ 0.001: *

We see that job seekers tend to watch more videos than the classes of the society (Table 6); job seekers consume on average more videos than lower management jobs (difference = 2.6, p-value $<$ 0.001), higher management jobs (difference = 1.5, p-value $<$ 0.01),employees (difference = 2.9,p-value $<$ 0.05) and students (difference = 2.7,p-value $<$ 0.001).

3.3 Identifying factors influencing course completion

Logistic regression is performed with respect to boolean variables, in this case, completing the course or not. The results of logistic regressions are reported as tables of odds ratios. An Odd-Ratio is calculated for a given variable in relation to one of the characteristics of this variable, which represents the reference, noted RC in the Table 7.

Variable	Category	Odds ratio	95% CI
Gender			
	(RC) Male		
	Female	1.1	1.0-1.2
HDI			
	(RC) Low		
	Intermediate	1.2	0.9-1.6
	Very High	**1.5*** **	1.3-1.8
Employment Status			
	(RC)Lower management jobs		
	Higher management jobs	**2.0 *** **	1.5-2.5
	Employees	**1.7 ** **	1.3-2.3
	Students	**1.6 ** **	1.2-2.1
	Jobseekers	**2.2 *** **	1.6-2.9
	Others	**1.8 ** **	1.3-2.5
Age group			
	(RC)0-30		
	31-50	**0.8 * **	0.7-1.0
	51-80	1.0	0.8-1.3

Table 7: Identification of factors influencing course completion of the course with Odd ratios(OR).
p-value $<$ 0.05: * / p-value $<$ 0.01: ** / p-value $<$ 0.001: *

For instance, we study the effect of gender, HDI, employment status and age group on the probability of completing the course(a boolean variable taking 1 for completion and 0 otherwise). For gender, there was no significant evidence that females are more likely to complete the course than males. However, we could observe that more developed countries are 1.5 more likely than least developed countries to complete the course(p-value $<$ 0.001), and as far as the employment status is concerned, higher managements jobs, employees, students, job seekers and others are 2.0 (p-value $<$ 0.001), 1.7(p-value $<$ 0.01), 1.6(p-value $<$ 0.01), 2.2(p-value $<$ 0.001), 1.8(p-value $<$ 0.01) respectively more likely to finish the course than lower management jobs. Falling into the age group of 0-30 suggests that the learners are 1.6 times more likely to reach the end of the course(p-value $<$ 0.05).

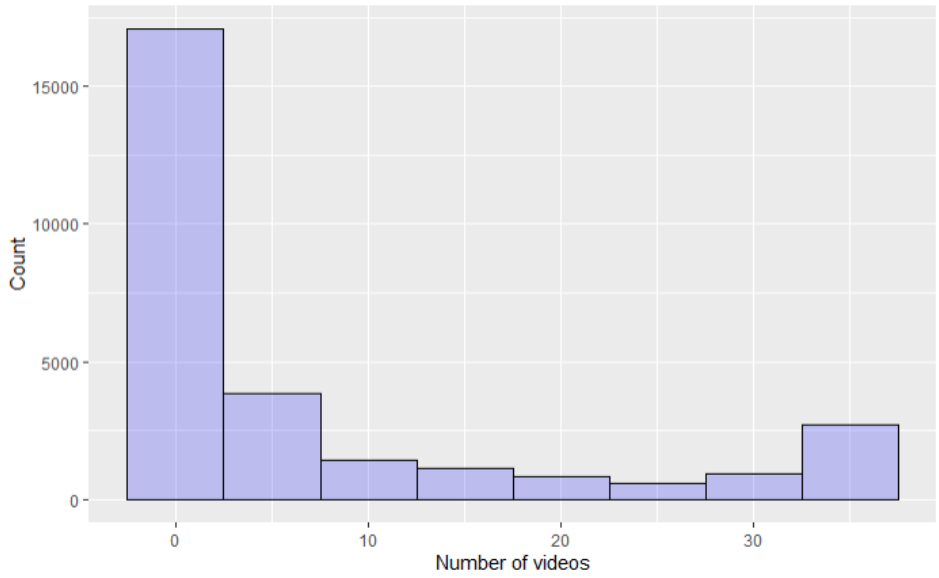

Figure 3: Distribution of the number of videos consumed.

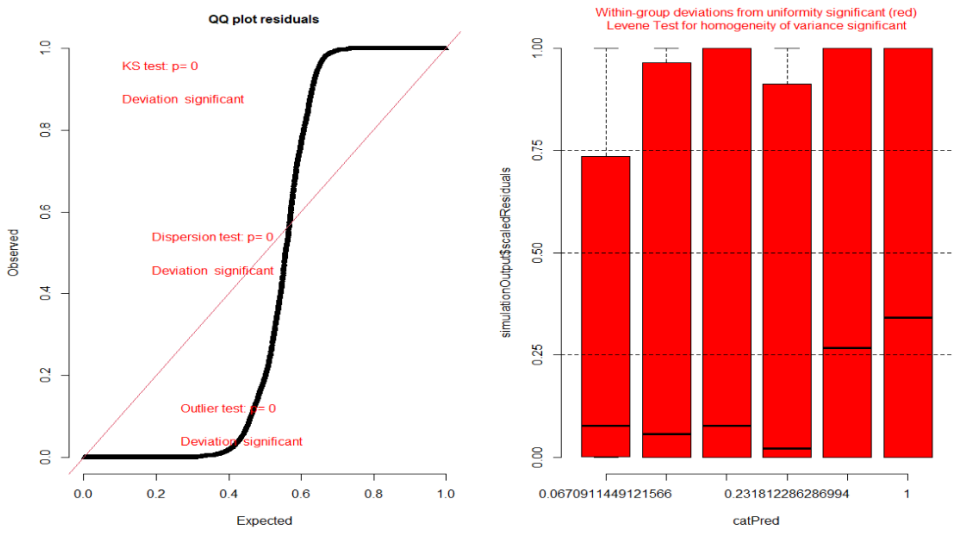

Figure 4: Diagnostic plots to analyze the residuals of the glm model with family equal to "poisson".

Figure 3 refers to the distribution of the number of videos watched. The shape of the distribution is positively skewed, i.e the mean is greater than the median. This positive skewness means that for a smaller number of videos, there are more views and for the higher number of videos, there are fewer views. The plot suggests that the distribution is not normal. We used a Q-Q plot to assess normality and residuals versus fit to check for constant variance of residuals i.e homoscedasticity. From Figure 4, we see that the data is non-normal and is not homoscedastic. The distribution suggests that the number of videos follows a Poisson distribution but with the diagnostic results, the data seems to be overdispersed(Residual deviance » resiual degrees of freedom), this can be adjusted by changing the family to "quasi-Poisson" or "negative binomial".

3.4 Comparison of video consumption behavior between learners with survival analysis

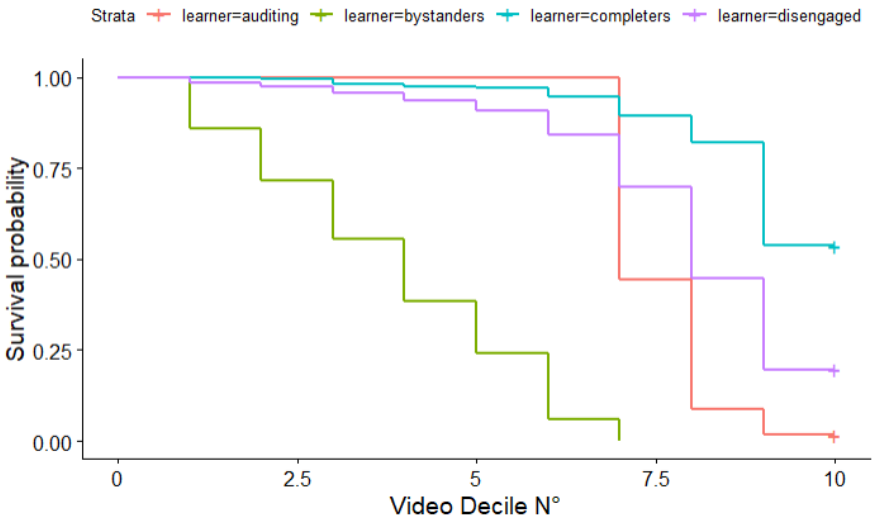

Figure 5: Video consumption behaviours by the different types of learners(auditing,bystanders,completers and disengaged learners).

Regarding the consumption of videos, we observed a clear difference between disengaged and auditing learners (Figure 5). Around the seventh decile, there was a sharp decline for auditors, also where we can see the distinction between auditors and disengaged learners. The difference between disengaged learners and auditors were statistically significant according to the survival analysis carried out(Table 8). We also observe that for bystanders, the first sharp decline occurs between the first and second decile, which is likely to correspond to the first week of the course.

Variable	Category	Hazard Ratio	95% CI
Gender			
	(RC) Male		
	Female	1.01	0.96-1.06
HDI			
	(RC) Low		
	Intermediate	**0.75** ***	0.68-0.83
	Very High	**0.53** ***	0.50-0.57
Learner			
	(RC) Auditing		
	Bystanders	**5.96** ***	5.31-6.68
	Completers	**0.13** ***	0.12-0.15
	Disengaged	**0.32** ***	0.29-0.36

Table 8: Differences in video consumption by gender,HDI and types of learners with hazard ratios.

We conducted log-rank tests,Hazard Ratio(H.R) to identify the factors with the greatest impact on video consumption behaviors within these different categories of learners,gender and HDI. From Table 8, we can see that the learners from more developed

countries tend to disengage twice slower than learners from least developed countries (H.R = 0.53, p-value < 0.001).Amongst the different learners, we can see that auditors consume around 3 times less videos than disengaged learners(H.R = 0.32, ,p-value < 0.001), approximately 10 times less than completers(H.R=0.13,p-value$<$0.001) and slightly more than 6 times than bystanders(H.R=5.96,p-value$<$0.001).

4 Discussion

Since the foundation of Coursera and edX in 2012, Massive open online courses (MOOCs) have been gaining traction at an incredible rate, launching a worldwide debate about its place in educational systems. However, one must be cautious not to rely on the number of enrollees as an indicator of success, since this number is likely to be driven by bystanders. Despite the importance of the number of registrations, the majority of the course activities are led by the highly-engaged learners, usually, the completers, which represent only a minority of the learners. The dropout rate of the learners may also depend on the course structure. For instance, a course with a large number of videos and a little number of activities/quizzes may increase the proportion of highly-engaged auditing learners.

We also found that factors such as socioeconomic status, country of residence and to some extent, the age group had an influence on video consumption and course completion rates. We have seen lower completion rates in the least developed countries, there might be several explanations for this, given the required technological infrastructure(low bandwidth, illiteracy level and language comprehension), however further investigations must be carried to understand the reasons.

As far as socioeconomic status is concerned, higher management jobs, jobseekers, students were relatively more engaged. Jobseekers and students may have similar motives of completing the course, obtaining certificates to increase the value of their resume, and for higher management jobs, advanced certificates may appeal to them to acquire more knowledge in their fields. Similarly to the country of residence phenomenon, further investigations must be performed.

Finally, other factors such as the amount of time that the registrants are willing to give, their learning intentions could prove valuable in understanding the factors leading to course completion. MOOCs have received criticism of low completion rates despite the relatively high number of registrants, studies only on completion rates will not suffice, more comprehensive studies must be designed. To what extent do enrollees benefit from the course despite not completing it? Tackling such issues might give a more holistic approach to the place of MOOCs in the educational system.

5 Annexes

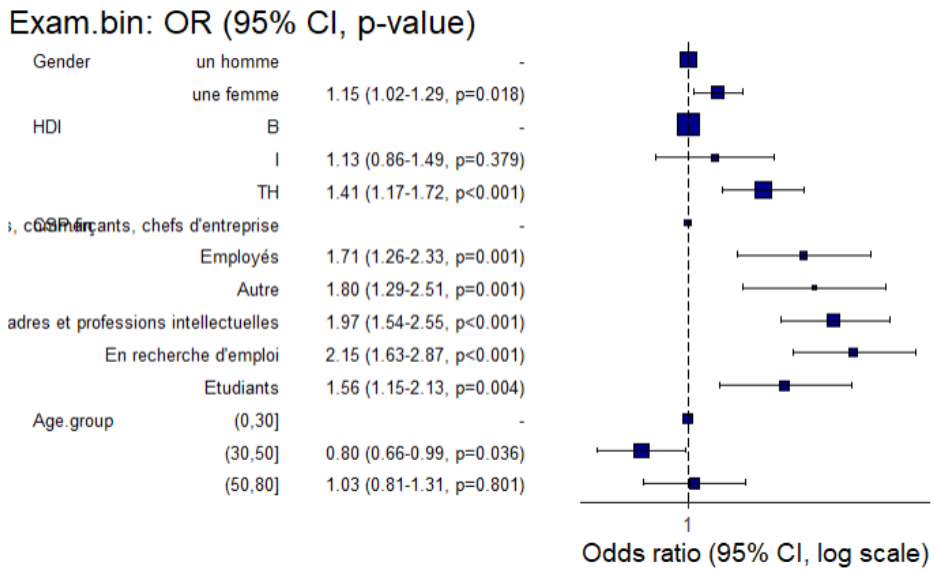

Figure 6: Forest plot with odd-ratios(OR) for gender, HDI, employment status and age group.

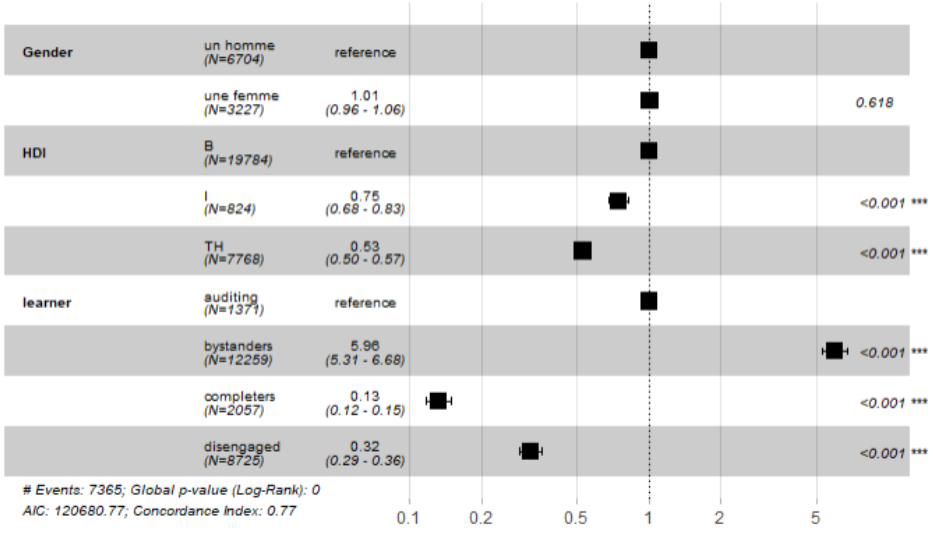

Figure 7: Forest plot with hazard ratios(H.R) for gender, HDI and types of learners.

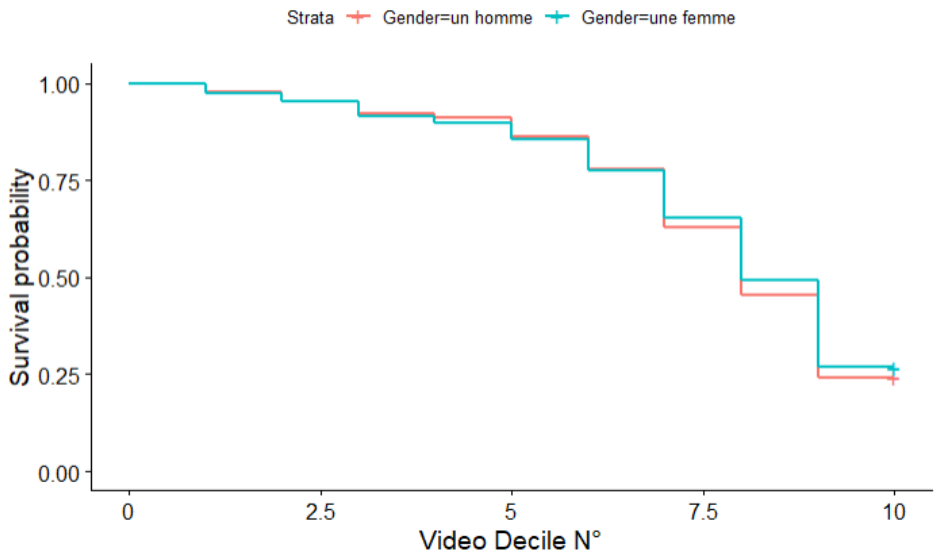

Figure 8: Video consumption behaviour by gender.

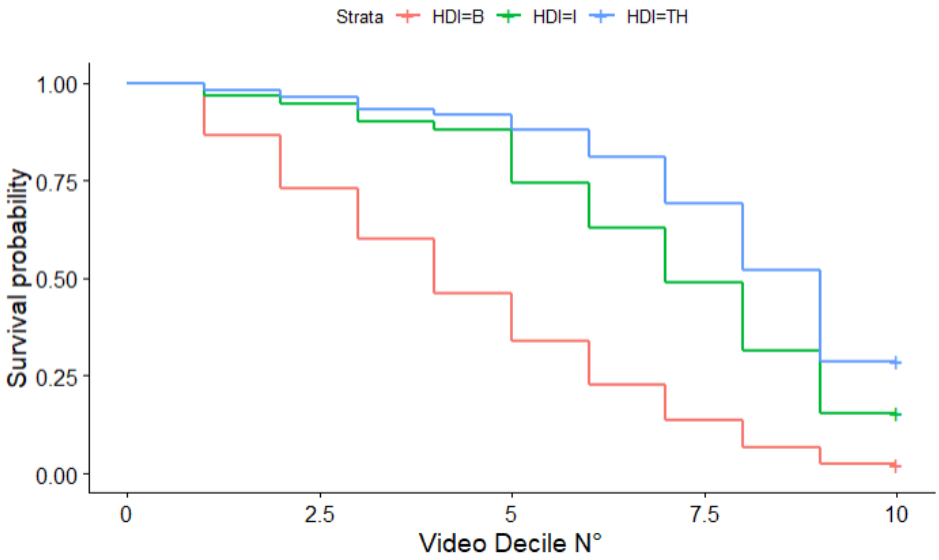

Figure 9: Video consumption behaviour by HDI.

	term	contrast	estimate	conf.low	conf.high	adj.p.value
1	Gender	une femme-un homme	0.89	0.33	1.45	$<2.2^{-16}$
2	HDI	I-B	4.22	2.76	5.68	$<2.2^{-16}$
3	HDI	TH-B	9.01	8.05	9.97	$<2.2^{-16}$
4	HDI	TH-I	4.79	3.59	5.99	$<2.2^{-16}$
5	Age group	(30,50]-(0,30]	0.48	-0.36	1.33	0.4
6	Age group	(50,80]-(0,30]	2.27	1.26	3.28	$<2.2^{-16}$
7	Age group	(50,80]-(30,50]	1.79	1.00	2.58	$<2.2^{-16}$

Table 9: Tukey's HSD (Honest Significant Difference) post hoc test on gender,HDI and age group.

6 References

Cisel, Matthieu, et al. "A Tale of Two Moocs: Analyzing Long-Term Course Dynamics." Centre De Recherche En Gestion (CRG), 14 Nov. 2017, https://tel.archives-ouvertes.fr/X-CRG/hal-01635080v1.

Two-Way (between-Groups) ANOVA in R - University of Sheffield. https://www.sheffield.ac.uk/polopoly_fs/1.536444!/file/MASH_2way_ANOVA_in_R.pdf.

Walker, Copyright 2018 Jeffrey A. "Elements of Statistical Modeling for Experimental Biology." Chapter 16 ANOVA Tables, https://www.middleprofessor.com/files/applied-biostatistics_bookdown/_book/anova-tables.html.

Yufeng. "Adjust for Overdispersion in Poisson Regression." Medium, Towards Data Science, 12 Oct. 2020, https://towardsdatascience.com/adjust-for-overdispersion-in-poisson-regression-4b1f52baa2f1.