edX participants' profile: analysis of the factors that lead to the search for certification

Rodrigo Lins Rodrigues

[1]Universidade Federal Rural de Pernambuco
Recife - PE – Brasil

rodrigo.linsrodrigues@ufrpe.br

Abstract. *Massive Online Open Courses (MOOCs) are freely accessible online courses with open registration. This term was first coined in 2008 when professors of University of Manitoba (Canada) started an online course free and open to anyone. In 2012, two platforms were launched, EdX and Coursera. Until now, these two platforms remain as the most popular MOOCs providers in the world attracting universities from all of the continents. The present study performs data analysis of Harvard and MIT courses available in EdX during the first four years of operation. The objective was to understand students' and courses' profiles and the factors that make certifications more attractive to the participants. This paper could identify some factors that contribute to students' motivation in obtaining formal certification. It was important to see that variables related to engagement impact in the inclination to obtain a certification. Furthermore, demographical characteristics as sex and age are relevant so that institutions can focus on specific targets.*

Resumo. *Massive Online Open Courses (MOOCs) são cursos on-line de acesso livre com registro aberto. Este termo foi criado pela primeira vez em 2008, quando os professores da Universidade de Manitoba (Canadá) iniciaram um curso on-line gratuito e aberto para qualquer um. Em 2012, foram lançadas duas plataformas, EdX e Coursera. Até agora, essas duas plataformas permanecem como os provedores de MOOC mais populares do mundo atraindo universidades de todos os continentes. O presente estudo realiza uma análise de dados dos cursos de Harvard e MIT disponíveis na EdX durante os primeiros quatro anos de operação. O objetivo foi compreender os perfis dos alunos e dos cursos e os fatores que tornam as certificações mais atraentes para os participantes. Este artigo pode identificar alguns fatores que contribuem para a motivação dos alunos na obtenção de certificação formal. Foi importante ver que as variáveis relacionadas ao envolvimento impactassem na inclinação para obter uma certificação. Além disso, as características demográficas como sexo e idade são relevantes, de modo que as instituições podem se concentrar em metas específicas.*

DOI: 10.22456/1679-1916.95708

1. Introduction

Since the beginning of 2010's, there was an exponential growth in MOOCs (Massive Open Online Course) offers. Initiatives from American universities Harvard and MIT (founders of EdX platform) and Stanford (precursor of Coursera platform) have made MOOCs more popular and attracted students from around the world. This main purpose of these platforms is to offer freely accessible online courses, so students attending to these courses do not need to be enrolled in a traditional education institute to obtain knowledge [LIYANA, 2013].

MOOC's structure is designed to allow the participation of an undefined number of students, once these classes are taught online. One thing MOOC's have in common is the relatively low completion rates. This is in part explained (1) due to low entry barriers and (2) the motivation why one enrolls for a MOOC. But this low completion rate does not mean that participants did not learn what they intended to earn. If someone wants to improve its knowledge in a specific subject field, it does not need to complete the entire course and earn a certificate, but only attend to specifics parts of the course. However, if someone wants to prove to others that it has completed the course and acquired the knowledge, it is important to have a valid certification issued by a trusted institution [KOPP, 2017].

These intentions may be summarized as [Reich, 2014]:

- Unsure: not sure about complete course activities;
- Browse: intention to take a look of the materials, without committing to complete course activities;
- Audit: willing to complete part of course activities;
- Complete: Willing to complete activities necessary to earn a certificate.

A survey ran with more than 1.8 million participants of HarvardX and MITx courses showed that, although 54% of students planned to complete the course, only 16% of them earned a certificate. On the other hand, some students change their minds and are convinced to earn a certificate after starting the course. This survey showed that 2% of participants planned to browse the course, but 5% of them got a certificate [KOPP, 2017], [REICH, 2014].

In addition to the free modality, institutions offer a paid professional certificate to students who wish to attest to their successful completion of the course [CHUANG, 2016], [MCAULEY, 2010]. The cost of this certificate varies by course and institution, but in general, are substantially lower than the costs of a traditional course at the same institutions. As the students' participation in these courses generates a great amount of data, there are great possibilities for research in this area.

Due to the online-nature of MOOC institutions, certificates are mostly electronic. Certification types do not differ radically among different providers. The most popular types are PDF-documents or electronic badges that can be shared through social networks, mainly in those networks focused on professional subjects [CALISE, 2017]. Institutions generally have two kinds of certificates [WITTHAUS, 2016]: certificates that confirm that a participant completed course e certificates that verify the identity of participants and evaluate their learning outcomes.

Certifications are a key success factor to MOOCs platforms business models. Free certifications were the first items to be shifted from free to paid. This changing to

DOI: 10.22456/1679-1916.95708

a premium model was observed in all of the most popular MOOCs providers, and it is helping to turn them into profitable institutions [SHAH, 2017] are investing in certified programs to brand themselves as non-traditional learning organizations. These programs offer three main advantages [UBELL, 2017].

- They are a curated selection of a variety of courses offered by a renowned partner University;
- Some of the programs can be used as college credits;
- They have lower prices than traditional courses.

These characteristics make an innovative business model and are key ingredients to make MOOCs platforms generate money [WITTHAUS, 2016], [SHAH, 2017]. The present study performs data analysis of Harvard and MIT courses available in EdX during the first four years of operation. The objective was to understand students' and courses' profiles and the factors that make certifications more attractive to the participants.

2. Definitions of MOOCs

MOOCs, in concept, are freely accessible online courses with open registration. They put together anyone who wishes to learn more about a subject and an expert who facilitates the learning [LIYANA, 2013]. Students attending these courses do not need to be enrolled in a traditional education institute. As all the classes are taught online, the course's structure is designed to the participation of an undefined number of students. Every resource needed for the completion of the course is accessible online, and the platform has its social networking, videos, texts, exercises and other resources [CHUANG, 2016].

Another characteristic of these courses is that the participants are not obligated to attend a MOOC – they choose when and which course they will attend based on their interest and needs [LIYANA, 2013]. MOOCs are designed to include models that interrelate political, epistemological, pedagogical and assessment components, innovatively [RAMIREZ, 2014]. 2.1.

2.1. The first MOOCs

The term MOOC was first coined in 2008 by Siemens and Downes, professors of a University of Manitoba (Canada). They first charged for an online course called "Connectivism and Connective Knowledge," but some months after they decided to open it to anyone for free, without giving the nonpaying students any credit for completing the course.

The paid course had 25 participants, while the nonpaying had 2.300 students registered to participate. The main reasons Siemens and Downes decided to open this course were to have a broader variety of perspectives and ideas about the topics studied and to adapt its structure to a new model of learning. After the good results from this first open course, Siemens and Downes offered additional MOOCs at their university in the following years [CORMIER, 2010].

2.2. Coursera and edX

To create a structure adapted to the needs of the current and the next generations, it is necessary to add a technological infrastructure. In this new model, it is

DOI: 10.22456/1679-1916.95708

important to create international alliances to create a worldwide network of institutions sharing their experience, their knowledge, pooling their resources, equipment, curricula, students and teachers [GARITO, 2016].

Offering MOOCs seems to be attractive even to well-known institutions and to newcomers. Coursera is a platform launched in April 2012 by Stanford University with the cooperation of Princeton, Penn and the University of Michigan. A month later, MIT and Harvard joined forces and launched the platform edX. Until now, these two platforms remain as the most popular MOOCs providers in the world attracting universities from all of the continents [BRAHIMI, 2015].

2.3. How do they make money

One of the main issues regarding MOOCs is that they lose money by providing free courses to anyone. In the first moment, universities and venture capitalists subsidize the platforms. However, investors expect returns on these investments, that will only happen if platforms can make revenues that surplus their costs [BRAHIMI, 2015], [BELLEFLAMME, 2016].

MOOC platforms have large fixed costs, like investment in the development of the platform itself and in acquiring large online storage and bandwidth. Moreover, the creation of audiovisual materials, mainly videos, and e-books, demand the use of professionals that universities usually don't need, like a cartoonist or a video editor [BELLEFLAMME, 2016].

However, these platforms have almost zero variable costs. Unlike traditional courses, there is no need to expand their infrastructure to attract more students, and they also can use the same content on various course offerings [BELLEFLAMME, 2016].

As many other disruptive initiatives in other markets, MOOCs platforms are trying to find a new business model. Usually, they offer free services and information to attract users and their activity to after charging for compliments. So, in addition to the free modality, institutions offer a paid professional certificate to students who wish to earn a certificate after completing the course. A certification signals to others that a student has mastered course content. The cost of this certificate varies by course and institution, but in general, the prices of an online certification are substantially lower than the costs of a traditional course at the same institutions [DELLAROCAS, 2013].

3. Methodology

The methodology used in this paper was based on CRISP/DM, an acronym for Cross Industry Standard Process for Data Minning. It is a six-phase framework to carry out data mining projects in any industry, independent of which technology will be used. It was conceived to standardize DM process in real-world situations and to help organizations and individuals to take better decisions [DELLAROCAS, 2013], [MORO, 2011].

In the business and data understanding the main goal is to understand the project objectives and define a data mining problem [WIRTH, 2000]. To this paper, it was studied the business model and information about EdX at its website and in other secondary sources.

The second step in CRISP/DM framework is the data understanding. Using the data available, it was carried out a descriptive analysis to understand participants'

DOI: 10.22456/1679-1916.95708

profile, the courses offered by each institution, the number of participants and certificates issued and other analyses. These findings will be detailed shown in section 4. The third step is data preparation phase. In this step, it is necessary to define, inspect and technically analyze the data which will be used. To this project, it was necessary to categorize the courses according to its percentage of certified participants. To this, it was calculated the mean average certification rate for all the courses that honors certificate (5.82% of participants). After that, were classified the courses that had a certification rate above the average. The final step was creating three dummy variables to classify to which subject of the courses.

In the modeling phase, the most suitable data mining techniques are selected and applied to achieve the objective defined in the first step. The outcome is a model that uses the knowledge learned in the previous phases, process the data and give an output [MORO, 2011]. In this paper, it was used the logistic regression technique to identify which factors can help to classify if a specific course will have a certification rate above the average. The fifth phase is the evaluation of model's outcome. The results were evaluated based on the accuracy of the confusion's matrix and the area under the ROC curve.

4. Data Analysis

The present study performs data analysis of Harvard and MIT courses available in EdX during the first four years of operation. The objective was to understand students and courses' profiles and the factors that make the certifications more attractive to the participants.

4.1. Courses offered

In the analyzed period, HarvardX and MITx offered 290 courses divided into four subjects: Computer Science (CS); Government, Health, and Social Science (GHSS); Humanities, History, Design, Religion, and Education (HHDRE); Science, Technology, Engineering, and Mathematics (STEM). It was observed that HarvardX offered most of HHDRE courses, while MITx offered most of CS and STEM courses. HarvardX accounted for 129 courses, while MITx offered 161 courses.

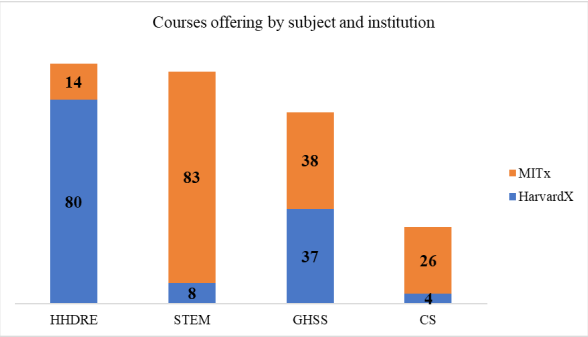

Figure 1. Courses offering by subject and institutions

4.2. Participants profile

In the first four years of EdX, 4.5 million students attended HarvardX and MITx courses. 71.66% of the participants are men. 73% of them hold at least a Bachelor's degree, and their median age is 29 years. It was found that male students focus more on

DOI: 10.22456/1679-1916.95708

CS and STEM courses, while women focus on HHRDE and GHSS courses. There was a notable difference between the number of male and female participants in Computer Science courses.

4.3. Most attractive courses

The top 10 courses, in a number of participants, are mainly from CS subject. The first four courses are related to introduction to computer science. Chuang & Ho noted that CS courses receive several participants that were previously enrolled in another subject courses. At the same time, several participants of CS courses enroll in other subject courses [CHUANG, 2016].

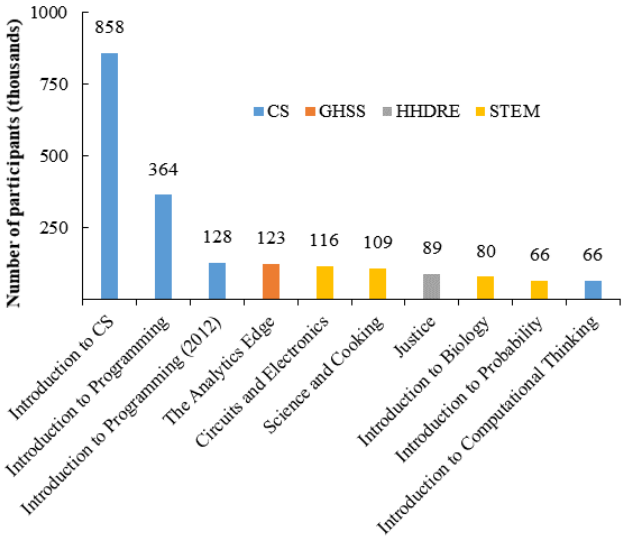

Figure 2. Most popular courses per subject – in number of participants

On the other hand, when we analyze the ten courses with higher average hours spent at the course per participant, only one of the most popular courses appears on this list. Although CS courses attract a huge number of participants, most of them do not spend much time studying it.

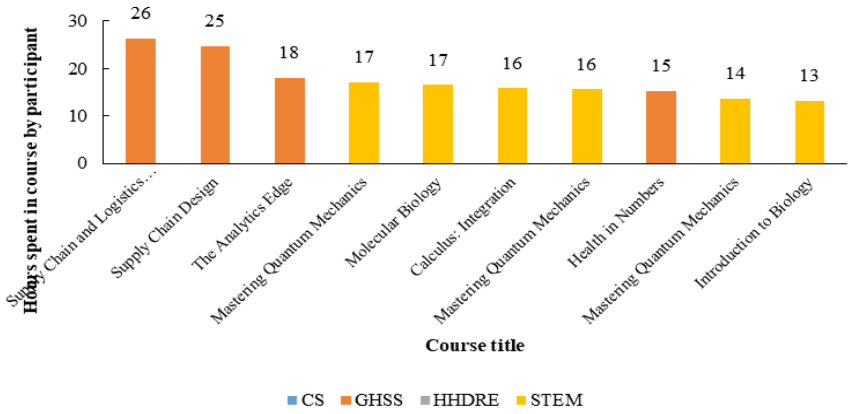

Figure 3. Courses with more engagement per subject – in hours spent in course by the participant.

DOI: 10.22456/1679-1916.95708

The difference between popular courses versus high retention courses may be explained by the students' intentions and demographics. CS and STEM courses students are younger and fewer bachelors than HHRDE and GHSS courses.

4.4. Certificates issued

Around 245.000 certificates were issued, representing 5.5% of the total of participants. It is important to note that some courses do not offer paid certificates. Moreover, the number of paying students is higher than the number of certificates issued – this is because a participant has to complete a final project and get a minimum grade to receive a certificate. Analyzing only the 236 courses that offer certifications, it is found that the median certification rate is 5.82%. This rate increases to 60% when are analyzed only the participants that paid for a certification [CHUANG, 2016].

It was observed that the most popular courses do not have a high proportion of certificates. The correlation between the number of participants and the percentage of certificates issued was -0.2586. Also, were identified positive correlations classified as moderate (from 0.40 to 0.60) between the proportion of certified students and variables related to student participation, the median age of the class and proportion of female students.

Table 1. Correlation between the proportion of certified participants and selected variables.

Variable	Correlation with the proportion of certified participants
The proportion of grade higher than zero	0,6155
Median age	0,5914
The proportion of audited participants	0,566
The proportion of participants who posted in forums	0,5059
Proportion of female participants	0,4569
The proportion of Bachelor's Degree or Higher	0,4123
The proportion of participants who played video	0,1856
Hours per participant	-0,0323
Audited (> 50% Course Content Accessed)	-0,1236
Total course hours	-0,1613
Participants (Course Content Accessed)	-0,2586
Proportion of male participants	-0,4569

Analyzing the 20 courses with the highest proportion of certified students, it was found that 19 are from HHDRE subject. To check if there was a statistical difference between certification rates from courses of different subjects, an ANOVA test was conducted to compare the effect of subject courses on the certification rates.

Table 2. P values of an ANOVA test between course subject and proportion of certified participants.

Variable	*Df*	*F*	*p value (α = 0)*
Course subject	3	31.41	<2e-16*

An analysis of variance showed that the effect of course subject on the proportion of certified participants was significant, F=31.41, p =< 2e-16. The results showed that the proportion of certified students varies according to the area of the course.

DOI: 10.22456/1679-1916.95708

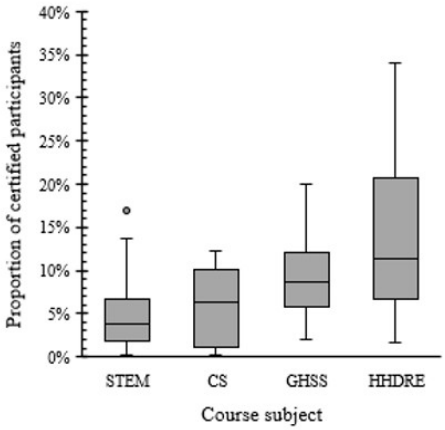

Figure 4. The proportion of certified participants per course subject.

Figure 4 shows the difference between the median percentage of certified participants between different subjects. It is possible to see that HHDRE courses have a proportion of certified participants more than two times than STEM courses.

These results are similar to the conclusions of other researches. Some studies have shown that participant behavior has wide variation across curricular areas [CHUANG, 2016].

5. Predictive Analysis

To understand and predict which factors lead to the search for a certification, it was built as a binary logistic regression model. First of all, it was calculated the mean average certification rate for all the courses that honors certificate (5.82% of participants). The following step was creating three dummy variables to classify to which subject the course belongs to. The final step was classifying the courses that had a certification rate above the average. It was found that 61% of the courses had a certification rate above the average.

To build the model, data was split into two parts: 70% of data was used to training and 30% to testing. After several calculations and results evaluations, it was found that the best logistic regression model used the following variables: proportion_audited, proportion_played_video, proportion_female, median_age, hours_per_participant and course subject (in this case, it was used the dummy variables).

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -7.64189    2.61566  -2.922  0.00348 **
proportion_audited    14.79058    2.46707   5.995 2.03e-09 ***
proportion_played_video  2.64600    2.06303   1.283  0.19964
proportion_female      6.19246    2.01813   3.068  0.00215 **
median_age             0.03756    0.08630   0.435  0.66341
hours_per_participant  0.27886    0.06928   4.025 5.69e-05 ***
d1                    -1.60284    0.81485  -1.967  0.04918 *
d2                    -1.58125    0.88115  -1.795  0.07273 .
d3                    -2.17086    0.71618  -3.031  0.00244 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5. The significance of the variables used in the model.

The figure above shows the significance of variables. It is possible to see that proportion_audited and hours_per_participant play an important role to predict if the

DOI: 10.22456/1679-1916.95708

course would have an above the average certification rate. The variable proportion_audited represents the percentage of participants in a course which audited more than 50% of the course, while hours_per_participant represents the average of the hours a participant spent in the course. The significance of the variables may be explained due to the student's necessity to spend more time watching the videos and completing the assignments to be able to receive a certificate.

The accuracy in predicting which courses had a proportion of certified participants higher than the average was 0.8475. The sensitivity statistic represents the ability of a model to correctly classify if a course has an above the average proportion of certified students [PARIKH, 2008] and is calculated as the proportion of true positives that are correctly predicted by the model [ALTMAN, 1994]. This means that among all of the courses above the average, the model could predict 88,89% of them. The Pos Pred Value (PPV) statistic is the proportion of the total of elements classified as true by the model and in fact had this classification [PARIKH, 2008]. The model obtained a PPV of 0.8469, which means that 84,69% of the courses predicted as above the average had more participants certified than the average.

A receiver operating characteristics (ROC) graph is a widely known technique for visualizing, organizing and selecting binary classifiers based on their performance. To quantify the performance of a classifier, it is measured by the area under the ROC curve (AUC). It is calculated as the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [FAWCETT, 2006].

The area under the ROC curve (AUC) is a metric that specifies the probability that when it is drawn one positive and one negative example at random, the model assigns a higher value to the positive than to the negative example. This value may range between 0.5 and 1, and the higher the result is, it is better [FAWCETT, 2006]. As the AUC obtained in this work was 0.9213, this model represents a good classifier.

6. Conclusion

In recent years, MOOCs have turned from experiments of single universities to a consolidated new way to spread formal education around the world. Due to the challenges of monetizing this new business model, it is important to take advantage of the data generated from participants' interaction. Researches from MOOCs allow developing courses more attractive to students and more profitable to educational institutions.

This paper could identify some factors that contribute to students' motivation in obtaining formal certification. It was important to see that variables related to engagement, like the proportion of participants who completed more than 50% of the course, the proportion of participants who played videos and the average hour per participant, impact in the inclination to obtain a certification. Furthermore, demographical characteristics as sex and age are relevant so that institutions can focus on specific targets. As suggestions to further studies, this same analysis could be applied to MOOCs from other institutions at EdX and to other platforms, like Coursera and Udacity.

DOI: 10.22456/1679-1916.95708

References

Arnold-Garza, S.: 'The Flipped Classroom Teaching Model and Its Use for Information Literacy Instruction'; Communications in Information Literacy, Vol. 8, No. 1 (2014), pp. 7–22.

LIYANA, GUNAWARDENA, Tharindu Rekha; ADAMS, Andrew Alexandar; WILLIAMS, Shirley Ann. MOOCs: A systematic study of the published literature 2008-2012. The International Review of Research in Open and Distributed Learning, v. 14, n. 3, p. 202-227, 2013.

KOPP, Michael; EBNER, Martin. La certificación de Los MOOC. Ventajas, desafíos y experiences practices. 2017

REICH, Justin. MOOC completion and retention in the context of student intent. EDUCAUSE Review Online, 2014.

CHUANG, Isaac; HO, Andrew Dean. HarvardX and MITx: Four Years of Open Online Courses--Fall 2012-Summer 2016. 2016.

MCAULEY, Alexander et al. The MOOC model for digital practice. 2010.

CALISE, Mauro; REDA, Valentina. In and Out. Federica experience in the rugged terrain of MOOCs inclusion in institutional strategies of university education. 2017

WITTHAUS, Gabi R. et al. Validation of non-formal MOOC-based learning: an analysis of assessment and recognition practices in Europe (OpenCred). 2016.

SHAH, Dhawal. Massive Open Online Courses used to be 100% free. But they didn't stay that way. 2017. <https://medium.freecodecamp.org/massive-open-online-courses-started-out-completely-free-but-where-are-they-now-1dd1020f59>. Accessed on Aug. 18th 2017.

UBELL, Robert. MOOCs Find Their Sweet Spot. 2017. <https://www.insidehighered.com/digital-learning/views/2017/04/12/three-steps-making-moocs-money-makers>. Accessed on Aug. 18th 2017.

RAMÍREZ, M. S. Guidelines and success factors identified in the first MOOC in Latin America. In: Edulearn14, 6 th International Conference on Education and New Learning Technologies, Barcelona. 2014.

CORMIER, D.; SIEMENS, G. Through the open door: Open courses as research, learning, and engagement. Educause, 45 (4), 30-39. 2010.

GARITO, Maria Amata. Alliances for Knowledge: MOOCs to Create New Professional Skills in a New Model of University (Positive and Negative Aspects). International Journal of Advanced Corporate Learning, v. 9, n. 1, 2016.

BRAHIMI, Tayeb; SARIRETE, Akila. Learning outside the classroom through MOOCs. Computers in Human Behavior, v. 51, p. 604-609, 2015.

DOI: 10.22456/1679-1916.95708