Master Thesis (2021)

The retention problem in *MOOC*:

Why do learners not complete their courses?

Graduate School of Applied Informatics

University of Hyogo

High Confidence Informatics Track, Applied Informatics Domain

IM19I501

Wang Lei

（Date of submission: 02/22/2021）

（Advisor: Danny Fernandes, Professor）

Abstract

The purpose of this research is to explore the patterns of retention in *MOOC* (Massive Open Online Course) and identify features affecting retention. This study is based on statistical analysis and machine learning on two publicly available datasets from edX. Through the data analysis we found that certification rate of all the registrants in *MOOC* is misleading. Instead, we need to explore retention in *MOOC* in the context of learners' intention. For those who intended to complete a course, the certification rate ranges from 13% to 100%, with a mean of 32% and standard deviation of 19.67%. The features affecting certification rate include course duration, number of chapters, country, education, gender, region, total number of events, active days, number of videos played, proportion of chapters accessed, proportion of user's lifetime in course duration. Among these features, active days and proportion of chapters accessed have a strong correlation with certification.

We identified those intended to complete a course based on our assumption regarding their learning activities. This may be not very accurate. Newer courses are introducing the pre-course surveys to identify learners' intention. The learners' reported intention can be combined with their activities data. This will help better understand the learners' intention to complete.

The findings of this research enable the *MOOC* instructors and platforms to narrow down the targets for intervention. Thus, resources can be used more efficiently. The features identified could serve as metrics for intervention.

Keywords: *MOOC*, Massive Open Online Course, retention, intention to complete

Table of Contents

Abstract 2

Glossary and acronyms 4

List of figures 6

List of tables 8

Section 1 Introduction 1

1.1 The problem 1

1.2 Research objective and research questions 1

1.3 Significance/Importance of the study 2

1.4 Scope and organization of thesis report 3

Section 2 A background on *MOOC* 4

2.1 The advent of *MOOC* 4

2.2 Trend in interest in *MOOC* 8

2.3 *MOOC* in the context of Gartner's Hype Cycle 10

2.3.1 The Gartner's Hype Cycle 10

2.3.2 Gartner's Hype Cycle for *MOOC* 12

Section 3 Related research 13

3.1 Definition of retention 13

3.2 Research on retention in *MOOC* 14

Section 4 Datasets 16

4.1 Original datasets 16

4.2 Working datasets 17

4.2.1 Year 1 dataset data wrangling and feature engineering 17

4.2.2 Year 4 dataset data wrangling 20

Section 5 Descriptive and inferential statistical analysis 24

5.1 Introduction 24

5.2 Preliminary analysis 25

5.2.1 Basic information 25

5.2.2 Demographics 26

5.3 Certification rate 29

5.4 Better understanding of certification rate 31

5.5 Operationalize intention to complete 31

5.6 Certification rate of intended to complete 33

5.7 Features affecting retention 35

5.7.1 Dependent features and independent features 35

5.7.2 Course related features 36

5.7.3 Demographic features 41

5.7.4 Learners' activity features 46

5.8 Exploration of year 4 dataset 48

5.8.1 Proxy of intended to complete. 48

5.8.2 Features related to certification 51

Section 6 *MOOC* retention prediction with machine learning 56

6.1 Dataset and features selection 56

6.2 Preprocessing 56

6.3 Train/ test split 57

6.4 Model selection 57

6.4.1 Logistic Regression 58

6.4.2 Random Forest. 59

6.4.3 Support Vector Machine 59

6.5 Baseline 60

6.6 Initial model evaluation 61

6.7 Optimized model 63

Section 7 Discussion and conclusions 64

7.1 Conclusions 64

7.2 Limitations and suggestions for additional research 66

Acknowledgements 68

Appendix 1

I. Python code for year 1 dataset analysis 1

II. Python code for year 4 dataset analysis 1

Glossary and acronyms

***Audited*:** Those who have accessed more than 50% of the course content at least once over the course duration. This is used interchangeably with *explored*.

Chapter: The highest organizational unit in the courseware.

***Explored*:** Those who accessed half or more of the chapters in the course.

Grade: The weighted average score from quizzes, assignments and tests.

Intended to complete: Those with intention to complete a course.

Intention to complete: We defined this based on three conditions: (a) The learner's lifetime is more than 13% of the course duration. (b) The learner interacted with at least 13% of the course chapters. (c) The learner's grade is greater than zero.

Lifetime: The duration from a learner's first login to last logout.

Participants: Those who have accessed the course content at least once over the duration of the course. This is used interchangeably with *viewed*.

***Registered*:** Those who registered for a *MOOC* course.

Retention: The fraction who enroll successfully finish a course to the standards specified by the instructor. This is also referred to as retention rate, completion rate and certification rate in this paper.

Subpopulation: Refers to different group of learners in this paper. For example, the *registered*, the *viewed*, the *explored*, the intended to complete.

***Viewed*:** Those who accessed the courseware tab.

***CS*:** Course subject group including computer science

***GHSS*:** Course subject group including government, health, and social science

***HHDRE*:** Course subject group including humanities, history, design, religion, and education

***MOOC*:** Massive open online course

***STEM*:** Course subject group including science, technology, engineering, and mathematics

***SVM*:** Support Vector Machine

List of figures

Figure 1: *MOOC* evolution (Source: Phil Hill, 2012) 7

Figure 2: Google Trends for *MOOC*/*MOOC*/Massive Open Online Course 9

Figure 3: Google Trends for *MOOC* 9

Figure 4: Hype Cycle (Source: Gartner research, May 2003) 11

Figure 5: Number of registrants by course ID 26

Figure 6: Proportion of registrants by country 27

Figure 7: Proportion of registrants by level of education 27

Figure 8: Distribution of age 28

Figure 9: Proportion of registrants by gender 28

Figure 10: Certification rate of intended to complete by course 35

Figure 11: Certification rate by institution 37

Figure 12: Certification rate by semester 39

Figure 13: Course duration vs. certification rate 39

Figure 14: Number of chapters vs. certification rate 40

Figure 15: Course subject vs. certification rate 41

Figure 16: Certification rate by country 42

Figure 17: Certification rate by level of education 43

Figure 18: Age vs. certified or not 44

Figure 19: Gender vs. certification rate 44

Figure 20: Region vs. certification rate 45

Figure 21: Proportion of chapters viewed vs. grade 48

Figure 22: Certification rate for participants 50

Figure 23: Certification rate for *audited* 50

Figure 24: Institution vs. certification 52

Figure 25: Course subject vs. certification 53

Figure 26: Year vs. certification 53

Figure 27: Percent grade > 0 vs. certification 55

Figure 28: Performance metrics for three supervised learning models 61

Figure 29: Parameters of the best classifier 63

Figure 30: Feature importance 63

List of tables

Table 1: Subpopulations 2

Table 2: List of features 3

Table 3: Year 1 dataset 18

Table 4: Year 4 dataset 19

Table 5: Cleaned year 1 dataset 21

Table 6: New features in year 1 dataset 22

Table 7: Cleaned year 4 dataset 23

Table 8: Number of registrants by course 25

Table 9: Overall certification rate of different subpopulations 29

Table 10: Certification rate of different subpopulations by course 30

Table 11: Overall certification rate 34

Table 12: Certification rate of intended to complete by course 34

Table 13 : Statistics of HarvardX and MITx courses certification rates 37

Table 14: Features selected 57

Table 15: Certification rate of different subpopulations 64

Section 1 Introduction

1.1 The problem

MOOC (Massive Open Online Course) used to attract a great deal of attention after its advent in 2012 (Daniel 2012; Gaebel 2014). It was expected to provide access of high quality, college level courses to anyone, anywhere, any time. Some researchers claimed that *MOOC* even had the potential to reform higher education (Khalil and Ebner 2014).

However, *MOOC* was not a salvation for the problems of higher education. *MOOC* had its own problems, too. The retention problem was one of those problems. Although so many learners enrolled in courses, only a very small proportion continued learning and completed these courses (Breslow et al. 2013; Belanger, Thornton, and Barr 2013; Jordan 2014, 2015).

Many researchers explored features affecting the retention in *MOOC* (Creelman 2013; Jordan 2014, 2015). Some researchers provided some possible strategies to improve the retention rate (Khalil and Ebner 2014). However, most of the researchers did not consider the learner's intention. In (Reich 2014) the author analyzed data of those who reported to intend to complete. However, the intention to complete was captured only in the pre-course survey. Some students may soon change their mind. In order to get a better understanding of this retention problem in *MOOC*, we also need to consider the students' actual learning activities.

1.2 Research objective and research questions

Our research objective in this study is to explore the patterns of retention in *MOOC* and identify features affecting retention in *MOOC*. To achieve this, we addressed the

following research questions: (a) What are the retention rates of different subpopulations in *MOOC*? Subpopulations refer to Table 1. (b) Do course related features affect retention in *MOOC*? Course related features refer to Table 2. (c) Do demographic features affect retention in *MOOC*? Demographic features refer to Table 2. (d) Do learners' activity related features affect retention in *MOOC*? Learners' activity related features refer to Table 2.

1.3 Significance/Importance of the study

This study explored retention in *MOOC* based on learners' intention to complete. Besides identifying features affecting retention with statistical analysis, we also built machine learning models for confirmation. Instead of using one dataset, this study analyzed two datasets. The second dataset serves as a longitudinal reference to the findings in the first dataset.

The findings of this research will enable the *MOOC* instructors and platforms to narrow down the targets for intervention. Thus, resources can thus be used more efficiently. The features identified could serve as metrics to guide intervention.

Table 1: Subpopulations

Subpopulation	Definition
Registered	Those who enrolled in the course
Viewed	Those who have accessed the course content at least once over the course duration
Explored	Those who have accessed more than 50% of the course content at least once over the course duration
Intended to complete	Those who have the intention to complete

Table 2: List of features

Feature category	Feature
Course related feature	Institution, semester, course duration, number of course chapters, course subject
Demographic feature	Country, level of education, age, gender, region
Learners' activity related feature	Total number of events, active days, number of videos played, proportion of chapters accessed, number of forum posts, enrollment days relative to launch, proportion of user's lifetime in course duration

1.4 Scope and organization of thesis report

The scope of this research is the retention problem in *MOOC*. It is not the retention problem in traditional education. It is not the retention problem in the online education other than *MOOC*.

This report has seven main sections. Section 1 is introduction. We briefly introduce the problem, research objective, research problems and importance of the research. Section 2 provides a background for *MOOC*. We first review the evolvement of *MOOC*. Then we explore *MOOC* in the context of Google Trends and Gartner's Hype Cycle. In Section 3 literature review, we review previous research on retention in *MOOC*. Section 4 describes our datasets. We introduce the original dataset and the working dataset. In section 5 data analysis we conduct descriptive statistics and inferential statistics. In section 6 machine learning modeling, we predict retention with the identified features. In section 7 the discussion and conclusions section, we discuss the findings and suggest additional research.

Section 2 A background on *MOOC*

In last section, we introduced the retention problem in *MOOC*. We then introduced our research objective and research questions as well as the importance of this study. To better understand the phenomenon of *MOOC*, we will go over the advent of *MOOC* as well as the evolvement of *MOOC* in the context of Google Trends and Gartner's Hype Cycle Model.

2.1 The advent of *MOOC*

MOOC, is an acronym which stands for massive open online course. George Siemens and David Cormier coined it to describe a course developed by Stephen Downes and George Siemens (Downes 2008). The name of the course was Connectivism and Connective Knowledge. It was considered to be the first *MOOC* (Liyanagunawardena, Adams, and Williams 2013). The course had 25 paid for-credit enrolments on campus as well as around 2200 free non-credit students online. It was special due to its large size, its openness, and the for-credit status (Downes 2008).

Although the first *MOOC* course was a milestone in the history of *MOOC*, the mainstream *MOOC* today is considered to be of another type. Based on the distinct pedagogical approach, *MOOC* was categorized into two types. One type is considered to be decentralized and more learner-centered, with a connectivist pedagogy. This type is called *cMOOC.* The other type is considered to be more teacher-centered and more traditional college styled, with a cognitive-behaviourist pedagogy (Rodriguez 2012). To distinguish from *cMOOC*, the other type of *MOOC* is called *xMOOC* (Daniel 2012).

The first x*MOOC* that really caught the attention of the public was *Introduction to Artificial Intelligence*. It was a graduate level course, taught by two Stanford professors Sebastian Thrun and Peter Norvig. The course was launched in year 2011 and finally attracted more than 160,000 enrolments from 209 countries. Even the professors themselves were thrilled to have such a large number of audiences. At the end of the course, 20,000 students completed the course and got statements to acknowledge their work. While the course was offered online, there were still 200 Stanford students taking the course on campus. What is impressive is that at the end of the on-campus course, only 30 students still came to the classroom. Many Stanford students preferred taking the course online instead (Norvig 2012).

It was not the first time the university put its courses online. Online courseware, like MIT OpenCourseWare, has a long history. Why did this *Introduction to Artificial Intelligence* course become so popular? One reason could be the way the course was designed.

As Peter Norvig put it in a TED talk (Norvig 2012), inspired by Benjamin Bloom's one-on-one tutoring idea (Bloom 1984), they explained the new point by writing on a piece of paper while talking. There was a camera taping from above. It felt like a smart friend was explaining that point right next to you. Inspired by Khan Academy, they chunked lectures into 2-6 minutes short videos. In the short videos, they embedded some quizzes. Learners needed to complete them before they could move forward with the video. There was even a deadline for the assignments, since as Peter Norvig argued in the same TED talk, "You can watch them any time you want. But if you can do it any time, that means you can do it tomorrow, and if you can do it tomorrow, you may not

ever get around to it."

Introduction to Artificial Intelligence was not the only course that attracted huge attention. The *Machine Learning* course was another one. It was offered by another Stanford professor Andrew Ng. This course ended up with 104,000 registered learners and 13,000 completed it. Although the design of these x*MOOC*s may not be exactly the same, elements like short videos, embedded quizzes, auto-graded or peer reviewed assignments are pretty common.

The experience mentioned above aspired the professors to build platforms to deliver more courses. In year 2011, Sebastian Thrun took the first step and founded Udacity, a for-profit company. The name was mashed up from two words: audacity and university. The earlier courses on Udacity mainly focused on computer science. Now their programs cover artificial intelligence, data science, business and so on. They call their program Nanodegree, which are packed with courses on various skills that you will need to land a job. Right after Udacity, Andrew Ng and his colleague Daphne Koller founded Coursera, another for-profit company. Afraid of legging behind, in year 2012, MIT and Harvard founded edX, a non-profit company. So, the big three *MOOC* players in the US all emerged. Year 2012 was also called the year of *MOOC* by New York Times (Pappano 2012). Figure 1 shows this period of history of *MOOC* evolvement. From Figure 1 we can see how the two types of *MOOC* evolved and how the main *MOOC* platforms emerged.

As the interest in *MOOC* was increasing, many people started to imagine the big picture of *MOOC* as a disruptive innovation to higher education. For a long time, the higher education in US has been criticized for class overcrowding and being costly.

Thus, *MOOC* seemed to have a great potential here. Most of these *MOOC*

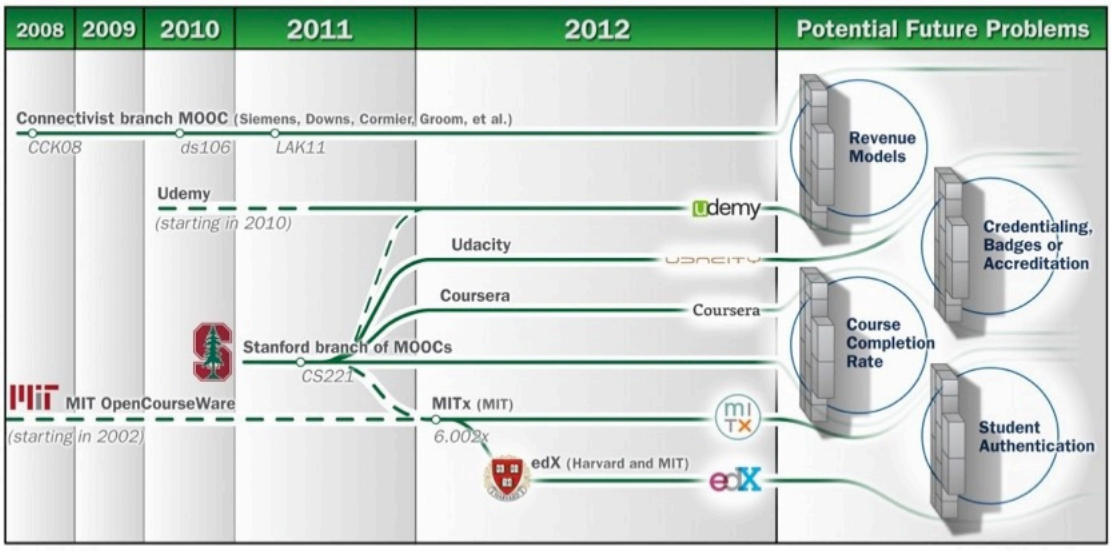

Figure 1: *MOOC* evolution (Source: Phil Hill, 2012)

courses are free to audit. If you would like to try all the assignments and get a certificate, you will need to pay some small amount of money, usually between 49 and 99 dollars. The once unavailable high-quality classes from the elite universities suddenly became available. As long as you have a high-speed broadband, a connected device, you are ready to take advantage of *MOOC*. *MOOC* was considered to distribute the education to the whole world and democratize education (Zawacki-Richter et al. 2018).

Although *MOOC* seemed to have so many potential benefits, many concerns regarding *MOOC* also arose at the same time. Although *MOOC* is considered to democratize education, research found that the majority of the *MOOC* takers already had a

Bachelor's degree or higher. *MOOC* was educating the well-educated instead of those who had no access to college (Emanuel 2013). The cost of taking a *MOOC* is pretty low, but if the motivation of students were taking a course for credit or for a job, maybe a certification of completion did not mean anything to them. Another big problem was the retention rate. Although so many students enrolled in a course, only a very small proportion ended up completing the course. As these kinds of problems persisted, the interest in *MOOC* decreased gradually. Below, we will explore the trend of *MOOC* interest using Google Trends analysis.

2.2 Trend in interest in *MOOC*

We used Google Trends to explore the trend of interest in *MOOC*. At first, we searched the following different key words: *MOOC*, *MOOCs*, Massive Open Online Course. We selected the US as the country, to first understand of which key words attracted more interest. The time period was set up from 2007/01/01 to 2020/10/01. We chose year 2007 instead of year 2008 as a start point although the word *MOOC* was first coined in 2008. We supposed that there may be possibly a few searches a little before year 2008 which would be a noise to our search. For the category and search dropdown menu, we selected All categories and Web Search. The result we got is shown in Figure 2. From Figure 2 we can see that, the key word *MOOC* has much more interest than the other 2 terms. *MOOC* seems to be a good term to explore further. For the term *MOOC*, we can see that the interest did not start increasing until early 2012. After a sharp increase, the peak appeared at around September, 2013. Then the interest started to drop down gradually. The final relative interest was less than 25% of the peak.

Then we explored the Google Trends in other countries. Besides US, the countries we

explored were China, Japan, Australia as well as Worldwide. The reason why we chose

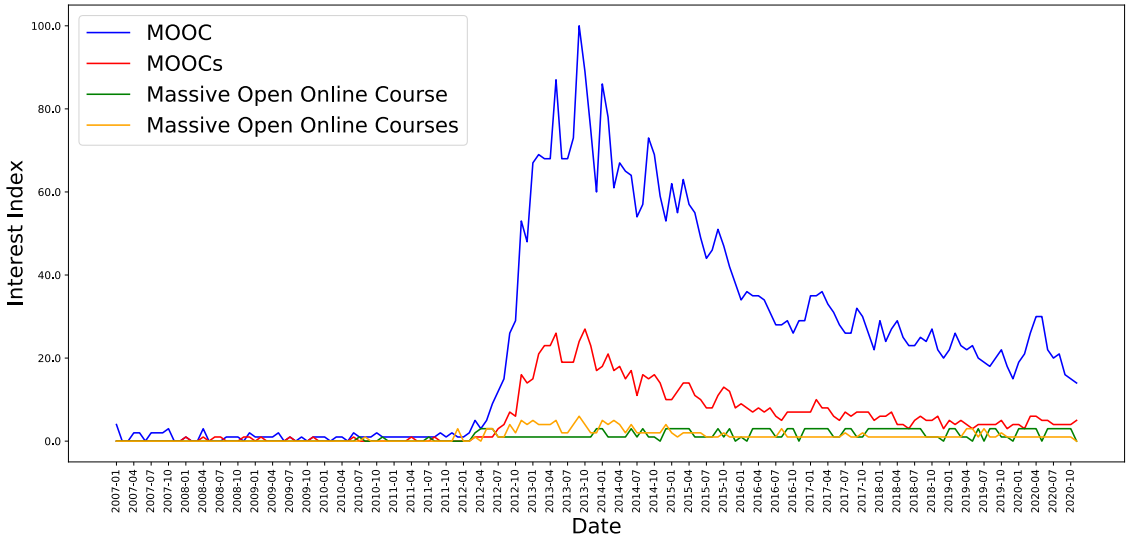

Figure 2: Google Trends for *MOOC*/*MOOC*/Massive Open Online Course

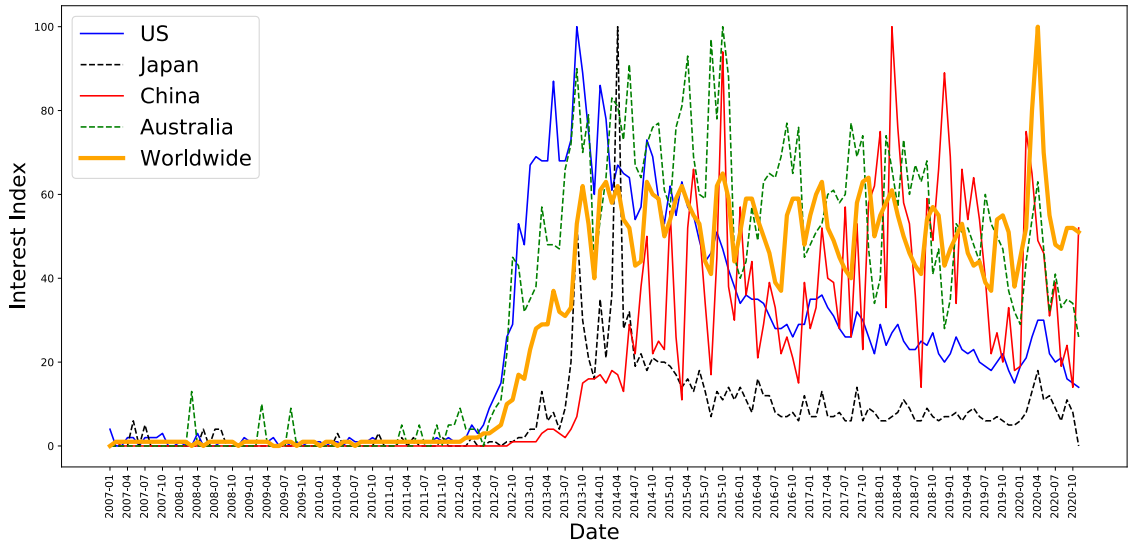

Figure 3: Google Trends for *MOOC*

China was that to the best of our knowledge, *MOOC* in China is still very popular now. We chose Japan simply because we are based in Japan. Many studies are based on *MOOC*s in Australia, so we also put Australia in the exploration list.

As shown in Figure 3, the Worldwide interest started to increase in early 2012, almost the same time as US. It reached the peak in about early 2014, a few months after US. Then the overall trend was almost stable unless there was a small decrease after mid 2018. The interest for *MOOC* in Australia started increasing at early 2012 and reached peak at around mid 2014. After that it gradually decreased to a lower level. Both Japan and China started increasing at around mid 2013 when the interest in US was about to peak. However, Japan reached the peak pretty fast, around April, 2014 while the peak in China did not appear until late 2018. Among all these trends, the decrease after the peak in Japan was the steepest. It did not take more than 3 months for the interest in Japan to suddenly drop from the very top to the bottom. Next, we will explore the trend of *MOOC* in the context of Gartner's Hype Cycle.

2.3 *MOOC* in the context of Gartner's Hype Cycle

2.3.1 The Gartner's Hype Cycle

The Hype Cycle is a model developed by Gartner to evaluate new technologies. This model divides the evolvement of a new technology into five phases: Technology Trigger, Peak of Inflated Expectations, Trough of Disillusionment, Slope of Enlightenment, and Plateau of Productivity. As shown in Figure 4.

The phase of Technology Trigger is when a breakthrough, public demonstration or other event generates public interest, venture capitalists start investing into this technology. As the awareness of this technology increases, the public start talking about all the potentials the technology may bring. More and more companies start

Figure 4: Hype Cycle (Source: Gartner research, May 2003)

exploring how this technology can be deployed to fit into their business strategies. High expectations are gradually generated. The technology moves to the second phase: Peak of Inflated Expectations. After this peak, public and companies find that this new technology does not meet their high expectations. Instead it may have a lot of problems. Because of this, interest in the technology suddenly drops to the bottom. This phase is Trough of Disillusionment. As some adopting companies still keep trying the new technology, a good business model which can make revenue is found. The public look at the technology more rationally. This technology is in its fourth phase, which is Slope of Enlightenment. When more and more companies and people are adopting the technology realistically, the technology reaches the last phase Plateau of Productivity (Dedehayir & Steinert, 2016; Linden & Fenn, 2003).

2.3.2 Gartner's Hype Cycle for *MOOC*

In the case of *MOOC*, we can also find a similar pattern of the Hype Cycle, In (Bozkurt, Keskin, and de Waard 2016) the authors identified key events such as Open Educational Practices, Open Educational Resources and first *MOOC* as the phase of Technology Trigger. *MOOC* reaches the Peak of Inflated Expectations when the first commercial *MOOC* platforms and the year of *MOOC* emerged. And Declaration of Anti-*MOOC* year was identified as the next phase: Trough of Disillusionment. After the trough, some more explorations like hybrid *MOOC* happened and meanwhile the research on *MOOC* were increasing. This made *MOOC* move into the Slope of Enlightenment phase. There was no significant event to indicate *MOOC* in the Plateau of Productivity phase. However, based on the consideration of the rapid progress, emerging business models, increasing educational adoption, year 2015 was identified as the beginning of Plateau of Productivity phase of *MOOC*. Putting aside all the hype and over-expectations, all the doubts and irony, we will reconsider the retention issue of *MOOC* in the next part.

Section 3 Related research

In last section, we introduced the background of *MOOC* to gain a better understanding of *MOOC*. In this section, we will review the related work regarding the retention in *MOOC*. We first explore the definition of retention and clarify the definition we are going to use in this paper. Then we introduce the current research related to retention in *MOOC* and identify the gap.

3.1 Definition of retention

One of the biggest challenges of *MOOC* is retention. Although so many learners enrolled in courses, only a very small proportion finally completed these courses (Breslow et al. 2013; Belanger, Thornton, and Barr 2013; Jordan 2014, 2015). The term retention or completion came from the traditional education. In (Levitz, Noel, and Richter 1999), retention is defined as "the percentage of first-time, full-time freshmen who return to the same institution for the second term or second year of study". In the context of e-learning (or online education), "retention rate was calculated by the number of students who withdrew from a course after the last date to drop from a course without any financial penalty" (El Said 2017). In the context of *MOOC*, many earlier discussion defined retention as "the fraction of individuals of those who initially enroll who successfully finish a course to the standards specified by the instructor" (Koller et al. 2013). The definition of retention in the context of *MOOC*/ traditional education/ e-learning are very similar. In this paper, we define retention as "The fraction who enroll successfully finish a course to the standards specified by the instructor". This is also referred to as retention rate, completion rate and certification rate in this paper.

3.2 Research on retention in *MOOC*

Why is retention a problem? The outcome of the education does not meet both the learners' and instructors' expectation. This is also a waste on the resource of *MOOC* especially when it takes hundreds of thousands dollars to produce one single *MOOC* (Hollands and Tirthali 2014). For platforms offering *MOOCs*, the users' churn will impact their sustainability.

Some research focused on exploring the features affecting retention. In (Dalipi, Imran, and Kastrati 2018), the authors identified that the reasons for dropout including chose another course instead, family issues and lack of time. They also found that the courses with a higher completion rate all had active discussion forums, complementing media and collaborative activities. (Adamopoulos 2013) found that: (a) the professor characteristic affect retention the most, (b) characteristics like assignments, course material and discussion forum also affect retention. Jordan (2014) found that completion rate is negatively correlated with course length. In a review paper on retention, (Khalil and Ebner 2014) identified that reasons of dropout including: lack of time, lack of motivation, feelings of isolation and the lack of interactivity, insufficient background and hidden costs. Besides the above-mentioned research, some research built prediction model to explore features affecting retention (Yang et al. 2013; Crossley et al. 2016; Dalipi, Imran, and Kastrati 2018; Gregori et al. 2018; Mongkhonvanit, Kanopka, and Lang 2019).

What's more, some researchers also moved forward and came up with some strategies that may improve the retention. In a review paper, (Khalil and Ebner 2014) identified a few techniques that increase the retention in *MOOC*. These techniques include

accommodating students on different time tables, promoting student completion and enhancing interaction.

However, many researchers argued that the definition of retention was not appropriate for *MOOC*. It needed to be re-operationalized and re-conceptualized. Some researchers studied learner's motivation and intention. (Koller et al. 2013) suggested that "Retention in *MOOC* should be considered in the context of learner intent, especially given the varied backgrounds and motivations of students who choose to enroll. When viewed in the appropriate context, the apparently low retention in *MOOC* is often reasonable". In another research, (Reich 2014) found that on average 22 percent of students who intended to complete a course earned a certificate, compared with 6 percent of students who did not intended to complete. The retention rate was much higher for those who intended to complete a course. However, the intention to complete was captured only in the pre-course survey. Some students may soon change their mind. In order to get a better understanding of this retention problem in *MOOC*, we also need to consider the students' actual learning activities. After interests in *MOOC* experienced big ups and downs, now it is stable in the worldwide scope. As in the context of the Hype Cycle, *MOOC* is identified in the Plateau of Productivity phase. It is really time for us to further explore the retention in *MOOC*. So that the *MOOC* stake-holders can get some insights when they are developing/ operating or learning a *MOOC*.

Section 4 Datasets

In last section we reviewed the related research on retention in *MOOC* and identified the gap. To eliminate the gap, we will need to conduct our own research. In this section we will first introduce the datasets we are going to use in this research. The data wrangling and feature engineering process will also be introduced in this section to illustrate how we get to the working dataset.

4.1 Original datasets

There are two datasets that we are going to use. One is the HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset[1]. The other one is the HarvardX-MITx Year 4 Report Appendix B dataset[2]. Since the HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset was used in the HarvardX-MITx year 1 report (Ho et al. 2014), this dataset will be labeled as year 1 dataset. The other dataset comes from the HarvardX-MITx year 4 report (Chuang and Ho 2016), it will be labeled as year 4 dataset.

The year 1 dataset is comprised of de-identified data from the first year (academic year 2013: fall 2012, spring 2013, and summer 2013) of MITx and HarvardX courses on the edX platform. These data are aggregate records, and each record represents one individual's activity in one edX course. The dataset is of version 2.0, and it was created on May 14, 2014. The data description of the year 1 dataset is shown in Table 3.

---

[1] https://doi.org/10.7910/DVN/26147
[2] http://year4.odl.mit.edu/appendix.html

The year 4 dataset is a course level statistical dataset. It includes 290 courses offered by HarvardX and MITx from fall 2012 to summer 2016. The data description of this year 4 dataset is shown in Table 4.

4.2 Working datasets

4.2.1 Year 1 dataset data wrangling and feature engineering

The original year 1 dataset has 20 features with 641,138 records. After preliminary exploration we found that the original dataset was pretty dirty. Some features had missing values, some had unreliable values. The data type of some features was not right. We used Python to clean the data. Numpy and pandas libraries were imported to do the data wrangling work. The data wrangling work that was done includes:

- **Subset**
  - Filter to get a subset without inconsistent records
  - Drop feature *registered*
  - Drop the unreliable *YoB* values
  - For feature *grade*, drop the unreliable values, drop the values in blank
  - Drop feature *roles*
- **Data type**
  - Convert the data type of *grade* from object type to numerical type
  - Convert *start_time_DI* and *last_event_DI* from object to date type.
- **Feature names**
  - Rename some of the columns names to be more concise and intuitive

Table 3: Year 1 dataset

Feature	Type	Description
course_id	String	ID of course, identifies institution (HarvardX or MITx), course name, and semester
userid_DI	String	First portion identifies dataset, second portion is a random ID number.
registered	Boolean	Whether registered in the course or not.
viewed	Boolean	Whether accessed the 'Courseware' tab (the home of the videos, problem sets, and exams) within the edX platform for the course or not
explored	Boolean	Whether accessed at least half of the chapters in the courseware or not
certified	Boolean	Whether earned a certificate or not. Certificates are based on course grades, and depending on the course, the cutoff for a certificate varies from 50% - 80%.
final_cc_cname_DI	String	Country of the student
LoE_DI	String	Highest level of education completed. Possible values: "Less than Secondary," "Secondary," "Bachelor's," "Master's," and "Doctorate."
YoB	Numerical	Year of birth.
gender	String	Possible values: m (male), f (female) and o (other).
grade	String	Final grade in the course, ranges from 0 to 1.
start_time_DI	String	Date of course registration.
last_event_DI	String	Date of last interaction with course, blank if no interactions beyond registration.
nevents	Numerical	Number of interactions with the course, recorded in the tracking logs; blank if no interactions beyond registration.
ndays	Numerical	Number of unique days student interacted with course.
nplay_video	Numerical	Number of play video events within the course.
nchapters	Numerical	Number of chapters (within the courseware) with which the student interacted.
nforum_posts	Numerical	Number of posts to the discussion forum.
roles	Boolean	Identifies staff and instructors, but blank as staff and instructors were removed from this release.
incomplete_flag	Boolean	Identifies records that are internally inconsistent.

Table 4: Year 4 dataset

Feature	Type	Description
Institution	String	Name of the institution offering the course
Course Number	String	The course number
Launch Date	String	Launch date
Course Title	String	Course title
Instructors	String	Instructors names
Course Subject	String	Course subject
Year	Integer	The year the course was launched
Honor Code Certificates	Boolean	Whether the course offers an honor code certificate or not, value 1 or 0
Participants (Course Content Accessed)	Integer	The number who have accessed the course content at least once over the duration of the course
Audited (> 50% Course Content Accessed)	Integer	The number who have accesses more than 50% of the course content at least once over the course duration
Certified	Integer	The number who have completed a specific proportion of course requirement over the course duration, the specified prop vary from course to course
% Audited	Numerical	The percent of the *audited* among the participants in a course, see *audited*
% Certified	Numerical	The percent of the certified among the participants in a course, see certified
% Certified of > 50% Course Content Accessed	Numerical	The percent of the certified among *audited*, see certified
% Played Video	Numerical	The percent of playing video at least once among participants
% Posted in Forum	Numerical	The percent of posting in forum at least once among participants
% Grade Higher Than Zero	Numerical	The percent of grade higher than zero among participant
Total Course Hours (Thousands)	Numerical	The total participant hours of the course
Median Hours for Certification	Numerical	Median hours for certification
Median Age	Numerical	Median age of the participants
% Male	Numerical	The percent of the male among participants
% Female	Numerical	The percent of the female among participants
% Bachelor's Degree or Higher	Numerical	The percent of bachelor's degree of higher among participants

- **Missing values**

    - Drop missing values in features *LoE_DI, YoB, gender* and *grade*
    - For missing values in features *nevents, ndays_act, nplay_video, nchapters*, fill them with 0
    - The value of *last_event_DI* is left blank if no interactions beyond registration, we fill with the value of *start_time_DI* which is the registration date.

After cleaning, our dataset has 402,750 records, with 17 features. The data description of the cleaned dataset is shown in Table 5.

After exploratory data analysis, we created new features for further exploration. Table 6 shows the description of the new features.

4.2.2 Year 4 dataset data wrangling

The original year 4 dataset has 23 features with 290 records. It is pretty clean compared to the year 1 dataset. However, there are still some problems with the dataset. Below is a summary of the problems and how we deal with them.

- The *Instructors* feature has a missing value, but this won't affect our analysis, we will leave it as it is.
- The data type of *% Played Video* shown as object. That is not correct. It should be float.
- The values of course subjects are too long. We will replace them.

Table 5: Cleaned year 1 dataset

Feature	Pandas Type	Description
course_id	Object	ID of course, identifies institution (HarvardX or MITx), course name, and semester
user_id	Object	User id
viewed	Int64	Whether accessed the 'Courseware' tab (the home of the videos, problem sets, and exams) within the edX platform for the course or not
explored	Int64	Whether accessed at least half of the chapters in the courseware or not
certified	Int64	Whether earned a certificate or not. Certificates are based on course grades, and depending on the course, the cutoff for a certificate varies from 50% - 80%.
country	Object	Country of the student
education	Object	Highest level of education completed. Possible values: "Less than Secondary," "Secondary," "Bachelor's," "Master's," and "Doctorate."
YoB	Float64	Year of birth.
gender	Object	Possible values: m (male), f (female) and o (other).
grade	Float64	Final grade in the course, ranges from 0 to 1.
time_registered	Datetime64[ns]	Date of course registration.
last_event	Datetime64[ns]	Date of last interaction with course, blank if no interactions beyond registration.
nevents	Float64	Number of interactions with the course, recorded in the tracking logs; blank if no interactions beyond registration.
ndays_act	Float64	Number of unique days student interacted with course.
nplay_video	Float64	Number of play video events within the course.
nchapters	Float64	Number of chapters (within the courseware) with which the student interacted.
nforum_posts	Int64	Number of posts to the discussion forum.

Table 6: New features in year 1 dataset

Feature	Pandas Type	Description
institution	Object	Identifies which institution the course is offered by. HarvardX or MITx. This is extracted from the course_id feature
course_code	Object	The short name of the course. This is also extracted from the course_id feature
semester	Object	The semester the course was offered. Extracted from the course_id feature
age	Float64	The age of the learner. Calculated from year_registered feature and YoB feature.
course_launch	Datetime64[ns]	The course launch date. This information was acquired from table 1 in HarvardX and MITx year 1 report.
course_wrap	Datetime64[ns]	The course wrap date. This information was acquired from table 1 in HarvardX and MITx year 1 report.
course_duration	Int64	The duration of the course. Calculated from course_launch and course_wrap
registered_launch_delta	Int64	Identifies the duration between registration date and course launch date
lifetime_proportion	Float64	The proportion of learner's lifetime in the course duration. Calculated from lifetime feature and course_duration feature

- Some feature names are too long, we will rename them.
- *Audited (> 50% Course Content Accessed)* is redundant with *% Audited*, *Certified* is redundant with *% Certified, % Male* is redundant with *% Female*. Since these redundant features won't provide us any additional information, we will drop them.
- If a course did not offer honor code certification, there are two situations: 1) The course offered no certificates at all or 2) Only offered verified certificates. For courses did not offer honor code certification, four courses had no students got certified. While the other courses all had students got certified. We can tell that the four courses with no one got certified offered no

certificates at all. We will need to delete these four records. And then we will drop the *Honor Code Certification* column since it is no longer useful.

The data description of the cleaned dataset is shown in Table 7.

Table 7: Cleaned year 4 dataset

Feature	Type	Description
Institution	String	Name of the institution offering the course
CourseNumber	String	The course number
LaunchDate	Datetime	Launch date
CourseTitle	String	Course title
Instructors	String	Instructors names
CourseSubject	String	Course subject
Year	Integer	The year the course was launched
Participants	Integer	The number who have accessed the course content at least once over the duration of the course
PercentAudited	Numerical	The percent of the *audited* among the participants in a course, see *audited*
PercentCertified	Numerical	The percent of the certified among the participants in a course, see certified
PercentcertifiedAudited	Numerical	The percent of the certified among *audited*, see certified
PercentPlayedVideo	Numerical	The percent of playing video at least once among participants
PercentPosted	Numerical	The percent of posting in forum at least once among participants
PercentGradeHigherThanZero	Numerical	The percent of grade higher than zero among participant
TotalCourseHours	Numerical	The total participant hours of the course
MedianHoursCertification	Numerical	Median hours for certification
MedianAge	Numerical	Median age of the participants
PercentFemale	Numerical	The percent of the female among participants
PercentBachelororHigher	Numerical	The percent of bachelor's degree of higher among participants

Section 5 Descriptive and inferential statistical analysis

5.1 Introduction

From last section, we get the working dataset. In this section we will delve deep into the dataset to explore the retention problem in *MOOC*. Specifically, we will conduct a descriptive statistical analysis, followed by an inferential statistical analysis.

The dataset we were using in this part is the cleaned dataset from last section. Since this dataset includes the information of those who registered in the courses, we will regard it as the *registered dataset*. Through analysis of the *registered dataset*, we found that completion rate of the *registered* was misleading. Because many registrants did not come back to the course after registration. The distributions of the learners' grade and activities are all heavily skewed. Many learner's grade and number of learning activities are close to zero. We argue that retention in *MOOC* needs to be explored in the context of learners' intention. Based on the exploration of the *registered*, we defined intention to complete with three conditions below:

- The learner's lifetime is more than 13% of the course duration
- The learner interacted with at least 13% of the chapters (chapters are the highest organizational units in the edX courseware)
- The learner's grade is greater than zero

Then we explored the certification rate of intended to complete and features affecting certification.

5.2 Preliminary analysis

5.2.1 Basic information

The dataset for *registered* has 402,750 records with 25 columns. There are 16 unique courses in our dataset. The number of registrants varies a lot from course to course. The minimum enrolment is 3,316, while the maximum is 84,402. As shown in Table 8 and Figure 5.

Table 8: Number of registrants by course

Course ID	Number of registrants
HarvardX/CS50x/2012	84,402
MITx/6.00x/2012_Fall	52,504
HarvardX/ER22x/2013_Spring	40,444
MITx/6.00x/2013_Spring	36,595
HarvardX/PH207x/2012_Fall	32,337
HarvardX/PH278x/2013_Spring	26,277
MITx/6.002x/2012_Fall	23,738
MITx/14.73x/2013_Spring	20,840
HarvardX/CB22x/2013_Spring	19,860
MITx/8.02x/2013_Spring	16,198
MITx/7.00x/2013_Spring	14,287
MITx/6.002x/2013_Spring	11,885
MITx/3.091x/2012_Fall	10,112
MITx/8.mrev/2013_Summer	6,193
MITx/2.01x/2013_Spring	3,762
MITx/3.091x/2013_Spring	3,316

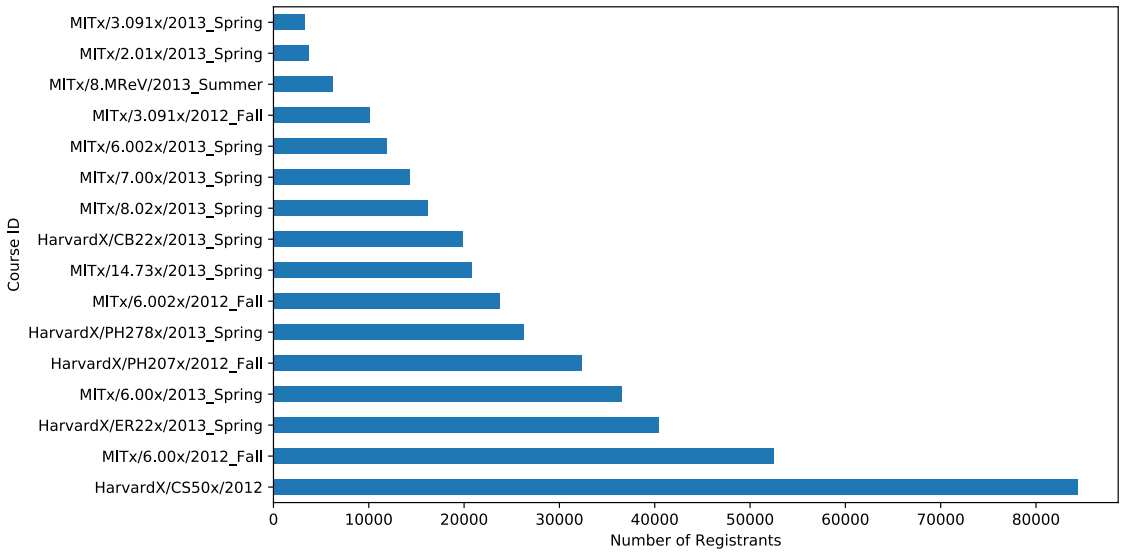

Figure 5: Number of registrants by course ID

Among all the 402,750 registrants, 239,49 viewed the course, which is about 60% of the registrants. 28,586 explored the course, this is less than 10% of the registrants.

5.2.2 Demographics

Country

As shown in Figure 6, the registrants come from a lot of different countries. US has the largest number of registrants, around 30%. India has the second largest number of about 15%.

Level of education

As much as 40% students hold a Bachelor's degree, while around 20% hold a Master's degree, 3% hold a Doctorate degree. That is to say, the majority of the students have an education level of Bachelor's or higher. We are educating the well educated in *MOOC*. Refer to Figure 7 below.

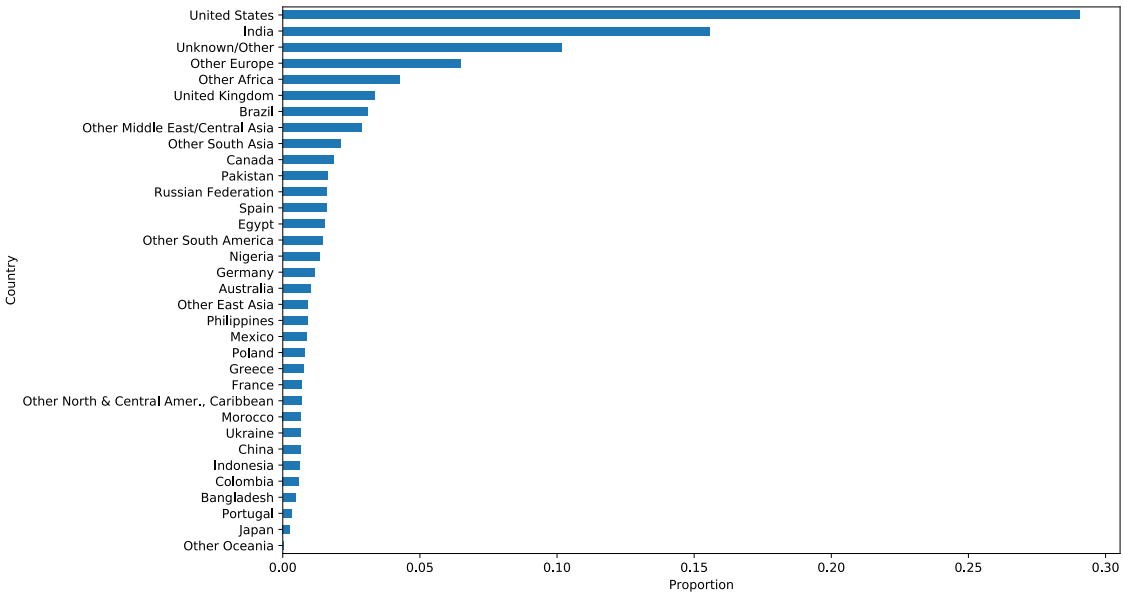

Figure 6: Proportion of registrants by country

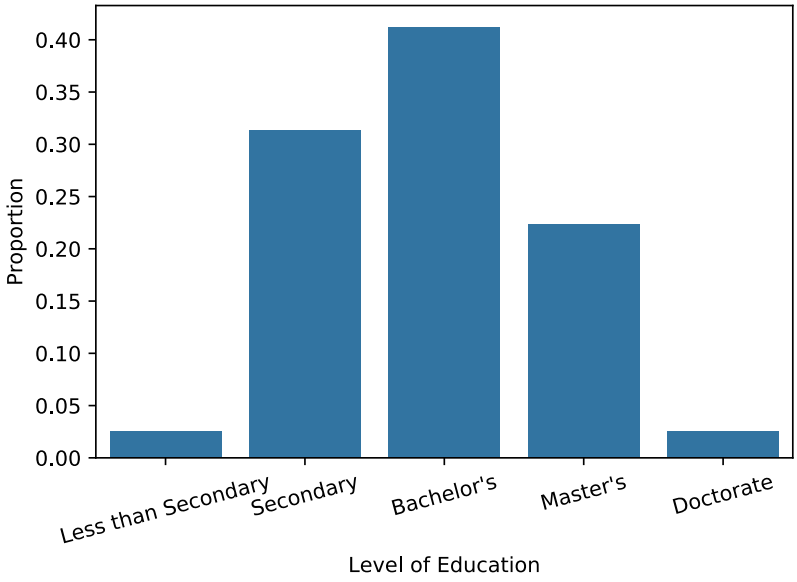

Figure 7: Proportion of registrants by level of education

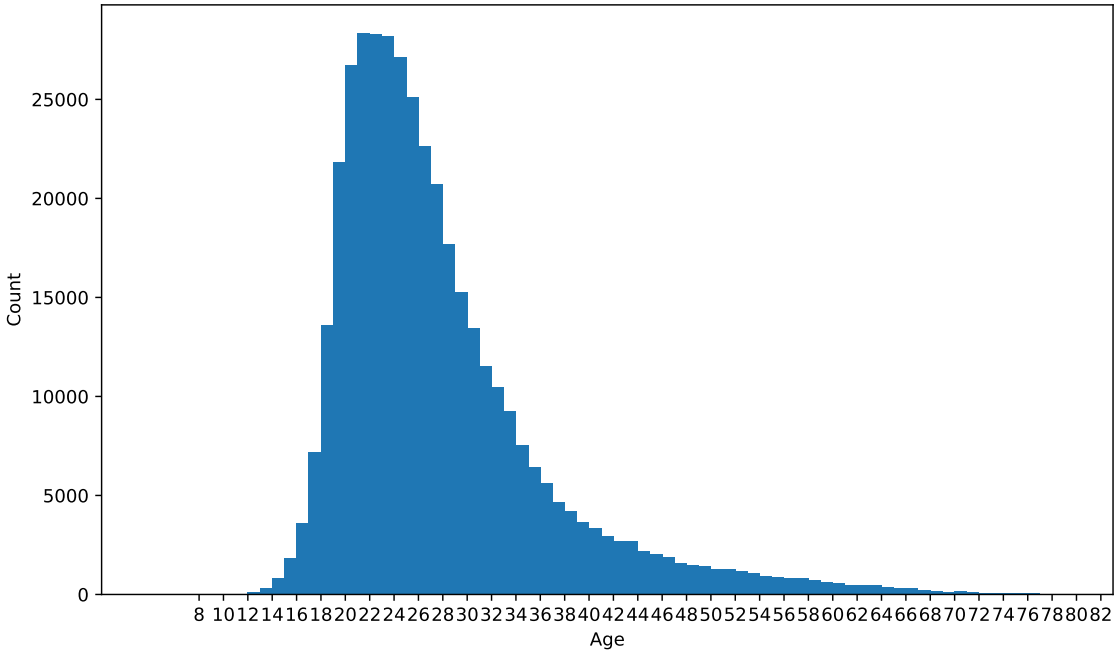

Figure 8: Distribution of age

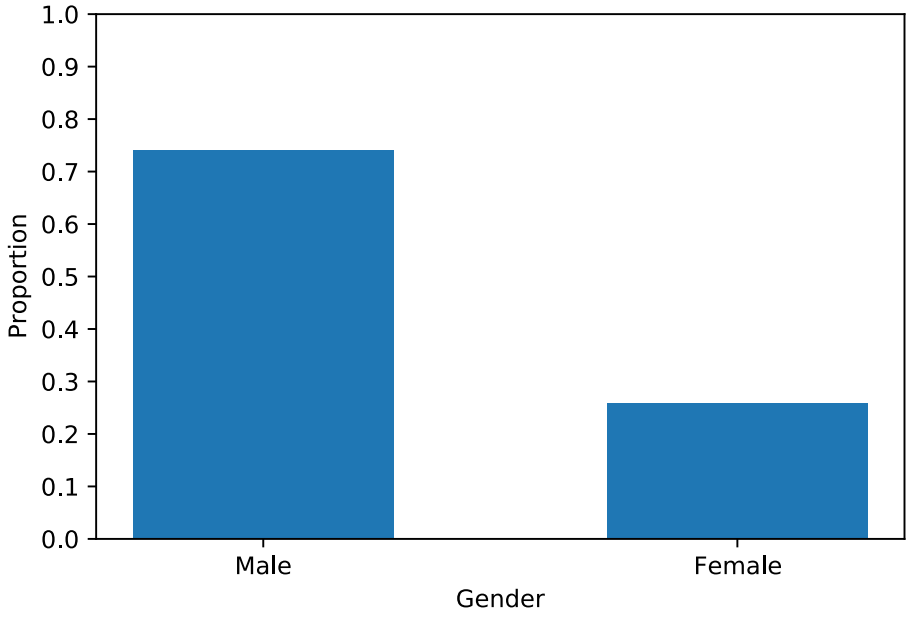

Figure 9: Proportion of registrants by gender

Age

The age of the learners varies from 10 years old to 82 years old. Most learners are in their 20s and 30s. The peak of the distribution appears at around 20 years old, and there is a long tail for more than 50 years old. The distribution of age is shown in Figure 8.

Gender

Per Figure 9, around 75% of the registrants are male, 25% are female.

5.3 Certification rate

Among 402, 750 who registered, only 13,673 finally got certified, the overall certification rate is about 3%. The certification rate of *viewed* is about 6%. The certification rate of *explored* is about 46%. The overall certification rates of different subpopulations are shown in Table 9.

The certification rate also varies a lot from course to course. As shown in Table 10. For the 16 courses, the certification rates of *registered* range from 1.26% to 9.04%, with a mean of 3.80% and a medium of 3.55%. The certification rates of *viewed* vary from 2.61% to 14.28%, with a mean of 6.15% and median of 5.86%. The certification rate of *explored* varies from 14.59% to 80.00%, with a mean of 57.21% and median of 57.45%.

Table 9: Overall certification rate of different subpopulations

Subpopulation	Count	Percentage of *registered* (%)	Overall certification rate (%)
Registered	402,750	100	3
Viewed	239,491	59	6
Explored	28,586	7	46

Table 10: Certification rate of different subpopulations by course

Course ID	Certification rate of registered (%)	Certification rate of viewed (%)	Certification rate of explored (%)
HarvardX/CB22x/2013_Spring	1.62	2.92	74.19
HarvardX/CS50x/2012	1.26	3.08	14.59
HarvardX/ER22x/2013_Spring	4.87	8.37	56.18
HarvardX/PH207x/2012_Fall	5.36	9.00	45.24
HarvardX/PH278x/2013_Spring	2.41	6.09	58.72
MITx/14.73x/2013_Spring	9.04	14.28	72.49
MITx/2.01x/2013_Spring	4.33	6.25	42.52
MITx/3.091x/2012_Fall	4.48	8.94	72.02
MITx/3.091x/2013_Spring	3.20	3.20	80.00
MITx/6.002x/2012_Fall	4.89	7.79	61.97
MITx/6.002x/2013_Spring	3.95	7.55	71.32
MITx/6.00x/2012_Fall	3.06	4.97	55.91
MITx/6.00x/2013_Spring	2.61	2.61	47.23
MITx/7.00x/2013_Spring	3.77	5.62	50.10
MITx/8.02x/2013_Spring	2.56	3.36	42.12
MITx/8.mrev/2013_Summer	3.33	4.37	70.80
Mean	3.80	6.15	57.21
Standard deviation	1.83	3.11	16.78
Minimum	1.26	2.61	14.59
First quantile	2.60	3.32	46.73
Median	3.55	5.86	57.45
Third quantile	4.58	7.94	71.49
Maximum	9.04	14.28	80.00

5.4 Better understanding of certification rate

As mentioned above, only 239,491 of the 402,750 registrants viewed the course. The percentage is about 59%. That is to say, 41% of the registrants never came back after registration.

The distribution of grade is seriously right skewed. Most learners have a grade of close to zero. As many as 86% learners have a grade of zero.

Similar to grade, the distribution of learners' activities is also heavily right skewed. The number of events, number of active days, number of videos played, number of chapters viewed, number of forum posts are close to zero for most of the learners.

All these indicate that many learners did not intend to complete the course. The completion rate of all the registrants is meaningless and misleading, we need to explore completion rate in the context of learners' intention.

5.5 Operationalize intention to complete

We totally agree that the *MOOC* learners are behaving differently from the traditional education learners. Some registered just to bookmark a course and many of them never came back; Some prospective college students registered to decide which major to take; Some registered to audit the course, did not mean to complete the course. However, we argue that there are still some students who intended to complete the course. Their behavior is similar to that of the traditional education. This subpopulation is our target of research. We will define intended to complete below to explore their retention pattern. Intended to complete is defined as those who have intention to complete a

course. The intention to complete is defined with the three conditions:

- Condition one: The learner's lifetime is more than 13% of the course duration
- Condition two: The learner interacted with at least 13% of the chapters
- Condition three: The learner's grade is greater than zero

As we know, in traditional on-campus courses, there are two weeks that the students can audit the course to decide whether to take the course or not. Two weeks is roughly 13% of a semester long course. In our case of *MOOC*, we will use this 13% to filter intended to complete. If the lifetime of a learner is more than 13% of the course duration, we assume he is more committed and more likely to intend to complete the course. This is condition one.

However, condition one alone may not be enough. Learner's lifetime is a point estimate only. We do not know the distribution of one's lifetime. Thus, we also need to make sure a learner has enough amount of activities during his lifetime. Especially during the 13% of the course duration. Here we will also use 13% of chapters viewed as a condition two to filter intended to complete.

With condition one and condition two above, we should be able to select most of intended to complete. But among these learners there are some learners who interacted a lot with the course but did not meant to complete the course. We assume that these learners are less likely to do the quizzes and assignments thus their grade may be pretty low. To exclude them from our subset, we will make sure all intended to complete have a grade greater than zero. This is our condition three.

We have the *lifetime_proportion* feature and *grade* feature in our dataset. Condition

one and condition three are very straightforward to conduct. However, condition two is not that straightforward. We have the number of chapters the learner viewed, but we do not have the number of chapters in the course. We assume that for each course at least one learner who certified view all the chapters. We will regard the maximum of number of chapters viewed as the number of chapters of the course. Then we will calculate the proportion of chapters viewed in total chapters of the course.

5.6 Certification rate of intended to complete

In the last part, we defined intended to complete and got a subset of intended to complete. In this part, we will explore the certification rate of this subpopulation.

Intended to complete dataset has 48,278 records. That is to say, 48,278 registrants intended to complete the course out of all 402,750 registrants. This is about 12%. Among these 48,278 registrants who intended to complete, 13,663 of them finally got certified. This is about 28%. The overall certification rate of intended to complete is about 28%.

The overall certification rate is shown in Table 11 below. The above is the overall certification rate. The certification rate by course is shown in Table 12 and Figure 10.

We can see that the certification rate by course varies a lot from course to course. The CS50X course from HarvardX even has a certification rate of 100%.

Table 11: Overall certification rate

Subpopulation	Count	Percentage of *registered* (%)	Overall certification rate (%)
Registered	402, 750	100	3
Viewed	239,491	59	6
Explored	28,586	7	46
Intended to complete	48,278	12	28

Table 12: Certification rate of intended to complete by course

Course ID	Certification rate of intended to complete (%)
HarvardX/CB22x/2013_Spring	25
HarvardX/CS50x/2012	100
HarvardX/ER22x/2013_Spring	44
HarvardX/PH207x/2012_Fall	29
HarvardX/PH278x/2013_Spring	26
MITx/14.73x/2013_Spring	42
MITx/2.01x/2013_Spring	22
MITx/3.091x/2012_Fall	36
MITx/3.091x/2013_Spring	28
MITx/6.002x/2012_Fall	32
MITx/6.002x/2013_Spring	28
MITx/6.00x/2012_Fall	19
MITx/6.00x/2013_Spring	13
MITx/7.00x/2013_Spring	26
MITx/8.02x/2013_Spring	21
MITx/8.mrev/2013_Summer	29

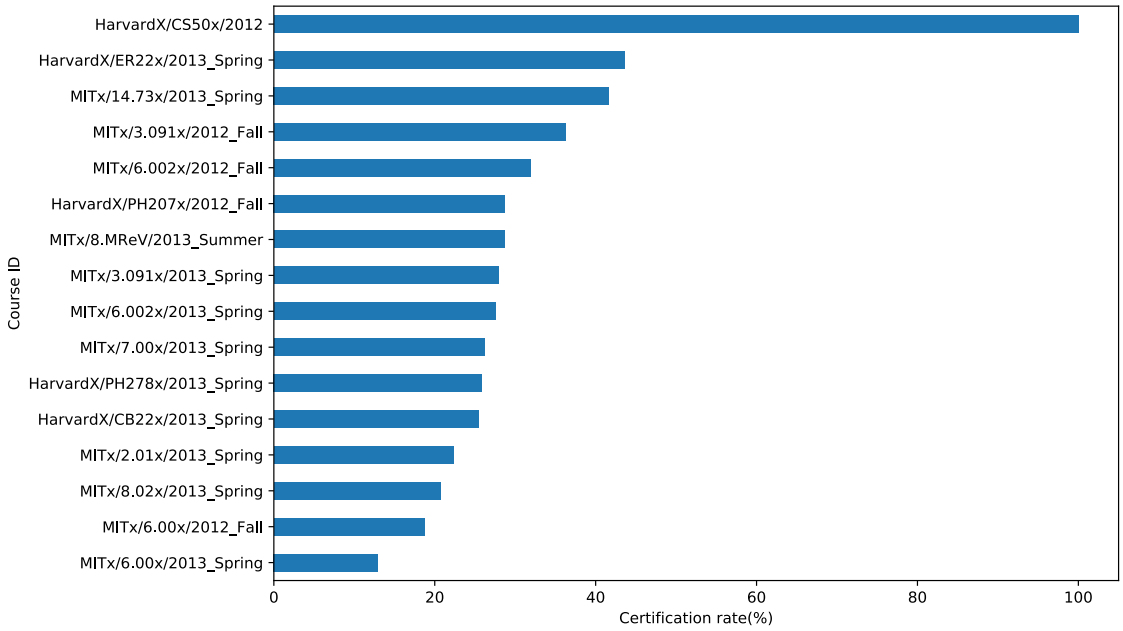

Figure 10: Certification rate of intended to complete by course

5.7 Features affecting retention

We explored the certification of intended to complete above. In this part, we will look at what features affect certification rate.

5.7.1 Dependent features and independent features

There are three dependent features: *certified, grade* and *certification rate*. *certified* identifies whether the learner completed the course and got a certification or not. The value would be 1 if the learner got a certification, 0 otherwise. This is on learner level.

grade is the final grade in the course, ranges from 0 to 1. Learners need to get at least a certain grade to be certified. This is also on learner level. *certification rate* is proportion of certified among intended to complete. This is on course level.

For independent features, we will divide them into three categories:

Course related features: Including *course_id, institution, semester, course_duration, course_chapters*

Learners demographic features: Including *country, level of education, age, gender, region.*

Learners activity features: Including *user_id, nevents, ndays_act, nplay_video, chapters_proportion, nforum_posts, registered_launch_delta, lifetime_proportion*

5.7.2 Course related features

In this part, we explore course related features: Including *course_id, institution, course_code, semester, course_duration, course_chapters*.

Certification rate by institution by course

For the 16 courses in our dataset, 5 are offered by HarvardX. The other 11 are offered by MITx. Figure 11 shows that:

- The median certification rates between the two institutions are close
- The minimum, first quantile, third quantile and maximum of HarvardX courses certification rate are all higher than those of MITx.
- There is an outlier for HarvardX. The certification rate is 100%. This outlier is not anomaly. We will leave it as it is.

Table 13 shows that:

- The 5 HarvardX courses have a mean certification rate of 45%, a median of 29%.
- The 11 MITx courses have a mean certification rate of 27% and a median of

28%

- The average certification rate of HarvardX courses is higher than MITx

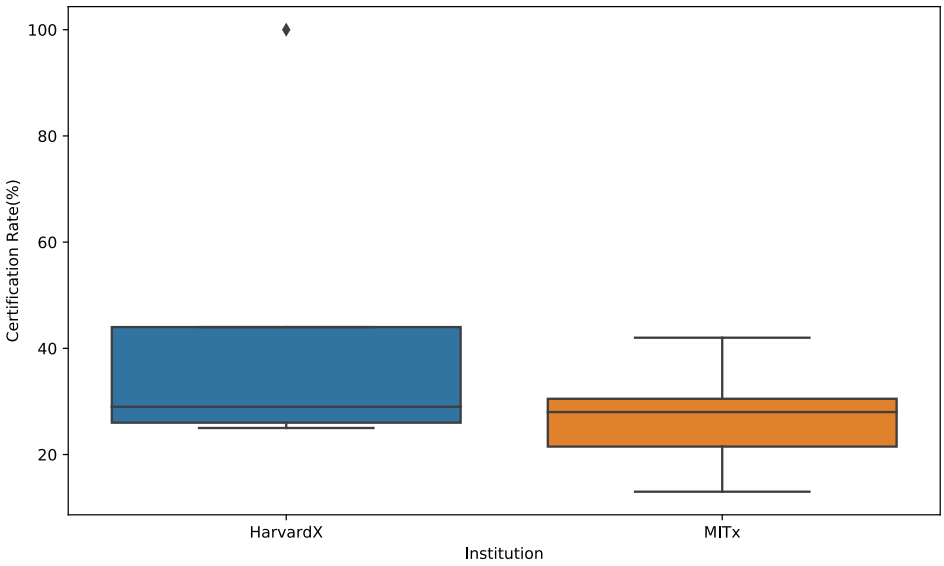

Figure 11: Certification rate by institution

Table 13 : Statistics of HarvardX and MITx courses certification rates

Statistics	HarvardX courses certification rate (%)	MITx courses certification rate (%)
Count	5	11
Mean	45	27
Standard deviation	32	8
Minimum	25	13
First quantile	26	22
Median	29	28
Third quantile	44	31
Maximum	100	42

To check how significant is the difference between mean certification rate of HarvardX and MITx, we conducted an independent *t*-test. The P-value is 0.09. The P-value is relatively large, there is no significant difference between the two means.

Certification rate by semester

There are three semesters in our dataset: semester of 2012 fall, semester of 2013 spring and semester of 2013 summer. There are five courses in semester of 2012 fall, 10 courses in semester of 2013 spring and only one course in semester of 2013 summer. Figure 12 is the boxplot of the certification rate of different semesters. From Figure 12 we can see that: There are some outliers. We will leave them as they are since they are not anomalies. The median certification rate for semester of 2013 spring is lower than semester of 2012 fall. The median certification rate for semester of 2013 summer is much higher than semester of 2013 spring. However, since there is only one course in semester of 2013 summer, this may be unreliable. We used *ANOVA* to test the significance. The P-value is 0.36. The P-value is relatively large, there is no sufficient difference between means of the three groups.

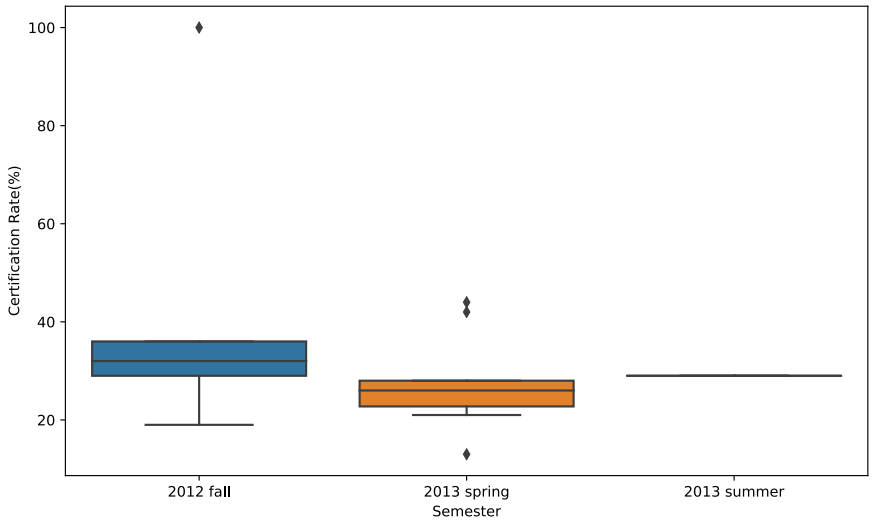

Figure 12: Certification rate by semester

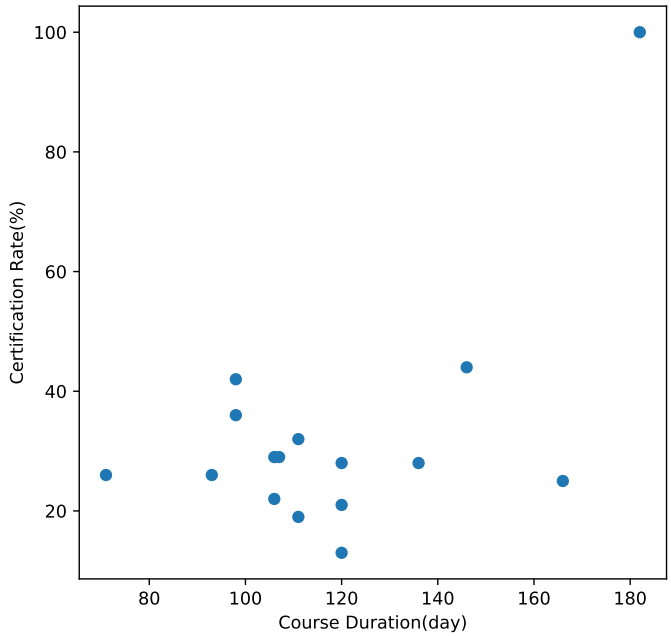

Figure 13: Course duration vs. certification rate

Course duration vs. certification rate

Will the course duration affect the certification rate? We will use a scatter plot to explore the relationship between course duration and certification rate. From Figure 13,

we can see that although there are some outliers, the overall trend is as course duration increases, certification rate decreases. However, the Pearson's correlation coefficient is 0.56 with a P-value of 0.02. This indicate that there is a moderate positive correlation between course duration and certification rate.

Number of course chapters vs. certification rate

Will the number of chapters in a course affect the certification rate? We explored the relationship between number of chapters in a course and certification rate in a scatter plot. Figure 14 shows that there is no obvious relationship between number of chapters in a course and certification rate. The Pearson's correlation coefficient is -0.16. This indicates a weak negative correlation. However, the P-value (0.54) indicates that the Pearson's correlation coefficient is not significant.

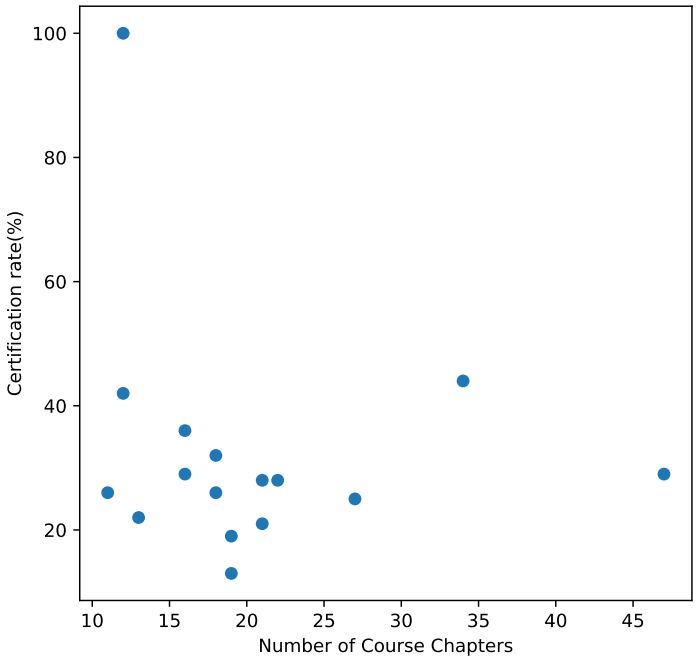

Figure 14: Number of chapters vs. certification rate

Course subject vs. certification rate

Do certification rates differ among course subjects? We compared the certification rate of courses from STEM and NON-STEM subject.

Figure 15 shows that there is one outlier with a certification rate of 100%. Courses from STEM subject have a relatively lower certification rate. How significant is this difference? Therefore, we conducted a *t*-test. The P-value of the *t*-test is about 0.12. The P-value is relatively large, there is no significant difference between the two means.

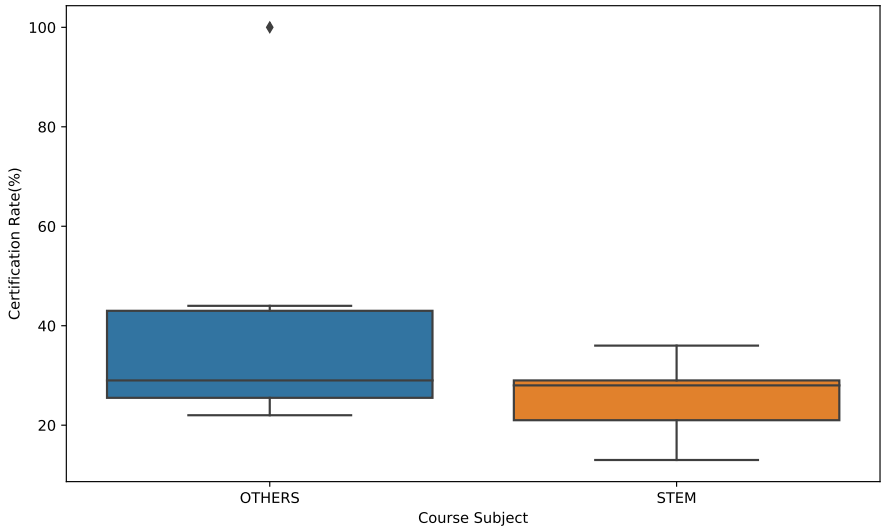

Figure 15: Course subject vs. certification rate

5.7.3 Demographic features

In this part, we explore learners' demographic features including *country, level of education, age, gender* and *region.* These features are on learner level. The sample size is very large, i.e. 48278. In inferential analysis, when the sample large is large, the P-values go quickly to zero. A small P-value may lead us to an unreliable conclusion. In (Lin, Lucas Jr, and Shmueli 2013), the authors proposed a few techniques to deal with large samples. Techniques include presenting effect size, reporting confidence intervals,

using charts. In this research we will report confidence intervals.

Country

Figure 16 shows the certification rate by country. We can see:

- The certification rate varies from country to country
- Most of the countries with a higher certification rate are European countries.

We conducted a chi-square test to check the significance of the difference among certification rates of different countries. The P-value < .001, we have sufficient evidence that there is a significant difference among certification rates of different countries.

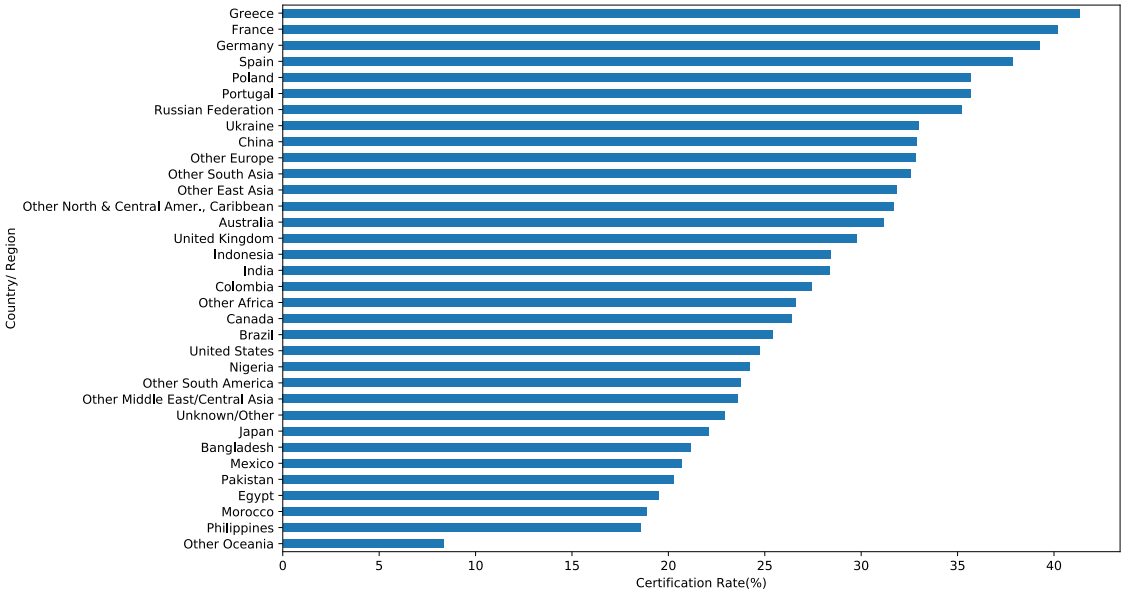

Figure 16: Certification rate by country

Level of education vs. certification rate

From Figure 17 we can see that the certification rates among different levels of education are close. Master has a relatively higher certification rate. We conducted a

chi-square test to check the significance of the difference among certification rates of different levels of education. The P-value P <.001, there is significant difference among certification rates of different levels of education.

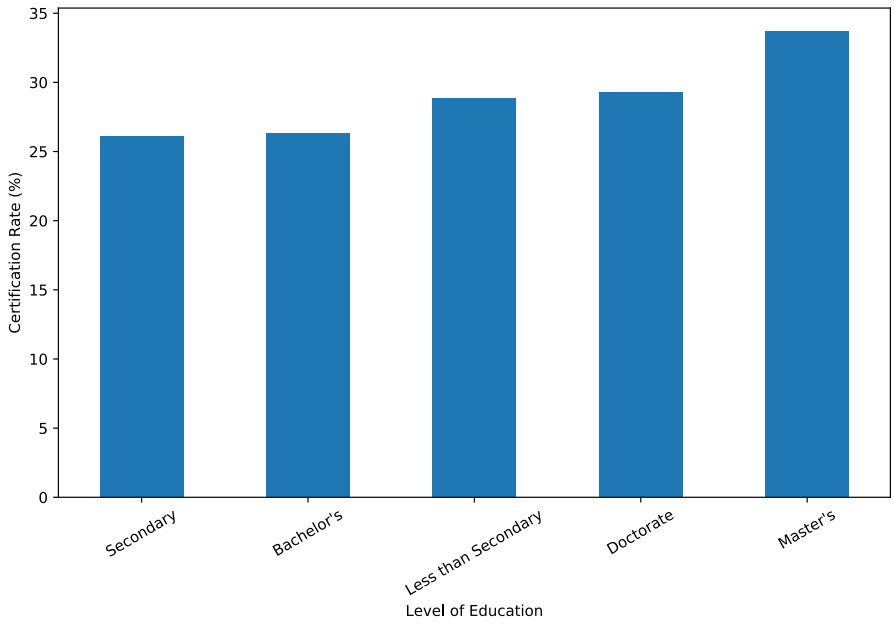

Figure 17: Certification rate by level of education

Age vs. certification

Figure 18 is the boxplot showing the distribution of age for certified and uncertified. There are some outliers. This is because some students have very old ages. We will include those outliers. We can tell that the distributions of age for certified and uncertified are similar to each other. The P-value we got from the two samples *t*-test is 0.19. The 95% confidence interval is $(-0.28, 0.06)$. This indicate that there is no difference between the average age of certified and uncertified.

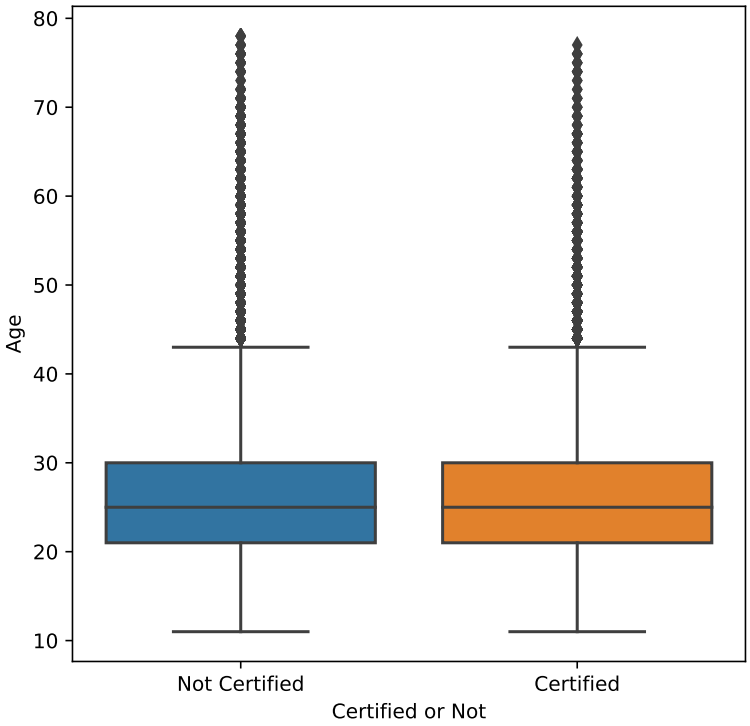

Figure 18: Age vs. certified or not

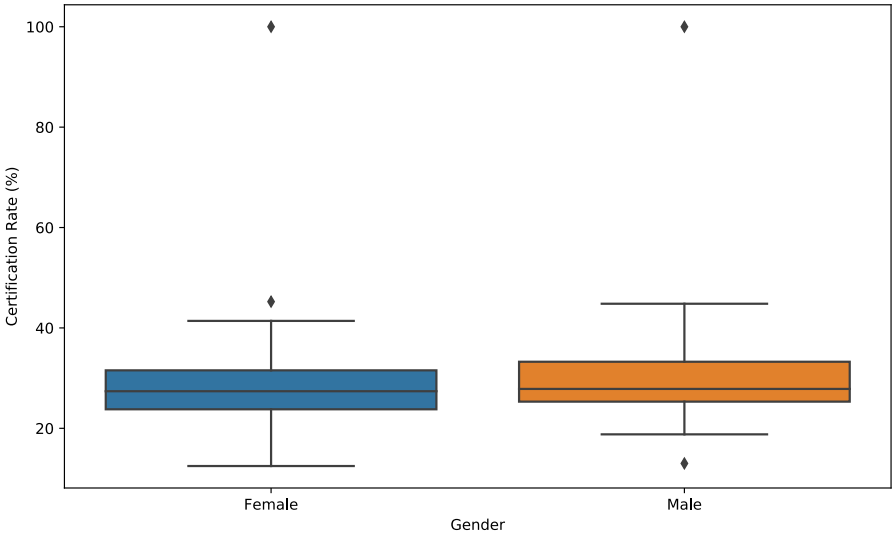

Figure 19: Gender vs. certification rate

Gender vs. Certification

Figure 19 shows that the certification rate of female learners is close to that of male learners. There are some outliers. We will include them since they are not anomalies. We conducted a chi-square test to check the significance. The P-value P<.001. The 95% confidence interval is (0.03, 0.05). There is a significant difference between certification rate of female learners and certification rate of male learners.

Region vs. certification

Based on the exploration above, we found that different countries have significant different certification rate. Does this hold true for different regions? We will explore the US region and Non-US region. Figure 20 shows that there are some outliers. Students from US region have a relatively higher certification rate than Non-US region. We also applied a chi-square test to check the significance. The P-value P<.001. The 95% confidence interval is (0.04, 0.06). There is a significant difference between certification rate of US region and Non-US region.

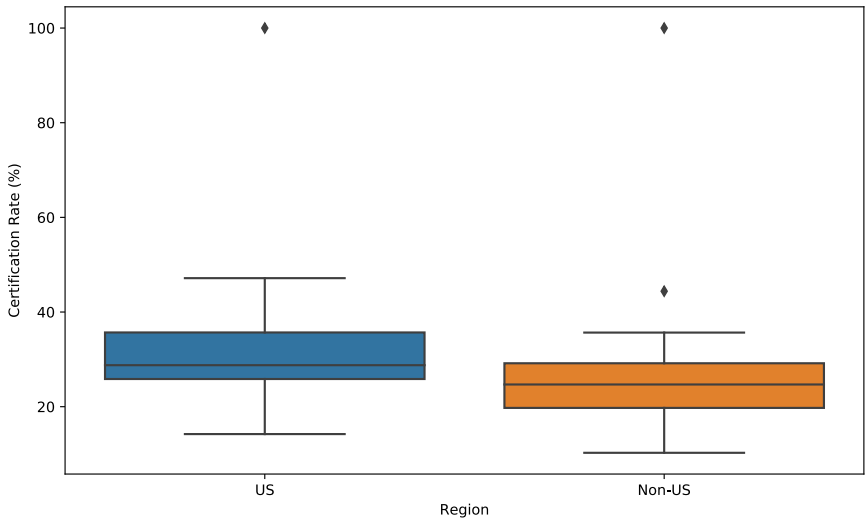

Figure 20: Region vs. certification rate

5.7.4 Learners' activity features

In this part, we explore learners' activity features: Including *user_id, nevents, ndays_act, nplay_video, chapters_proportion, nforum_posts, registered_launch_delta, lifetime_proportion*

User vs. semester vs. certification

Most learners were not certified for any course. While there were also some learners certified for as many as five courses. For those who had completed one course, are they more likely to complete a subsequent course?

We will first select the users who completed at least one course in the 2012 fall semester. Then we will look at the certification rate of those who registered in the 2013 spring semester among those who completed a course in the 2012 fall semester. We will also look at the overall certification rate of 2013 spring semester for comparation. We found that the 2013 spring semester certification rate for learners who completed at least one course in the 2012 fall semester is 47.91%. This is much higher than 2013 spring semester certification rate for all the learners which is 27.58%. We then conducted a chi-square test to test the significance. The P-value $P<.001$. There is sufficient evidence that these two certification rates are significantly different.

Number of events vs. grade

Regarding the relationship between number of events and grade, the Pearson's correlation coefficient is 0.54 with P-value $P<.001$. The 95% confidence interval is (0.53, 0.54). There is a moderate positive correlation between number of events and grade.

Number of active days vs. grade

The Pearson's correlation coefficient is 0.63 with a P-value P<.001. The 95% confidence interval is (0.63, 0.64). There is a strong positive correlation between number of active days and grade.

Number of videos played vs. grade

The Pearson's correlation coefficient is 0.28 with a P-value P<.001. The 95% confidence interval is (0.27, 0.29). This indicates a weak positive correlation between number of videos played and grade.

Proportion of chapters viewed vs. grade

Figure 21 shows learners who viewed more chapters are more likely to get a higher score. The Pearson's correlation coefficient is 0.84 with a P-value P<.001. The 95% confidence interval is (0.836, 0.841). This indicates a very strong positive correlation between proportion of chapters viewed and grade.

Number of forum posts vs. grade

The Pearson's correlation coefficient is 0.03 with a P-value P<.001. The 95% confidence interval is (0.02, 0.04). This indicates a very weak correlation between number of forum posts and grade.

Enrollment day relative to launch vs. grade

The Pearson's correlation coefficient is -0.06 with a P-value P<.001. The 95% confidence interval is $(-0.07, -0.05)$. This indicates a very weak correlation between these two features.

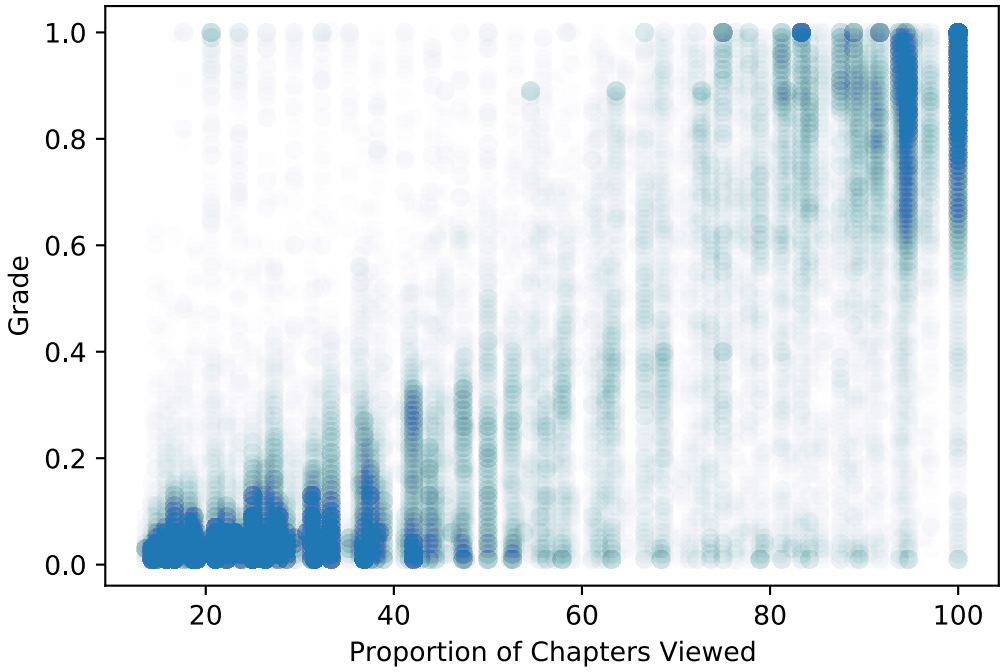

Figure 21: Proportion of chapters viewed vs. grade

Percentage of lifetime in course duration vs. grade

The Pearson's correlation coefficient is 0.41 with a P-value P<.001. The 95% confidence interval is (0.40, 0.41). This indicates a moderate positive correlation between these two features.

5.8 Exploration of year 4 dataset

In last part, we analyzed retention in year 1 dataset. What's the trend of retention over the years? What features are still correlated to retention? In this part, we will explore the year 4 dataset to get some insights.

5.8.1 Proxy of intended to complete.

We define intended to complete in the year 1 dataset. However, in year 4 dataset, there is no such information that we could directly define intended to complete. What the year

4 dataset has is the certification rate of participants as well as certification rate as *audited*. Participants refers to those who accessed the course content at least once in the course duration. It is corresponding to the *viewed* in the year 1 dataset. *Audited* refers to those who accessed more than 50% of the course content in the course duration. This is corresponding to the *explored* in the year 1 dataset.

Certification rate of participants

From Figure 22 we can see that most courses have a certification rate less than 10%. There are a few outliners with certification rate more than 30%.

Certification rate of audited

As shown in Figure 23, for *audited*, the average certification rate is about 30%. The highest certification rate is almost 80%.

Participants vs. intended to complete

Are certification rates of participants significantly different from intended to complete? We conduct a *t*-test check this out. The P-value P<.001. There is significant difference between the certification rate of the participants and intended to complete.

Audited vs. intended to complete

We also conduct a *t*-test to check the significance between the certification rates of the *audited* and intended to complete. The P-value is 0.95. The P-value is relatively large. There is no significant difference between the certification rates of the *audited* and intended to complete. We will use *audited* in the year 4 dataset as a proxy of intended to complete.

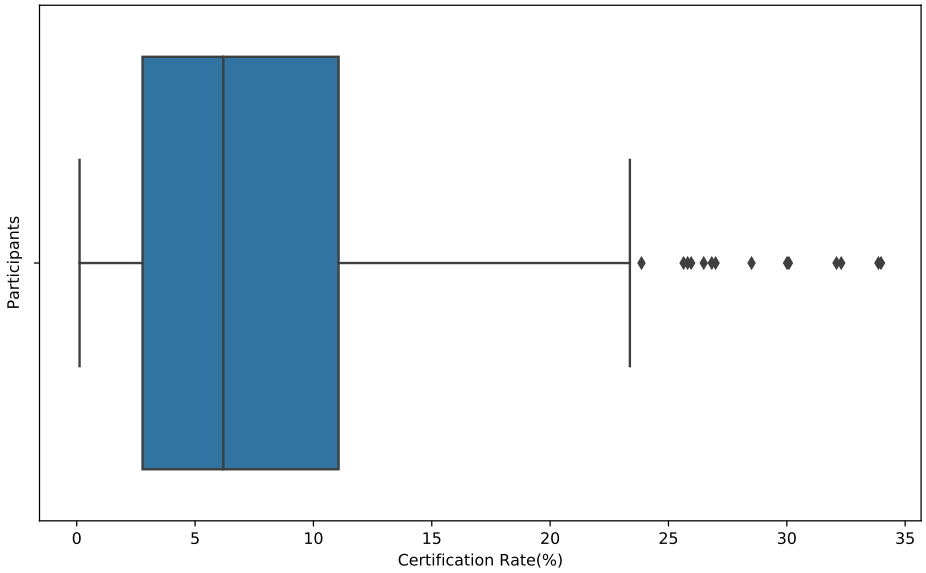

Figure 22: Certification rate for participants

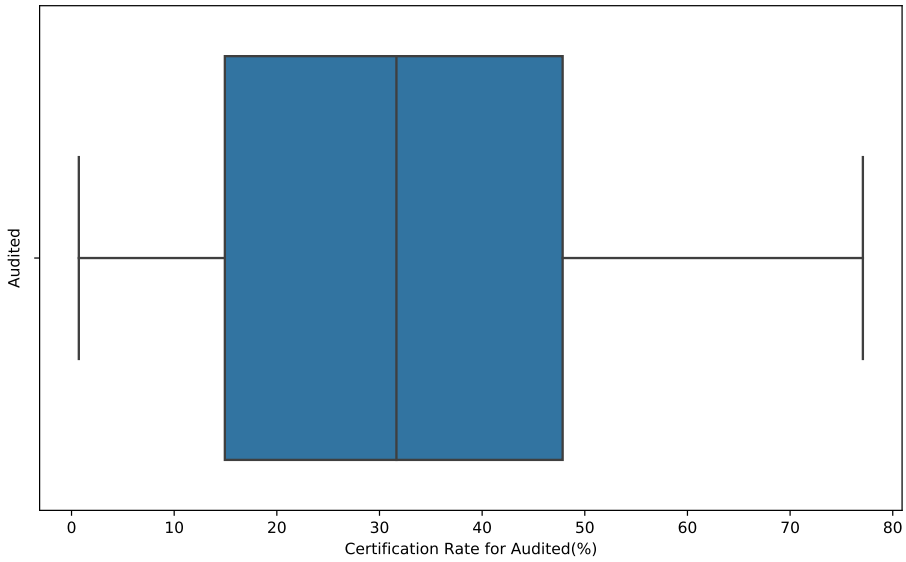

Figure 23: Certification rate for *audited*

5.8.2 Features related to certification

In the exploration of year 1 dataset, we found that:

- Features like *institution, semester, age, nforum_posts, registered_launch_delta, course_subject* have no correlation with certification.
- Features like *country, education, gender, region, nplayvideo, lifetime_proportion, course_duration, nevents, course_chapters, ndays_act, chapters_proportion* are correlated with certification. Among these variables, *course_chapters, ndays_act, chapters_proportion* have a strong correlation with certification.

In this part, we explore the year 4 dataset. We try to figure out what still hold true over time and what are new.

Similar to the exploration in year 1 dataset, we also divided the features into 3 categories for year 4 data. The categories are shown as below:

- **Course related features:** Including *Institution, CourseNumber, LaunchDate, CourseTitle, Instructors, CourseSubject, Year*
- **Learners demographic features:** Including *MedianAge, PercentFemale, PercentBachelororHigher*
- **Learners activity features:** Including *Participants, PercentAudited, PercentPlayedVideo, PercentPosted, PercentGradeHigherThanZero, TotalCourseHours, MedianHoursCertification*

Institution vs. certification

From Figure 24, we can see that courses from HarvardX have a relatively higher average certification rate than MITx. We conducted a *t*-test. The P-value is 1. There is no significant difference between the average certification rate of the two institutions.

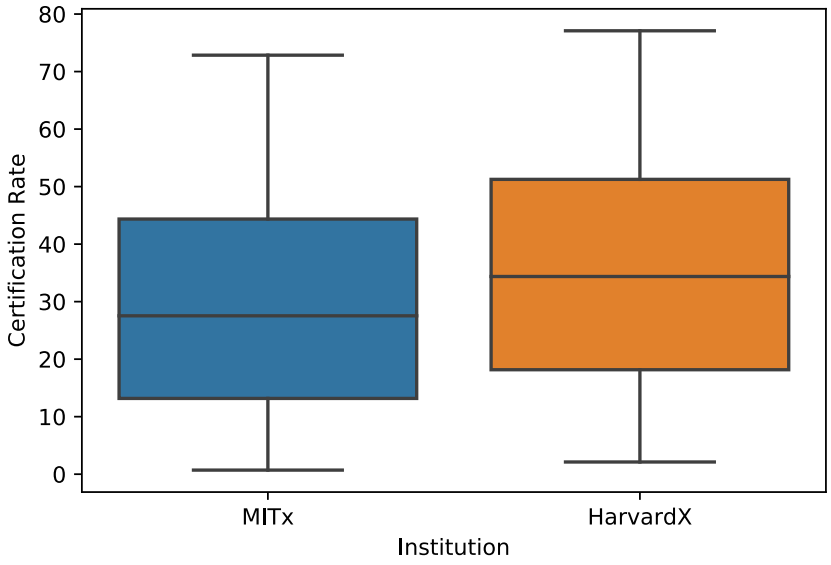

Figure 24: Institution vs. certification

Course subject vs. certification

Figure 25 shows that STEM (science, technology, engineering, and mathematics) has a relatively lower average certification rate. We conducted an *ANOVA* test to check how significant are the differences between the four groups. The P-value P<.001. This indicates that there is a significant difference between the means of the four groups.

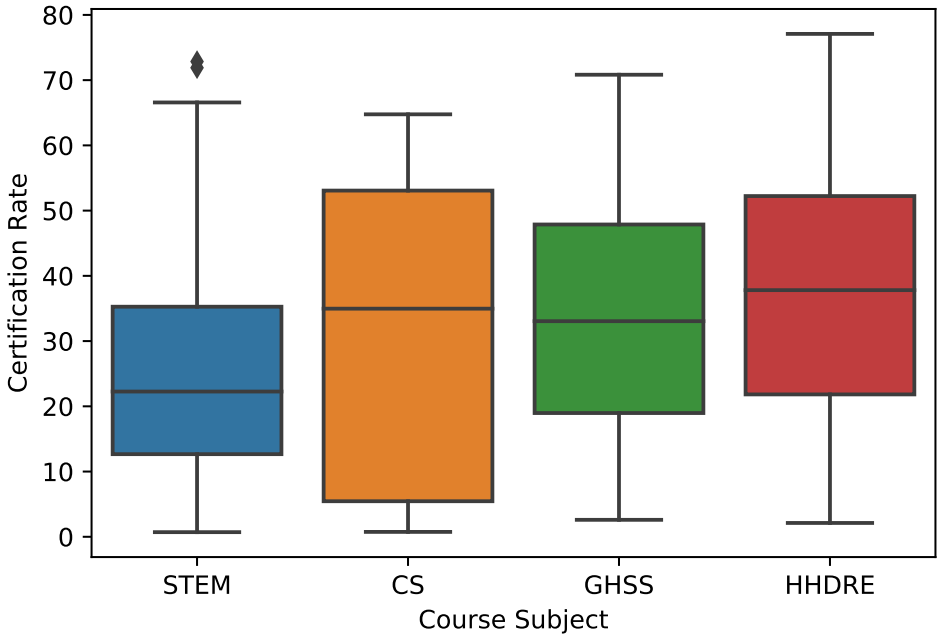

Figure 25: Course subject vs. certification

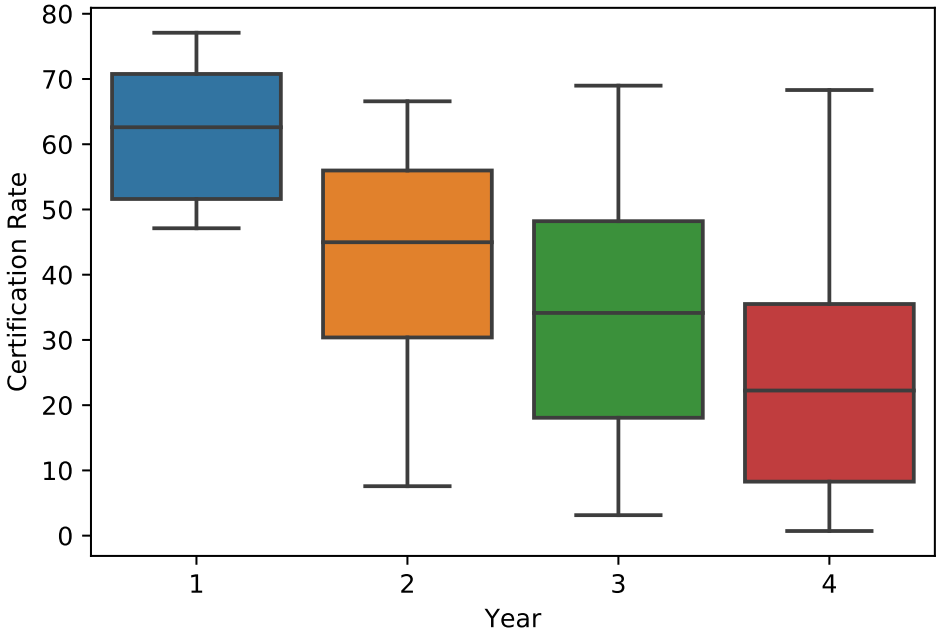

Figure 26: Year vs. certification

Year vs. certification

Figure 26 shows that the average certification rate is decreasing over the years. The P-value of the *ANOVA* test P<.001. There is significant difference between the means of the four years.

Demographic features vs. certification

For median age vs certification, the Pearson's correlation coefficient is about 0.21 with a P-value P<.001. This indicates a weak positive correlation. For percentage of female vs. certification, the Pearson's correlation coefficient is about 0.24 with a P-value P<.001. This indicates a weak positive correlation. For percentage of Bachelor's degree or higher vs. certification, the Pearson's correlation coefficient is about 0.1. however, the P-value is 0.08. There is no sufficient evidence that the correlation coefficient is 0.1.

Percent grade higher than zero vs. certification

Figure 27 shows the relationship between percent grade higher than zero and certification. The Pearson's correlation coefficient is about 0.66. This indicates a strong positive correlation.

Other learners' activity related features vs. certification

For number of participants vs. certification, the Pearson's correlation coefficient is about -0.03 with a P-value of 0.6. For percentage of *audited* among participants vs. certification, the Pearson's correlation coefficient is about -0.1 with a P-value of 0.08. For percent played video vs. certification, the Pearson's correlation coefficient is about

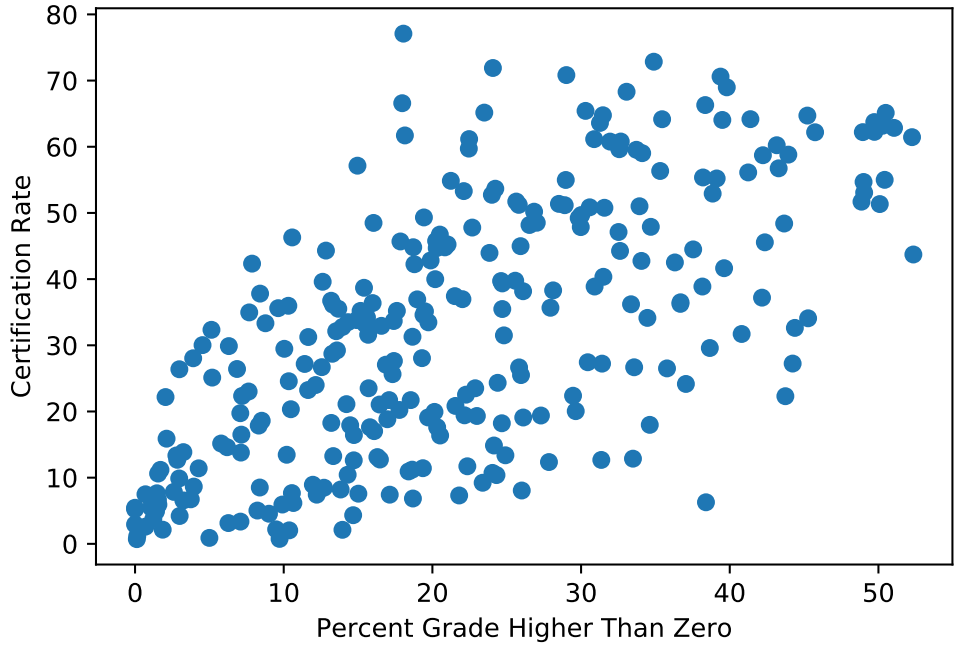

Figure 27: Percent grade > 0 vs. certification

0.27 with a P-value of 0.004. This indicates a weak positive correlation. For percent posted vs. certification, the Pearson's correlation coefficient is about 0.43 with a P-value P<.001. This indicates a weak positive correlation. For total hours vs. certification, the Pearson's correlation coefficient is -0.09. However, the P-value is 0.15, which is relatively large. For median hours for certification vs. certification the Pearson's correlation coefficient is about -0.02. However, the P-value is 0.15, which is relatively large.

Based on the exploration above, we found that features like course subject, launch year, age, gender, percent played video, percent posted and percent of grade higher than zero are correlated to certification. Among these, percent grade higher than zero has a relatively strong correlation. Others only have weak correlation with certification.

Section 6 *MOOC* retention prediction with machine learning

In the last section we explored the patterns of retention in *MOOC*. We also identified statistically significant features affecting retention in *MOOC*. How well can we predict retention with these identified features? In this section, we evaluate selected machine learning algorithms to identify how well these features can predict retention.

6.1 Dataset and features selection

There are two datasets we are using: the year 1 dataset and the year 4 dataset. The year 1 dataset is a learner level dataset. The year 4 dataset is a course level dataset. Since our prediction is learner level, we will use the year 1 dataset to train and test the models.

Table 14 shows the features we select to train and test the model. The dependent feature is *certified*. And the other features are independent features. The independent features are categorized into course related, demographics related, learners' activity related.

6.2 Preprocessing

In feature selection, we have selected the features to fit into the model. In this part, we preprocess the features. The package we used was the Scikit-Learn's *preprocessing* package. Some features are categorical features, we used LabelEncoder library to convert string to numbers. In addition to categorical features encoding, it is often good practice to perform some type of scaling on numerical features. Applying a scaling to the data does not change the shape of each feature's distribution. Normalization ensures that each feature is treated equally when applying supervised learners. In our case, we used MinMaxScaler function to normalize the numerical features.

Table 14: Features selected

Dependent/ Independent	Feature
Dependent feature	Certified
Independent features (course related)	course duration, number of course chapters, course subject
Independent features (demographics related)	Country, level of education, gender, region
Independent features (learners' activity related)	Total number of events, active days, number of videos played, proportion of chapters accessed, number of forum posts, proportion of user's lifetime in course duration

6.3 Train/ test split

One big concern with machine learning is the generalization. To pursue better generalization, we split our dataset into training dataset and test dataset. So that we can train the model with the training dataset and test the model with the test dataset. The *train_test_split* function from Scikit-Learn's *model_selection* module splits arrays or matrices into random training and test subsets. When the *shuffle* parameter is set to True, the data will be shuffled before splitting. For test dataset size, in our case, we set it to be 0.2. If we do not set the *random_state* parameter, for every iteration, the training set and test set would be different with another iteration. This will somewhat reduce the random variance when selecting the training and test set. We will set the *random_state* here, for the statistical test on classifiers we will not set the *random_state*.

6.4 Model selection

Since our target is labeled (1 for certified 0 for not certified), this is a supervised learning problem. And what we are predicting is a binary feature, so we could use classifiers.

As indicated in the machine learning map in Scikit-Learn[3], if our training examples are not too large (>100,000), then the classifiers we could choose are *SVM*, *KNN*, Naive Bayes and Ensemble Methods. Naive Bayes is good for text data. Since our data is not text data, we will skip Naive Bayes. *KNN* is a lazy learner. Although it is simple, it could be very slow, takes a lot of memory. We will also skip KNN. We will select *SVM* and an Ensemble method - Random Forest. Besides these, we will also try a logistic regression model which could also be a good candidate.

Below we will briefly introduce the real-world application, strength and weakness of each algorithm as well as the reason why we select it.

6.4.1 Logistic Regression

- Real-world application in industry: Online transaction fraud detection
- Strength: Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting. Logistic models can be updated easily with new data using stochastic gradient descent.
- Weakness: Logistic regression tends to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships
- Reason for choosing this model: In our case our m (number of training examples) is large (m=38,622), while n (number of features) is small (n=13).

[3] https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

6.4.2 Random Forest.

- Real-world application in industry: Cancer detection
- Strength: One decision tree model tends to overfit. The strong point of Random forest is that it is a collection of lots of decision tree models. It lets each decision tree model vote and pick up the best model.
- Weakness: Random Forests are much harder and time-consuming to construct due the complexity, and they also require more computational resources and are also less intuitive.
- Reason for choosing this model: First it's an Ensemble method, which could effectively prevent overfitting. Random Forest derives from decision tree models, it could automatically select features.

6.4.3 Support Vector Machine

- Real-world application in industry: Handwriting recognition
- Strength: It works really well in complicated domains where there is a clear margin of separation.
- Weakness: Does not perform well in very large dataset, does not perform well with lots of noise, could be very slow.
- Reason for choosing this model: *SVM* with a linear kernel applies to the condition when the number of training example (m) is large (m=38, 622), the number of features (n) is small (n=13)

6.5 Baseline

For the naïve predictor, we always predict negative. That is, the learner is not certified. Since we are always predicting negative, the true positive and false positive are both 0. False negative would be the 1s in the dataset that we predicted as 0s, then the size of False negative should be the number of the 1s in our dataset. True negative would be all the 0s in the dataset, and the size should be the number of 0s in the dataset.

Precision is defined as for those we predicted as certified, how many are actually certified. In the naive model we are predicting 0 students are certified, and among these 0 students predicted certified, 0 are really certified. The precision is somewhat 0/0 = 100%. The precision of 0/ 0 looks a little weird, but we could think it this way: we made no mistake in our prediction, so our precision is 100%.

Recall is defined as for those actually certified, how many do we correctly predicted as certified.

Our objective is to identify those who will not be certified and suggest possible intervention accordingly. We would like to make sure for those our model predicts as certified, are certified. Otherwise we may let go a student who needs intervention since we predict him as certified. That is to say, we want our F-score to be close to precision. We can achieve this by setting beta close to 0. We will try beta = 0.1 to see how well our model can predict. We found that for the naive predictor, the accuracy score is 0.9661, the F-score: 0.

Other researchers who also work on predicting retention achieved accuracy scores of 82% - 90% (Fonti 2015) and 89% (Dalipi, Imran, and Kastrati 2018). Based on the related research, we will choose an average accuracy, i.e. 87% as the baseline. The

related research did not measure the F-score, so we are not going to take a F-score baseline from the earlier research.

6.6 Initial model evaluation

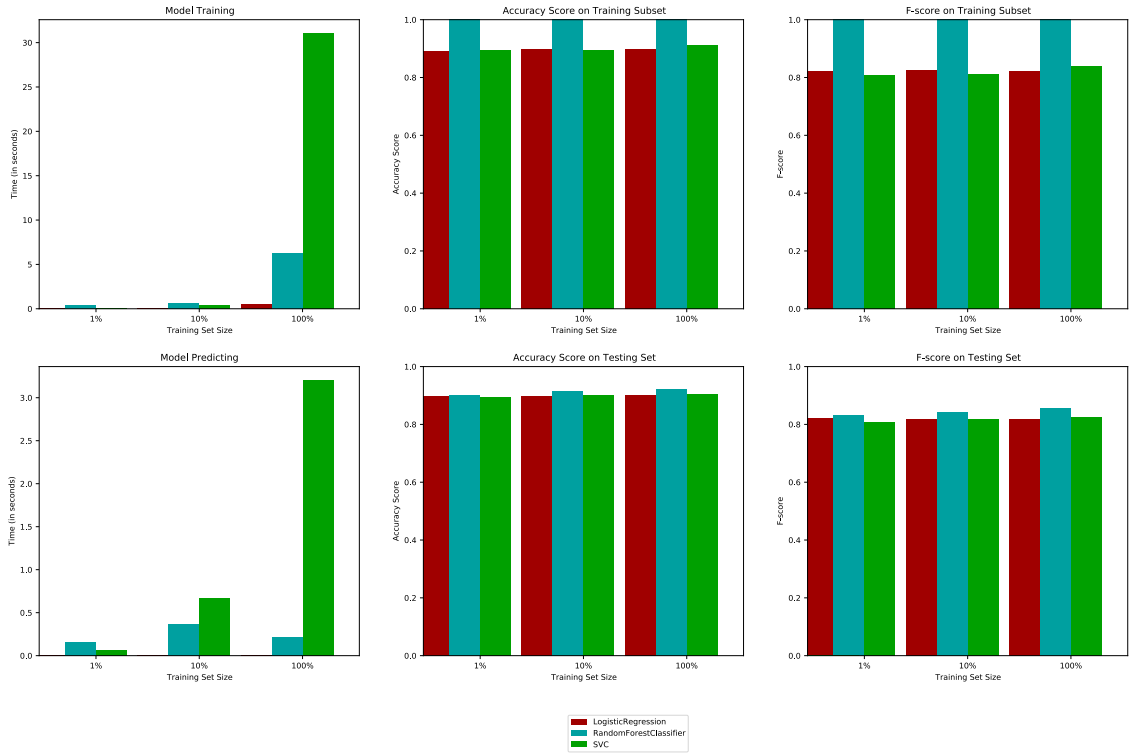

Figure 28: Performance metrics for three supervised learning models

We train and test our dataset on the three classifiers mentioned above. The result is show in Figure 28. The evaluation metrics we are using include running time, accuracy score and F-score. We can see that the Random Forest classifier outperforms the other two. But how reliable is this result? We will do a statistical test below to check the significance of the difference between the classifiers.

To confirm whether one classifier really over-perform another, we use statistical test to check the significance difference between two classifiers. In this paper we are going to use Mcnemar's test. In (Dietterich 1998), the author reviewed five statistical tests

including Mcnemar's test. He found that for algorithms that can be executed only once, Mcnemar's test was the only test with acceptable Type I error (false positive).

To apply the Mcnemar's test, we randomly split our dataset into training set and test set. We train both classifiers A and B on the training set and test them on the test set. Then we record the errors in a contingency table, like shown in Table below. Then we will use a chi-squared test to test the difference. To avoid the training set and test set variance, we run 5 iterations. For each iteration, both the training set and the test set will be different.

Our sample size of the training dataset is large (38,622). This large sample size would have an effect on P-value. To deal with the large sample, we train the models with 1% 10% and 100% of the training dataset. Then we will check the P-values of each sample size.

The null hypothesis (H_0) is there is no significant difference between the performance of the two classifiers. While the alternative hypothesis (H_1) is that there is significant difference between the performance of the two classifiers. The results are as following:

- Logistic Regression vs. Random Forest: For all 5 iterations, for each sample size, the largest P-value is 0.019. We reject the H_0 and accept that there is significant difference between the performance of the two classifiers.
- *SVM* vs. Random Forest: When the sample size is 1% of thee training dataset, the largest P-value is 0.4. For 10% and 100% training dataset, P-value $P<.001$. We reject the H_0 and accept that there is significant difference between the performance of the two classifiers.

6.7 Optimized model

Based on explorations above, Random Forest is the best classifier. We will tune the parameters to come up with an optimized Random Forest classifier. We will explore how well the best model can predict. After that we explore the feature importance of the best model. The parameters of the optimized model are shown in Figure 29. The final accuracy score on the test data is 0.92. The final F-score on the test data is 0.84.

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=10, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=300,
                       n_jobs=None, oob_score=False, random_state=42, verbose=0,
                       warm_start=False)
```

Figure 29: Parameters of the best classifier

The feature importance is shown in Figure 30. Learners' activity related features play an important role in predicting certification. Course related features are of some importance in predicting certification. While demographics features have subtle impact on predicting certification.

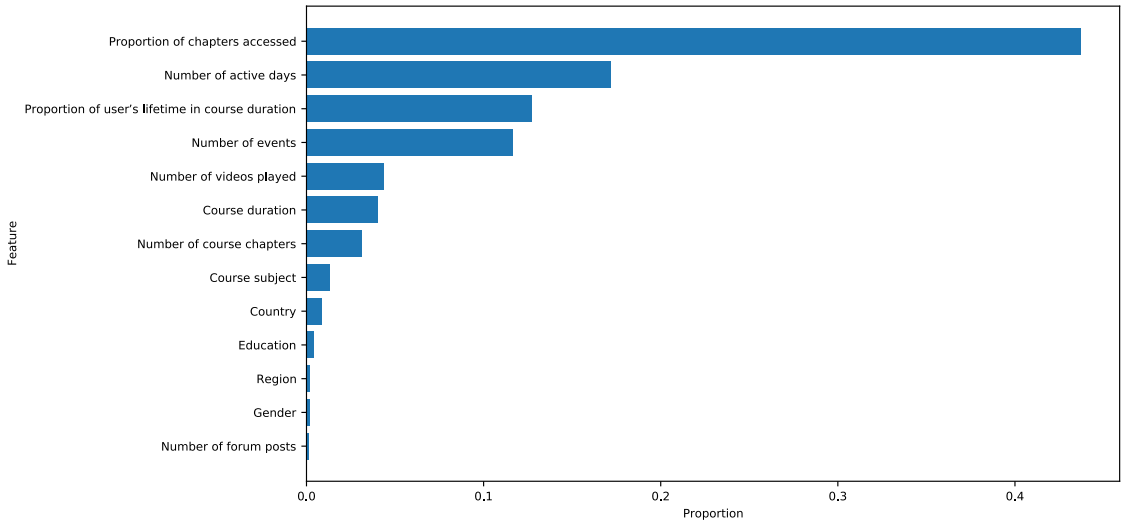

Figure 30: Feature importance

Section 7 Discussion and conclusions

7.1 Conclusions

MOOC has the potential to provide high quality college level education at scale. Like many other new technologies, it has followed the Gartner's Hype Cycle pattern. In the Hype Cycle model, *MOOC* is now at the Plateau of Productivity phase. There is still a lot of research needs to be done on *MOOC*. The retention problem is one such research area.

In this study, we explored the pattern of retention in *MOOC*. Specifically, we explored the certification rate of each subpopulation and features affecting retention in *MOOC*. We first conducted descriptive analysis on the datasets. We then conducted inferential statistical analysis to test the significance of the findings from the descriptive analysis. Finally, we built machine learning models to check how well the identified features can predict retention.

The result of the certification rate of each subpopulation is shown in Table 15 below.

Table 15: Certification rate of different subpopulations

Learners subpopulation	Minimum (%)	Maximum (%)	Mean (%)	Standard deviation (%)
Registered	1.26	9.04	3.80	1.83
Viewed	2.61	14.28	6.15	3.11
Explored	14.59	80.00	57.21	17.00
Intended to complete	13.00	100.00	32.00	19.67

The certification rates of *registered* range from 1.26% to 9.04%, with a mean of 3.80% and a standard deviation of 1.83%. About 60% of the *registered* viewed the course. The certification rates of *viewed* vary from 2.61% to 14.28%, with a mean of 6.15% and standard deviation of 3.11%. Among the *registered*, less than 10% explored the course. The certification rate of *explored* varies from 14.59% to 80.00%, with a mean of 57.21% and standard deviation of 17.00%.

To better understand the phenomenon with less bias, we defined intended to complete as those who intended to complete the course. This was defined with three conditions:

- Condition one: The learner's lifetime is more than 13% of the course duration
- Condition two: The learner interacted with at least 13% of the chapters
- Condition three: The learner's grade is greater than zero

We found that 12% of the *registered* in our dataset intended to complete a course. The certification rate of intended to complete ranges from 13% to 100%, with a mean of 32% and standard deviation of 19.67%.

For course related features we found that institution, semester and course subject do not affect retention. There is a moderate positive correlation between course duration and certification rate. There is a weak negative correlation between number of course chapters and certification rate.

For demographic features we found that country, education, gender, region affects retention rate. Age do not affect retention.

For learners' activity related features, we found that total number of events, active days, number of videos played, proportion of chapters accessed and proportion of user's

lifetime in course duration are correlated with certification. Among these features, active days and proportion of chapters accessed have a strong correlation with certification. Number of forum posts and enrollment days relative to launch do not affect certification.

Based on the exploration above, we suggest that *MOOC* should foster more active participation. Instructors could work on getting a more appropriate course duration.

7.2 Limitations and suggestions for additional research

In this research we used secondary data. The year 1 dataset comes with very good data descriptions, but the year 4 dataset doesn't. For features in year 4 dataset, we interpreted them according to the workpaper where the dataset is published (Chuang and Ho 2016). There are chances that we may interpret any feature in a slightly different way with the original intention. This also applies to the year 1 dataset although there is a data description.

The datasets do not collect the information for those who intended to complete the course. We have to define this with our own conditions. The true cohort of intended to complete may not be well represented. However, in future work, this can be achieved in a few ways. A pre-course survey could acquire learners' level of intention to complete the course. Besides, by checking learns' activities on *MOOC* platforms, we could tell whether they are actively interacting with the course. Combining pre-course survey and learners' learning activities, we should be able to better capture the learners' intention to complete. There is also another technique that could be used. Our target cohort could have significantly different behaviors. We could also use an unsupervised machine learning algorithm to cluster the target cohort, which is those who intended to complete

the course.

In this research we used an analogy method to explore retention in *MOOC* based on our understanding of the real-world high education. However, it should be noted that there are also some features specific to *MOOC*. For example, the length of the chunked lecture (short video), the way the assignments are reviewed (peer-reviewed or auto-graded), the embedded quizzes, whether a course acknowledge credit transfer or not. These features are not available in our datasets, but they worth further exploring.

In this research, we have focused on the observable features. However, behind these observable features, there are some latent factors that are not observable. For example, the learners' motivation, isolation, busyness, society status and so on. Some of the latent factors could be the real reason behind our identified features. Further research needs to be done to deeper explore this. In this case, structural equation model could serve as a powerful tool.

Acknowledgements

My sincere thanks to my advisor Professor Danny Fernandes, for all the patient guidance and invaluable advices he has given me. I would also like to thank Professor Nakamoto Yuki and Professor Ohshima Hiroaki for reviewing my paper, attending my presentation and providing insightful suggestions.

References

Adamopoulos, Panagiotis. 2013. “What Makes a Great MOOC? An Interdisciplinary Analysis of Student Retention in Online Courses.”

Belanger, Yvonne, Jessica Thornton, and Roger C Barr. 2013. “Bioelectricity: A Quantitative Approach--Duke University’s First MOOC.” *EducationXPress* 2013 (2): 1–1.

Bloom, Benjamin S. 1984. “The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring.” *Educational Researcher* 13 (6). Sage Publications Sage CA: Thousand Oaks, CA: 4–16.

Bozkurt, Aras, Nilgun Ozdamar Keskin, and Inge de Waard. 2016. “Research Trends in Massive Open Online Course (MOOC) Theses and Dissertations: Surfing the Tsunami Wave.” *Open Praxis* 8 (3). International Council for Open and Distance Education: 203–21.

Breslow, Lori, David E Pritchard, Jennifer DeBoer, Glenda S Stump, Andrew D Ho, and Daniel T Seaton. 2013. “Studying Learning in the Worldwide Classroom Research into EdX’s First MOOC.” *Research & Practice in Assessment* 8. ERIC: 13–25.

Chuang, Isaac, and Andrew Ho. 2016. “HarvardX and MITx: Four Years of Open Online Courses--Fall 2012-Summer 2016.” *Available at SSRN 2889436*.

Creelman, Alastair. 2013. “COMPLETION RATES – A FALSE TRAIL TO MEASURING COURSE QUALITY? LET’S CALL IN THE HEROES INSTEAD” 16 (2): 10.

Crossley, Scott, Luc Paquette, Mihai Dascalu, Danielle S. McNamara, and Ryan S. Baker. 2016. “Combining Click-Stream Data with NLP Tools to Better Understand MOOC Completion.” In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK ’16*, 6–14. Edinburgh, United Kingdom: ACM Press.

Dalipi, Fisnik, Ali Shariq Imran, and Zenun Kastrati. 2018. “MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges.” In , 1007–14. IEEE.

Daniel, John. 2012. “Making Sense of MOOCs: Musings in a Maze of Myth, Paradox and Possibility.” *Journal of Interactive Media in Education* 2012 (3). Ubiquity Press.

Dietterich, Thomas G. 1998. “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms.” *Neural Computation* 10 (7). MIT Press: 1895–1923.

Downes, Stephen. 2008. “Places to Go: Connectivism & Connective Knowledge.” *Innovate: Journal of Online Education* 5 (1): 6.

El Said, Ghada Refaat. 2017. “Understanding How Learners Use Massive Open Online Courses and Why They Drop Out: Thematic Analysis of an Interview Study in a

Developing Country." *Journal of Educational Computing Research* 55 (5): 724–52.

Emanuel, Ezekiel J. 2013. "Online Education: MOOCs Taken by Educated Few." *Nature* 503 (7476). Nature Publishing Group: 342.

Fonti, Mary L. 2015. "A Predictive Modeling System: Early Identification of Students at-Risk Enrolled in Online Learning Programs."

Gaebel, Michael. 2014. *MOOCs: Massive Open Online Courses*. EUA.

Gregori, Elena Barberà, Jingjing Zhang, Cristina Galván-Fernández, and Francisco de Asís Fernández-Navarro. 2018. "Learner Support in MOOCs: Identifying Variables Linked to Completion." *Computers & Education* 122 (July): 153–68.

Ho, Andrew Dean, Justin Reich, Sergiy O Nesterko, Daniel Thomas Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang. 2014. "HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013." *SSRN Electronic Journal*.

Hollands, Fiona M, and Devayani Tirthali. 2014. "MOOCs: Expectations and Reality." *Center for Benefit-Cost Studies of Education, Teachers College, Columbia University* 138.

Jordan, Katy. 2014. "Initial Trends in Enrolment and Completion of Massive Open Online Courses." *The International Review of Research in Open and Distributed Learning* 15 (1).

Jordan, Katy. 2015. "Massive Open Online Course Completion Rates Revisited: Assessment, Length and Attrition." *The International Review of Research in Open and Distributed Learning* 16 (3).

Khalil, Hanan, and Martin Ebner. 2014. "MOOCs Completion Rates and Possible Methods to Improve Retention - A Literature Review," 9.

Koller, Daphne, Andrew Ng, Chuong Do, and Zhenghao Chen. 2013. "Retention and Intention in Massive Open Online Courses: In Depth." *Educause Review* 48 (3): 62–63.

Levitz, Randi S, Lee Noel, and Beth J Richter. 1999. "Strategic Moves for Retention Success." *New Directions for Higher Education* 1999 (108). Citeseer: 31–49.

Lin, Mingfeng, Henry C Lucas Jr, and Galit Shmueli. 2013. "Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem." *Information Systems Research* 24 (4). INFORMS: 906–17.

Liyanagunawardena, Tharindu Rekha, Andrew Alexandar Adams, and Shirley Ann Williams. 2013. "MOOCs: A Systematic Study of the Published Literature 2008-2012." *The International Review of Research in Open and Distributed Learning* 14 (3): 202.

Mongkhonvanit, Kritphong, Klint Kanopka, and David Lang. 2019. "Deep Knowledge Tracing and Engagement with MOOCs." In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, 340–42. Tempe, AZ, USA: ACM Press.

Norvig, Peter. 2012. “The 100,000 Student Classroom.” *TED Talk. June* 22.

Pappano, Laura. 2012. “The Year of the MOOC.” *The New York Times* 2 (12): 2012.

Reich, Justin. 2014. “MOOC Completion and Retention in the Context of Student Intent.” *EDUCAUSE Review Online* 8.

Rodriguez, C Osvaldo. 2012. “MOOCs and the AI-Stanford Like Courses: Two Successful and Distinct Course Formats for Massive Open Online Courses.” *European Journal of Open, Distance and E-Learning*. ERIC.

Yang, Diyi, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. 2013. “Turn on, Tune in, Drop out: Anticipating Student Dropouts in Massive Open Online Courses.” In , 11:14.

Zawacki-Richter, Olaf, Aras Bozkurt, Uthman Alturki, and Ahmed Aldraiweesh. 2018. “What Research Says About MOOCs – An Explorative Content Analysis.” *The International Review of Research in Open and Distributed Learning* 19 (1).

Appendix

I. Python code for year 1 dataset analysis

Refer to the Github repository:

https://github.com/RayWongGit/HarvardX_MITx_Person_Course_Dataset_Exploration

II. Python code for year 4 dataset analysis

Refer to the Github repository:

https://github.com/RayWongGit/Edx_year4_dataset_exploration